ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Συστήματα Αυτόματης Οργάνωσης Κειμένων με Βάση το Περιεχόμενο και το Ύφος

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Τσιμπουκάκης Κ. Νικόλαος

Αθήνα, Νοέμβριος 2009

.....
Νικόλαος Κ. Τσιμπουκάκης
Διδάκτωρ Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΣΗΜΑΤΩΝ, ΕΛΕΓΧΟΥ ΚΑΙ ΡΟΜΠΟΤΙΚΗΣ

Ηρώων Πολυτεχνείου 9
157 73 Ζωγράφου
Αθήνα

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

DEPARTMENT OF ELECTRICAL AND COMPUTER ENGINEERING

DIVISION OF SIGNALS, CONTROL AND ROBOTICS

9, Iroon Polytechniou Str.
157 73 Zografou
Athens, GREECE

ΔΙΔΑΚΤΟΡΙΚΗ ΔΙΑΤΡΙΒΗ

Τσιμπουκάκης Κ. Νικόλαος
Διπλωματούχος της Σχολής Ηλεκτρολόγων Μηχανικών & Μηχανικών Υπολογιστών
του Εθνικού Μετσοβίου Πολυτεχνείου

«Συστήματα Αυτόματης Οργάνωσης Κειμένων με Βάση το Περιεχόμενο και το Ύφος»

Τριμελής Συμβουλευτική Επιτροπή: Γ. Καραγιάννης (Επιβλέπων)
Ν. Μαράτος
Γ. Ταμπουρατζής

Εγκρίθηκε από την Επταμελή Εξεταστική Επιτροπή στις 23 Νοεμβρίου 2009

Γ. Καραγιάννης
Καθηγητής Ε.Μ.Π.

Ν. Μαράτος
Καθηγητή Ε.Μ.Π.

Γ. Ταμπουρατζής
Ερευνητή Α΄ στο Ερευνητικό Κέντρο Αθηνά.

Π. Μαραγκός
Καθηγητή Ε.Μ.Π.

Π. Τσανάκας
Καθηγητή Ε.Μ.Π.

Β. Μέρτζιος
Καθηγητή Δ.Π.Θ.

Χ. Χαραλαμπάκης
Καθηγητή Ε.Κ.Π.Α.

ΚΕΦΑΛΑΙΟ 1ο

1.1 Εισαγωγή

Αντικείμενο της παρούσας εργασίας είναι η μελέτη και υλοποίηση συστημάτων ικανών να κατηγοριοποιούν με αυτόματο τρόπο κείμενα γραμμένα σε ελληνική γλώσσα. Η κατηγοριοποίηση μπορεί να αφορά στο θεματικό περιεχόμενο της συλλογής κειμένων, αλλά μπορεί και να βασίζεται σε άλλου είδους κριτήρια διαχωρισμού, όπως, παραδείγματος χάριν, η αναγνώριση – ταυτοποίηση του συγγραφέα, τα οποία καθορίζονται από τις ανάγκες του εκάστοτε χρήστη. Τα συστήματα που αναπτύχθηκαν προσαρμόζονται εύκολα στο εκάστοτε σύνολο δεδομένων προς επεξεργασία χωρίς να απαιτούνται αλλαγές στην υλοποίησή τους. Χρησιμοποιούν μόνο ένα σύνολο δεδομένων εκπαίδευσης βάσει του οποίου προσπαθούν να επιλύσουν το πρόβλημα. Σημειωτέον ότι τα εν λόγω συστήματα εφαρμόσθηκαν σε πραγματικά δεδομένα, έτσι ώστε τα αποτελέσματα της συγκεκριμένης έρευνας να είναι αξιοποιήσιμα σε πρακτικά προβλήματα.

Η παρούσα εργασία αποσκοπεί στην ανάπτυξη μεθόδων για την αποτελεσματική εξαγωγή πληροφορίας από κειμενικά δεδομένα, με αυτόνομο τρόπο, με στόχο την ανάλυση των κειμένων αυτών και τη διευκόλυνση της ανάκτησης πληροφορίας. Για το λόγο αυτό αξιοποιούνται τεχνικές από τους τομείς της στατιστικής, αλλά και της εκμάθησης μηχανών, οι οποίες επιτρέπουν την αυτόματη εξαγωγή μορφολογικής ή σημασιολογικής πληροφορίας από κείμενα (ο όρος «αυτόματη εξαγωγή» υποδηλώνει ότι η λειτουργία αυτή θα πραγματοποιηθεί με την ελάχιστη δυνατή ανθρώπινη παρέμβαση / καθοδήγηση). Τεχνικές από τους χώρους αυτούς αξιοποιήθηκαν για το διαχωρισμό των κειμένων βάσει των χαρακτηριστικών των συγγραφέων τους. Για τον προσδιορισμό των αποτελεσματικότερων τεχνικών μελετήθηκαν διάφορες εναλλακτικές μέθοδοι βασισμένες κυρίως σε βιολογικά μοντέλα. Τέλος, σημαντικό ρόλο στην παρούσα εργασία διαδραμάτισαν γλωσσολογικές μελέτες (κάποιες από τις οποίες έχουν υλοποιηθεί και από το ΙΕΛ / Ε.Κ. ΑΘΗΝΑ), με βάση τις οποίες προσδιορίσθηκαν τα πιο πρόσφορα χαρακτηριστικά για την αποτύπωση του ύφους.

Τα συστήματα που αναπτύχθηκαν αποσκοπούσαν στην επίτευξη δύο διακριτών στόχων: α) την αυτόματη κατηγοριοποίηση κειμένων με βάση το δημιουργό τους και β) την αυτόματη κατηγοριοποίηση κειμένων ανάλογα με το θεματικό τους περιεχόμενο, δηλαδή το θέμα που πραγματεύονται. Για το λόγο αυτό διεξήχθησαν μία σειρά από διαφορετικά είδη πειραμάτων πάνω σε ποικίλα σύνολα δεδομένων, κατάλληλα χαρακτηρισμένων ώστε να είναι συμβατά με τους επιθυμητούς στόχους κάθε πειράματος.

Η δομή της εργασίας έχει ως εξής: Στο κεφάλαιο 2 γίνεται μία συνοπτική παρουσίαση εργασιών που πραγματεύονται αντίστοιχα προβλήματα οργάνωσης κειμένων με βάση το περιεχόμενο και το ύφος, καθώς επίσης και αναφορά στις τεχνικές που χρησιμοποιούνται στη βιβλιογραφία. Στο κεφάλαιο 3 αναπτύσσεται το σύστημα οργάνωσης κειμένων με βάση το συγγραφέα και παρουσιάζονται τα πειραματικά αποτελέσματα. Επίσης γίνεται αξιολόγηση της προτεινόμενης μεθοδολογίας και σύγκριση του εν λόγω συστήματος με άλλα. Στο κεφάλαιο 4 παρουσιάζεται το σύστημα κατηγοριοποίησης κειμένων με βάση το περιεχόμενο και αναλύονται εκτενώς τα πειραματικά αποτελέσματα της εφαρμογής του. Στο κεφάλαιο 5 περιγράφονται κάποιες εναλλακτικές υλοποιήσεις του συστήματος του κεφαλαίου 4 βασισμένες σε διαφορετικές τεχνικές, ενώ γίνεται και σύγκριση με καθιερωμένες μεθόδους. Η εργασία ολοκληρώνεται με τα γενικά συμπεράσματα, αλλά και τις μελλοντικές κατευθύνσεις προς την ανάπτυξη αντίστοιχων συστημάτων (κεφάλαιο 6).

1.2 Χρησιμότητα των συστημάτων αυτόματης οργάνωσης κειμένων

Ο όγκος της πληροφορίας σε ηλεκτρονική μορφή αυξάνεται συνεχώς με όλο και γρηγορότερους ρυθμούς, εξαιτίας της ραγδαίας εξάπλωσης της χρήσης των ηλεκτρονικών υπολογιστών σε διάφορους τομείς δραστηριοτήτων. Για παράδειγμα, οι περισσότεροι οργανισμοί χρησιμοποιούν μεθόδους οργάνωσης των διαδικασιών τους και διαχείρισης των πόρων βασισμένες σε ηλεκτρονικά συστήματα αποθήκευσης. Επιπλέον η αλληλογραφία στα σύγχρονα περιβάλλοντα εργασίας γίνεται κυρίως ηλεκτρονικά, ενώ και οι εμπορικές συναλλαγές ή οι διαπραγματεύσεις ακόμη και μεταξύ μεγάλων οργανισμών λαμβάνουν χώρα σε εικονικές αγορές (e-marketplaces). Ομοίως, και οι πληροφορίες που γίνονται διαθέσιμες στο ευρύ κοινό μέσω ελεύθερων δημοσιεύσεων αυξάνονται με γοργούς ρυθμούς μέσα από τους ιστότοπους του διαδικτύου (blogs), με αποτέλεσμα η αναζήτηση της χρήσιμης για το χρήστη πληροφορίας να γίνεται περίπλοκη. Σε όλους αυτούς τους ψηφιακούς τόπους η γλώσσα που χρησιμοποιείται ως μέσο επικοινωνίας είναι η φυσική γλώσσα (natural language).

Η αύξηση του όγκου των δεδομένων είναι δυνατό να οδηγήσει πρακτικά στην αχρήστευση πολύ μεγάλου μέρους τους, καθώς είναι δύσκολο να εντοπισθεί και να αξιοποιηθεί η εκάστοτε αξιόλογη πληροφορία. Συνεπώς, η αναζήτηση της υπάρχουσας πληροφορίας θα πρέπει να γίνεται με αποδοτικό και φιλικό προς το χρήστη τρόπο, ώστε να εξασφαλίζεται η αξιοποίησή της. Το μεγαλύτερο μέρος της πληροφορίας έχει τη μορφή απλού κειμένου και, κατά συνέπεια, οι προσπάθειες για την εξόρυξη της πληροφορίας στρέφονται κατά κύριο λόγο στην επεξεργασία φυσικής γλώσσας.

Παράλληλα, καθίσταται αναγκαία η καλύτερη οργάνωση-αρχειοθέτηση μεγάλων σωμάτων κειμένων, τα οποία παράγονται και διακινούνται από διαφόρους οργανισμούς, με στόχο την ευκολότερη και προσαρμοσμένη πρόσβαση σε αυτά χρηστών με ποικίλες ανάγκες. Για παράδειγμα, θα ήταν χρήσιμο τα μηνύματα παραπόνων σχετικά με προϊόντα να δρομολογούνται αυτόματα στο αρμόδιο τμήμα μίας μεγάλης εταιρείας.

Η ταξινόμηση κειμένων επιτρέπει την οργάνωση ενός σώματος κειμένων με βάση κάποιο κοινό χαρακτηριστικό, διευκολύνοντας την αναζήτηση στο σώμα αυτό. Συχνά η ταξινόμηση κειμένων γίνεται με βάση προεπιλεγμένες κατηγορίες, με αποτέλεσμα να απαιτείται η προσαρμογή – εκπαίδευση του χρήστη στους κανόνες λειτουργίας του εκάστοτε συστήματος. Αυτό δυσκολεύει την αναζήτηση, αφού δεν είναι πάντοτε αυτονόητο σε ποια θεματική κατηγορία ανήκει ένα κείμενο. Πολλές φορές καμία κατηγορία του συστήματος δεν μπορεί να καλύψει τις απαιτήσεις του χρήστη, ιδιαίτερα όταν αυτές είναι αρκετά εξειδικευμένες. Σε άλλες εφαρμογές η λύση στο συγκεκριμένο πρόβλημα δίνεται με την κατάταξη ενός κειμένου σε περισσότερες από μία κατηγορίες. Ωστόσο, μία τέτοια επιλογή έχει συχνά ως αποτέλεσμα την επιστροφή από το σύστημα πλεονάζουσας ή ακόμα και άχρηστης πληροφορίας σε σχέση με αυτή που αναμένει ο χρήστης.

Γενικά η ταξινόμηση κειμένων είναι μία αρκετά δύσκολη και επίπονη διαδικασία, ενώ συχνά ο όγκος των κειμένων είναι τόσο μεγάλος που καθιστά απαγορευτική την εκτέλεσή της από τον άνθρωπο. Είναι λοιπόν αναγκαίο η ταξινόμηση να γίνεται από μηχανές, δηλαδή από ηλεκτρονικούς υπολογιστές, με αυτόματο τρόπο, ώστε να μην απαιτείται συνεχής ανθρώπινη παρέμβαση. Από την άλλη μεριά, ωστόσο, αυτή η διαδικασία θα πρέπει να σχεδιαστεί με ιδιαίτερη προσοχή, ώστε τα αποτελέσματά της να συμβαδίζουν κατά το δυνατό με τον τρόπο σκέψης του ανθρώπου. Με άλλα λόγια, οι κατηγορίες που θα επιλεγούν πρέπει να είναι, στο βαθμό που αυτό είναι εφικτό, σε αρμονία με την ανθρώπινη αντίληψη. Για παράδειγμα, είναι εύλογο για τον άνθρωπο να αναζητά κείμενα με βάση το συγγραφέα ή το περιεχόμενό τους παρά με βάση το μέγεθός τους, που είναι μία παράμετρος ιδιαίτερα εύκολη στο χειρισμό από τους ηλεκτρονικούς υπολογιστές.

Πέρα όμως από τις ιδιαίτερες ανάγκες που μπορεί να καλύπτουν τα συστήματα οργάνωσης κειμένων, το ζητούμενο είναι πάντοτε η μέγιστη δυνατή ακρίβεια και αποτελεσματικότητα, ώστε να επιστρέφεται στο χρήστη η μέγιστη ποσότητα δεδομένων σχετικών με τις απαιτήσεις του. Η αποτελεσματικότητα των αλγόριθμων αναζήτησης καθορίζεται από δύο πολύ σημαντικές παραμέτρους, την **ακρίβεια** (precision) και την **ανάκληση** (recall) (Yiming Yang, 1999). Ως ακρίβεια ορίζεται το ποσοστό των κειμένων που επιστρέφει ο αλγόριθμος αναζήτησης, τα οποία ανταποκρίνονται στα κριτήρια αναζήτησης του χρήστη, ενώ η ανάκληση πληροφορίας ορίζεται ως το ποσοστό των κειμένων που ανευρέθηκαν ως σχετικά με το θέμα αναζήτησης, σε σχέση με το σύνολο των κειμένων. Το επιθυμητό σε έναν αλγόριθμο αναζήτησης είναι οι αρκετά υψηλές τιμές ταυτόχρονα και στις δύο αυτές μετρικές, αν και κάτι τέτοιο δεν είναι πάντα εφικτό.

Γενικά οι έννοιες της ακρίβειας και της ανάκλησης είναι αντιστρόφως ανάλογες και οι τιμές τους εξαρτώνται κάθε φορά από το δεδομένο πρόβλημα. Πιο συγκεκριμένα, ορισμένες φορές είναι προτιμότερο ένας αλγόριθμος αναζήτησης να μην επιστρέφει τίποτα, όταν δεν βρίσκει κάτι που είναι επαρκώς σχετικό με τα κριτήρια αναζήτησης, ενώ άλλες φορές είναι πιο σημαντικό να επιστραφεί το πιο κοντινό αποτέλεσμα στα κριτήρια αναζήτησης, το οποίο όμως μπορεί και να μην ταιριάζει σε μεγάλο βαθμό με αυτά. Τα τελευταία χρόνια έχουν γίνει πολλές προσπάθειες στον τομέα της επεξεργασίας, οργάνωσης και ταξινόμησης κειμένων με χρήση μεθόδων τεχνητής νοημοσύνης, εκ των οποίων οι κυριότερες χρησιμοποιούν ιδέες προερχόμενες από το χώρο της στατιστικής ανάλυσης δεδομένων και των τεχνητών νευρωνικών δικτύων.

Συνοψίζοντας, αναφέρθηκε ότι η αυξημένη χρήση υπολογιστικών συστημάτων παρέχει μεν καλύτερες δυνατότητες αποθήκευσης πληροφορίας, ταυτόχρονα όμως οδηγεί σε δυσκολίες στην ανάκτηση χρήσιμης πληροφορίας, η οποία συνήθως αποθηκεύεται σε φυσική γλώσσα. Κατά συνέπεια, είναι σκόπιμο αλλά και επιβεβλημένο να αναπτυχθούν μέθοδοι αυτόματης διαχείρισης των δεδομένων, ώστε αυτά να ταξινομούνται σύμφωνα με τα κριτήρια που αντιστοιχούν στις ανάγκες του εκάστοτε χρήστη, αλλά και να διευκολύνουν την προσπέλασή τους με χρήση εύχρηστων καταλόγων-δεικτών (index). Στα πλαίσια αυτού του στόχου η παρούσα εργασία περιγράφει τη δημιουργία συστημάτων ταξινόμησης κειμένων με βάση (α) το θεματικό τους περιεχόμενο και (β) το δημιουργό τους.

1.3 Συνεισφορά

Στην παρούσα παράγραφο δίνεται μία σύνοψη της βασικής συνεισφοράς της διατριβής αυτής στο αντικείμενο της αυτόματης οργάνωσης κειμένων σε κατηγορίες. Στην εργασία αυτή προτάθηκαν συστήματα αυτόματης οργάνωσης κειμένων σε κατηγορίες, με βάση (α) το θεματικό τους περιεχόμενο και (β) το ύφος για την Ελληνική γλώσσα. Συγκεκριμένα, υλοποιήθηκαν αντίστοιχα δύο διακριτά συστήματα, τα οποία χρησιμοποιούν διαφορετικό τρόπο απεικόνισης των κειμένων στο χώρο προτύπων.

Το σύστημα αναγνώρισης ύφους βασίζεται κατά κύριο λόγο σε γλωσσικά χαρακτηριστικά, τα οποία αντικατοπτρίζουν τον ιδιαίτερο τρόπο με τον οποίο ο κάθε συγγραφέας χρησιμοποιεί το λόγο. Στο σύστημα αυτό χρησιμοποιήθηκε το ευρέως διαδεδομένο νευρωνικό δίκτυο MLP (Multi-Layer Perceptron) για το οποίο επιλέχθηκε ο αλγόριθμος εκπαίδευσης Levenberg-Marquardt. Με τη χρήση του αλγορίθμου Levenberg-Marquardt έλαβε χώρα η αξιολόγηση των χαρακτηριστικών με αυτόματο τρόπο η οποία και οδήγησε σε σημαντική απλοποίηση του μοντέλου με μείωση του αριθμού των χαρακτηριστικών του και βελτίωση της απόδοσής του. Επισημαίνεται ότι η τεχνική αυτή δεν έχει εφαρμοσθεί σε πειράματα υφολογίας κατά το παρελθόν σύμφωνα με τη διεθνή βιβλιογραφία. Επιπλέον το προτεινόμενο σύστημα δίνει υψηλή ακρίβεια διαχωρισμού, με απόδοση που προσεγγίζει τις ευρέως διαδεδομένες τεχνικές, οι οποίες θεωρούνται και τεχνικές αναφοράς (state-of-the-art), όπως είναι ο συνδυασμός χαρακτηριστικών TF-IDF (Term Frequency – Inverse Document Frequency) με το SVM (Support Vector Machine) και η LDA (Linear Discriminant Analysis). Τέλος, η μέθοδος επιλογής χαρακτηριστικών που βασίστηκε στον αλγόριθμο Levenberg-Marquardt αποδείχθηκε ικανή να βελτιώσει το αποτέλεσμα της κατηγοριοποίησης ακόμη και όταν χρησιμοποιήθηκε διαφορετικός αλγόριθμος κατηγοριοποίησης από το MLP (πχ το SVM για το οποίο δεν έχει ως τώρα προταθεί στη βιβλιογραφία μέθοδος μείωσης των χαρακτηριστικών).

Το σύστημα αναγνώρισης θεματικών κατηγοριών βασίζεται αρχικά σε μετρήσεις λημμάτων και εν συνεχεία σε χάρτες λέξεων, οι οποίοι προβάλλουν τις νοηματικές συσχετίσεις μεταξύ λέξεων σε ένα χώρο δύο διαστάσεων στηριζόμενοι στο μοντέλο SOM. Οι χάρτες λέξεων δημιουργούνται χωρίς επίβλεψη και μπορούν να δημιουργηθούν για πολύ μεγάλο σύνολο δεδομένων χωρίς καμία ανθρώπινη παρέμβαση. Ο τρόπος επιλογής των χαρακτηριστικών που βασίζεται στη χρήση του κανόνα ABC για τη δημιουργία των χαρτών και η αναπαράσταση των κειμένων στο χώρο των προτύπων είναι καινοτόμοι καθώς δεν

έχουν προταθεί στο παρελθόν για κείμενα καμίας γλώσσας. Παράλληλα με το σύστημα που αναπτύχθηκε βάσει του μοντέλου SOM, προτάθηκαν εναλλακτικά συστήματα που εφάρμοζαν τεχνικές βελτιστοποίησης σμήνους (PSO – Particle Swarm Optimization), μίγματος Γκαουσσιανών κατανομών με δίκτυα ακτινικών συναρτήσεων (GMM/RBF, Gaussian Mixture Models / Radial Basis Functions) και κρυφά Μαρκοβιανά μοντέλα (HMM, Hidden Markov Models). Τα συστήματα PSO και HMM αποτελούν καινοτόμες προτάσεις, αφού τα μοντέλα αυτά δεν έχουν δοκιμασθεί σε προβλήματα κατηγοριοποίησης κειμένων. Κύρια συνεισφορά στην εφαρμογή των μεθόδων αυτών ήταν η απεικόνιση του χώρου προτύπων στη μέθοδο, ώστε να επιλυθεί το αντίστοιχο πρόβλημα. Τα συστήματα PSO και GMM/RBF δεν κατέληξαν σε εξίσου επιτυχημένα αποτελέσματα σε αντίθεση με το HMM, το οποίο κατέστη ικανό να αντικαταστήσει το σύστημα SOM με ελάχιστη απώλεια της ακρίβειας κατηγοριοποίησης. Τα πειράματα έλαβαν χώρα σε μεγάλα σώματα κειμένων τα οποία και συλλέχθηκαν με αυτόματο τρόπο. Οι συνθήκες που δοκιμάσθηκαν τα συστήματα αυτά προσεγγίζουν σε μεγάλο βαθμό τις απαιτήσεις των πρακτικών εφαρμογών. Τέλος το σύστημα αναγνώρισης περιεχομένου παρουσίασε πολύ καλύτερα αποτελέσματα, ακόμη και με λιγότερα χαρακτηριστικά σε σχέση με το συνδυασμό χαρακτηριστικών TF-IDF με το SVM που χρησιμοποιείται ευρέως σε αντίστοιχα προβλήματα.

ΚΕΦΑΛΑΙΟ 2ο

2.1 Αλγόριθμοι επεξεργασίας δεδομένων στο χώρο των προτύπων

Για την οργάνωση κειμενικών δεδομένων οι πιο συχνά χρησιμοποιούμενες τεχνικές είναι οι τεχνικές αναγνώρισης προτύπων (Pattern Recognition) (Duda R.O., Hart P.E. & Stork D.G. 2001) (Καραγιάννης Γ., Σταϊνχάουερ Γ. 2001). Οι τεχνικές αυτές απαντώνται σε ένα πλήθος εφαρμογών αναγνώρισης και αποτελούν το βασικό τρόπο εισαγωγής των υπολογιστικών συστημάτων σε πραγματικά προβλήματα που απαιτούν ευφυΐα πέραν της απλής αναζήτησης και της απόλυτης ταύτισης (exact matching). Ουσιαστικά αποτελούν την διαδικασία που επιτρέπει σε έναν υπολογιστή να επεξεργασθεί ερεθίσματα του πραγματικού κόσμου, με τρόπο που προσεγγίζει την ανθρώπινη αντίληψη, στο βαθμό που αυτό είναι εφικτό. Απώτερος στόχος αυτών των τεχνικών αποτελεί η επίλυση δύσκολων για τους υπολογιστές προβλημάτων κατανόησης, τα οποία όμως αντιμετωπίζονται εύκολα από τους ανθρώπους, όπως είναι η επεξεργασία εικόνων, η αναγνώριση φωνής ή η αναγνώριση χειρόγραφων χαρακτήρων.

Αρχικά οι τεχνικές αναγνώρισης προτύπων προϋποθέτουν την αναπαράσταση κάθε αντικειμένου προς εξέταση με τέτοιο τρόπο, ώστε να μπορεί να εισαχθεί στους ψηφιακούς υπολογιστές. Για το λόγο αυτό χρησιμοποιείται ένα σύνολο από χαρακτηριστικά, τα οποία αντικατοπτρίζουν τις βασικές τους ιδιότητες. Τα χαρακτηριστικά αυτά μπορεί να είναι αριθμητικές τιμές ή συμβολικά δεδομένα και η επιλογή τους είναι βασισμένη στην εμπειρία του σχεδιαστή του συστήματος ή κάποιου ειδικού. Τα χαρακτηριστικά που επιλέγονται για κάποιο αντικείμενο μπορεί να διαφέρουν ανάλογα με τις ανάγκες της εκάστοτε εφαρμογής. Για παράδειγμα, προτιμάται διαφορετική αναπαράσταση για ένα σήμα φωνής σε μία εφαρμογή αναγνώρισης του ομιλητή από ό,τι σε μία εφαρμογή διαχωρισμού μουσικής από φωνή. Η κατάλληλη επιλογή των χαρακτηριστικών συνιστά βασικό παράγοντα επιτυχίας της εφαρμογής. Επιπλέον η αναγνώριση προτύπων βασίζεται στην παραδοχή ότι τα αντικείμενα μίας τάξης έχουν χαρακτηριστικά περισσότερο όμοια από αυτά τα οποία δεν ανήκουν στην τάξη («Αρχή του συμπαγούς»).

Τα συστήματα αναγνώρισης προτύπων έχουν την ικανότητα να προσαρμόζονται στις ανάγκες των εκάστοτε δεδομένων. Συνήθως απαιτούν ένα σύνολο δεδομένων για να προσαρμόσουν τις παραμέτρους τους σε αυτό. Η διαδικασία προσαρμογής των δεδομένων, η οποία ονομάζεται **εκπαίδευση**, είναι αναγκαία για συστήματα επίλυσης πρακτικών προβλημάτων. Η εκπαίδευση μπορεί να απαιτεί χαρακτηρισμένα, ως προς την έξοδο του συστήματος, δεδομένα, οπότε ονομάζεται **επιβλεπόμενη** (supervised), ή να μην απαιτεί κάποιο χαρακτηρισμό, οπότε ονομάζεται **μη επιβλεπόμενη** (unsupervised) (Haykin S. 1999). Παράλληλα, υπάρχουν και συστήματα που βασίζονται σε κανόνες, αλλά και αυτά χρειάζονται κάποια προσαρμογή των κανόνων. Περιπτώσεις συστημάτων βασιζόμενων σε στατικούς κανόνες που δεν απαιτούν εκπαίδευση είναι σχετικά σπάνιες και μη αποδοτικές.

Οι αλγόριθμοι που χρησιμοποιούνται σε προβλήματα αναγνώρισης προτύπων έχουν εμπνευσθεί από ποικίλους κλάδους των μαθηματικών, της επιστήμης της πληροφορίας, της βιολογίας αλλά και από εμπειρικές τεχνικές (Jain A.K., Duin R. P. & Mao J. 2000). Δημοφιλείς στατιστικές μέθοδοι αναγνώρισης προτύπων είναι η **διαχωριστική ανάλυση** (discriminant analysis) (Duda R.O., Hart P.E. & Stork D.G. 2001), η **ανάλυση σε πρωτεύουσες συνιστώσες** (principal components), πολλοί **αλγόριθμοι ομαδοποιήσεων** (cluster analysis) (Everitt B. S., Landau S. & Leese M. 2001) καθώς επίσης και τα **Μαρκοβιανά μοντέλα** (Markov models) (Rabiner L.R. 1989).

Από την επιστήμη της πληροφορίας έχουν αντληθεί μέθοδοι που μετρούν την **εντροπία** (Maximum Entropy), αλλά και μέθοδοι που οδηγούν στην κατασκευή **δένδρων αποφάσεων** (Decision Trees) (Quinlan J. R. 1993). Επιπλέον, πολλές τεχνικές βελτιστοποίησης χρησιμοποιούνται στα συστήματα αναγνώρισης προτύπων ως αλγόριθμοι εκπαίδευσης με σκοπό την εύρεση μίας βέλτιστης κατάστασης για το σύστημα, η πιο απλή από τις οποίες είναι η μέθοδος της **βαθμωτής κατάβασης** (Gradient Descent) (Haykin S. 1999). Πολλές αλγοριθμικές λύσεις έχουν χρησιμοποιηθεί σε συστήματα αναγνώρισης προτύπων, οι οποίες δεν υπάγονται σε κάποια άλλη κατηγορία, όπως είναι οι τεχνικές **σύγκρισης συμβολοακολουθιών**, όπως για παράδειγμα η μετρική Edit-distance (Duda R.O., Hart P.E. & Stork D.G. 2001).

Ο χώρος της βιολογίας έχει εξίσου προσφέρει στην ανάπτυξη αντίστοιχων τεχνικών. Ο τρόπος λειτουργίας των φυσικών συστημάτων, όπως οι κοινωνίες των εντόμων, ή ο μηχανισμός αναπαραγωγής και εξέλιξης στη φύση έχουν αποτελέσει έμπνευση για τη δημιουργία της οικογένειας των **αλγορίθμων βελτιστοποίησης σμήνους** (Particle Swarm Optimization – PSO (Kennedy J. & Eberhart R. C. 2001)), αλλά και των **γενετικών αλγόριθμων** (Genetic Algorithms (Goldberg D. E., 1989 και Holland, J. H. 1975)).

Τέλος, μία αρκετά μεγάλη κατηγορία αλγορίθμων συνιστούν τα **νευρωνικά δίκτυα** (Neural Networks), τα οποία επιχειρούν να προσομοιώσουν τον τρόπο λειτουργίας του ανθρώπινου εγκεφάλου (Haykin S. 1999). Οι αλγόριθμοι αυτοί καταλαμβάνουν σημαντικό τμήμα της παρούσας εργασίας. Το ερευνητικό ενδιαφέρον για τα νευρωνικά δίκτυα προήλθε από την παρατήρηση ότι το ανθρώπινο μυαλό λειτουργεί τελείως διαφορετικά από τις ψηφιακές υπολογιστικές μηχανές και έχει την ικανότητα να είναι αρκετά πιο αποτελεσματικό από αυτές σε προβλήματα αναγνώρισης προτύπων. Ο ανθρώπινος εγκέφαλος αποτελείται από δομικές μονάδες που ονομάζονται **νευρώνες** ή **νευρώνια** (neurons). Οι δομικές αυτές μονάδες συνδέονται και επικοινωνούν μεταξύ τους μεταφέροντας ηλεκτρικά σήματα με συνάψεις που φέρουν βάρη. Οι νευρώνες έχουν την δυνατότητα να επεξεργάζονται τις εισόδους που δέχονται από άλλους νευρώνες και να δίνουν νέα αποτελέσματα στις εξόδους τους. Οι διάφορες περιοχές του εγκεφάλου έχουν τη δυνατότητα να προσαρμόζονται ανάλογα με τις ανάγκες της εκάστοτε λειτουργίας που επιτελούν, ώστε να σημειώνουν τη βέλτιστη απόδοση. Είναι επίσης σημαντικό να αναφερθεί ότι ο ανθρώπινος εγκέφαλος αποτελείται από πολλά διαφορετικά, ως προς τη δομή και τη λειτουργία, είδη νευρωνικών δικτύων.

Μέσα από την προσπάθεια να εξομοιωθεί η λειτουργία του ανθρώπινου εγκεφάλου με τα υπολογιστικά συστήματα για την επίλυση δύσκολων προβλημάτων αναγνώρισης προτύπων, σχεδιάστηκαν τα τεχνητά νευρωνικά δίκτυα. Ένα από τα σημαντικά πλεονεκτήματα των τεχνητών νευρωνικών δικτύων είναι η ικανότητά τους να επιλύουν σύνθετα προβλήματα, αφού τα δομικά τους στοιχεία, οι νευρώνες, περιέχουν μη γραμμικές συναρτήσεις ενεργοποίησης. Τα νευρωνικά δίκτυα μπορούν να κατασκευάζουν, μέσα από τη διαδικασία εκπαίδευσης, μία απεικόνιση της εισόδου στην έξοδο, χωρίς να «γνωρίζουν» από πριν τον όποιο συσχετισμό υπάρχει. Επιπλέον, έχουν τη δυνατότητα να μαθαίνουν εύκολα καινούρια δεδομένα και να προσαρμόζονται στις νέες συνθήκες, χωρίς να απαιτούνται ριζικές αλλαγές. Συνήθως τα νευρωνικά δίκτυα εκτελούν υποσυμβολική επεξεργασία (subsymbolic processing) της πληροφορίας και μπορούν να δώσουν στην έξοδο κάποιο μέτρο, που αντιπροσωπεύει το βαθμό εμπιστοσύνης για το «συμπέρασμα» (π.χ. το δεδομένο πρότυπο ανήκει με 60% βεβαιότητα στην κατηγορία Α) στο οποίο κατέληξαν. Είναι συστήματα παράλληλης επεξεργασίας, καθώς οι δομικές μονάδες λειτουργούν ανεξάρτητα, στοιχείο που τους δίνει το πλεονέκτημα να είναι ανθεκτικά σε περιπτώσεις μερικής βλάβης. Έχει παρατηρηθεί πειραματικά ότι, ακόμα και αν καταστραφεί κάποιο τμήμα ενός τεχνητού νευρωνικού δικτύου, το δίκτυο δεν καταρρέει ολοκληρωτικά, αλλά εξακολουθεί να λειτουργεί και να εξάγει συμπεράσματα, με περισσότερα όμως σφάλματα.

Επιπλέον τα τεχνητά νευρωνικά δίκτυα έχουν τη δυνατότητα να εμφανίζουν ανοχή σε θόρυβο σε αρκετά μεγάλο βαθμό. Το μόνο αρνητικό στοιχείο που τους προσάπτεται είναι ότι λειτουργούν σε μεγάλο βαθμό σαν «μαύρο κουτί» (black box) και δεν έχουν τη δυνατότητα να δώσουν πληροφορίες για το πώς κατέληξαν σε κάποιο συμπέρασμα – έξοδο.

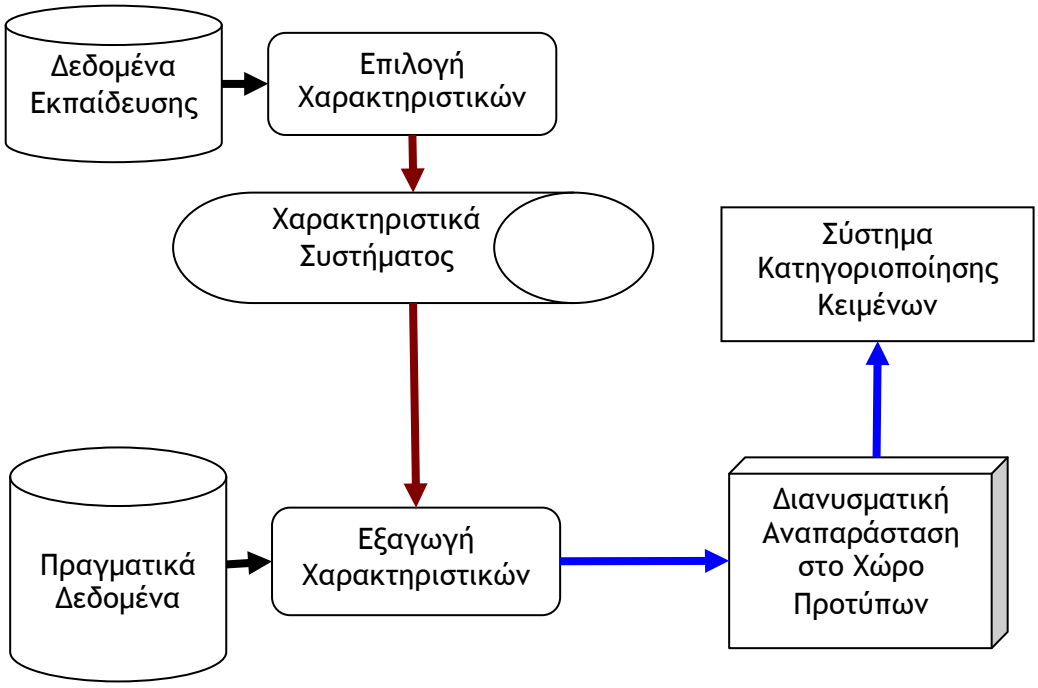
Με βάση αυτές τις ιδιότητες τα τεχνητά νευρωνικά δίκτυα παρουσιάζουν πολύ καλές αποδόσεις σε προβλήματα αναγνώρισης προτύπων. Συχνά χρησιμοποιούνται ως φίλτρα για περιορισμό του θορύβου. Άλλες εφαρμογές τους είναι α) οι **συσχετιστικές μνήμες** (associative memories), οι οποίες αντιστοιχούν κάθε διάνυσμα εισόδου σε μία έξοδο, και β) η **προσέγγιση άγνωστης συνάρτησης** (function approximation or interpolation), δεδομένου ότι είναι γνωστές ορισμένες τιμές της. Τέλος, διάφορα μοντέλα νευρωνικών μπορούν να χρησιμοποιηθούν και για μείωση της διάστασης ενός προβλήματος (dimensionality reduction), την εύρεση σχέσης μεταξύ άγνωστων δεδομένων καθώς και την οπτικοποίηση των δεδομένων σε δύο ή τρεις διαστάσεις.

Ανάλογα με τον τρόπο επεξεργασίας των δεδομένων από τα τεχνητά νευρωνικά δίκτυα, δηλαδή τον τρόπο με τον οποίο μεταφέρονται τα σήματα διαμέσου των συνδέσεων, διαχωρίζουμε τα δίκτυα αυτά σε δίκτυα **προσωτροφοδότησης** (feedforward neural networks) και δίκτυα **αναδρομικά** ή **ανατροφοδότησης** (recurrent neural networks). Στα δίκτυα προσωτροφοδότησης οι έξοδοι από το ένα επίπεδο νευρώνων μεταδίδονται αποκλειστικά και μόνο στα επόμενα επίπεδα. Αντίθετα, στα αναδρομικά δίκτυα ή δίκτυα ανατροφοδότησης οι έξοδοι των νευρώνων ενός επιπέδου μπορεί να συνδέονται ως είσοδοι σε οποιοδήποτε άλλο νευρώνα οποιουδήποτε επιπέδου, ακόμα και προηγούμενου. Αυτό έχει ως αποτέλεσμα η συνολική έξοδος του δικτύου να μεταβάλλεται συνεχώς (**ταλαντώνεται**) μέχρι να καταλήξει σε κάποια ισορροπία (σύγκλιση) μετά από κάποιο αριθμό βημάτων. Με βάση αυτό το διαχωρισμό είναι δυνατόν να αναπτυχθούν νευρωνικά δίκτυα με αρκετές παραλλαγές και με πολύ διαφορετικές αρχιτεκτονικές. Για το λόγο αυτό τα ιδιαίτερα χαρακτηριστικά υλοποίησης του κάθε μοντέλου που χρησιμοποιήθηκε στην παρούσα μελέτη θα αναφέρονται με λεπτομέρεια στο αντίστοιχο κεφάλαιο.

Η μάθηση ή εκπαίδευση είναι μια θεμελιώδης ιδιότητα των τεχνητών νευρωνικών δικτύων, η οποία τους επιτρέπει να εξάγουν πληροφορία από το περιβάλλον τους και να βελτιώνουν την συμπεριφορά τους. Ανάλογα με τον τύπο του δικτύου υπάρχουν διάφορα είδη μάθησης: Μάθηση με επίβλεψη ή επιβλεπόμενη και μη επιβλεπόμενη μάθηση.

Όπως είναι γνωστό, η πλειονότητα των αλγόριθμων αναγνώρισης προτύπων (μεταξύ αυτών και τα νευρωνικά δίκτυα) δέχεται ως είσοδο αριθμητικά δεδομένα (σε αντίθεση με τα συμβολικά). Συνεπώς σε εφαρμογές επεξεργασίας κειμένων καθίσταται αναγκαία η μετατροπή της φυσικής γλώσσας των κειμένων σε ένα σύνολο μετρήσιμων χαρακτηριστικών. Για την εν λόγω μετατροπή ακολουθείται η διαδικασία που απεικονίζεται στο σχήμα 1.1.

Αρχικά, σε ένα μικρό δείγμα κειμένων, σε σχέση με το σύνολο των διαθέσιμων δεδομένων, μετράται ένα αρκετά ευρύ σύνολο από υποψήφια χαρακτηριστικά, τα οποία στη συνέχεια αναλύονται περαιτέρω ως προς τα στατιστικά τους γνωρίσματα (μέση τιμή, διασπορά, εντροπία κ. ά.), πάντοτε όμως σε σχέση με τη επιθυμητή ομαδοποίηση. Ιδιαίτερα σημαντικό είναι το μικρό δείγμα κειμένων να είναι αντιπροσωπευτικό του συνόλου, ώστε τα επιλεγμένα χαρακτηριστικά να μπορούν να αποδώσουν με αποτελεσματικό τρόπο την πληροφορία του συνόλου. Για την αποτίμηση της αξίας των χαρακτηριστικών χρησιμοποιούνται μέθοδοι όπως η **Γραμμική Διακριτική Ανάλυση** (Linear Discriminant Analysis – LDA (Duda R.O., Hart P.E. & Stork D.G. 2001)) ή η **Ανάλυση Πρωτευουσών Συνιστωσών** (Principal Component Analysis – PCA (Duda R.O., Hart P.E. & Stork D.G. 2001), ενώ πολλές φορές αποτιμώνται όλα τα χαρακτηριστικά με τη βοήθεια κάποιου ειδικού. Εν συνεχεία το σύνολο των τελικών χαρακτηριστικών μετράται (εξάγεται) σε όλα τα διαθέσιμα κείμενα και χρησιμοποιείται για την απεικόνιση των κειμένων στο χώρο προτύπων.



Σχήμα 1.1. Τυπικό σύστημα εξαγωγής χαρακτηριστικών

Πριν γίνει οποιαδήποτε μέτρηση στα κείμενα είναι αναγκαία κάποια προεπεξεργασία τους ανάλογη με τη μορφή που είναι αποθηκευμένα. Πολύ συχνά τα κείμενα βρίσκονται στη μορφή κάποιας γλώσσας επισημείωσης (Markup Language), όπως html ή xml, ή είναι αποθηκευμένα σε αρχεία με ειδική μορφοποίηση (pdf, doc, rtf), οπότε καθίσταται απαραίτητο να εξαχθεί από αυτά μόνο το κείμενο, στην απλούστερη δυνατή μορφή. Από τα κείμενα συνήθως αφαιρούνται στοιχεία όπως τα URLs, οι αριθμοί, οι διευθύνσεις e-mail, οι ταχυδρομικές διευθύνσεις, καθώς αυτά τα στοιχεία προσφέρουν λίγη έως μηδαμινή πληροφορία σχετικά με το είδος του κειμένου. Επίσης, λέξεις που βρίσκονται σε διαφορετική γλώσσα από αυτή του κειμένου είναι πιθανότατα άχρηστες για την επεξεργασία. Εξάλλου, αρκετές φορές είναι απαραίτητος και ο ορθογραφικός έλεγχος, καθώς τα τυχόν ορθογραφικά λάθη όχι μόνο δεν παρέχουν πληροφορία σχετικά με το κείμενο, αλλά επιβαρύνουν ως θόρυβος την επεξεργασία.

Οι πολύ σπάνιες λέξεις αγνοούνται κατά την επεξεργασία, η οποία δεν είναι εφικτό να προχωρήσει σε τόσο μεγάλη λεπτομέρεια, καθώς ενδέχεται να οδηγήσουν το σύστημα σε αστάθεια, όταν αυτό βασίζει τα συμπεράσματά του σε δεδομένα με πολύ μικρή συχνότητα εμφάνισης. Αντίστοιχα αγνοούνται και οι πολύ συχνές λέξεις, εκ των οποίων οι περισσότερες είναι οι λεγόμενες λειτουργικές λέξεις (functional words), όπως άρθρα, αντωνυμίες κ.λπ., οι οποίες δεν προσδίδουν κάποιο στοιχείο στο νόημα του λόγου αλλά απλά βοηθούν στην «εξομάλυνσή» του (Manning C. D. & Schütze H. 1999, Freeman R.T. & Yin. H. 2005 και Georgakis A., Kotropoulos C., Xafopoulos A., & Pitas I., 2004). Επιπλέον, σε γλώσσες με πλούσια μορφολογία, όπως η Ελληνική, είναι απαραίτητη η αναγωγή των κλιτών μορφών μίας λέξης σε μία, δηλαδή στο αντίστοιχο λήμμα: για παράδειγμα, οι λέξεις “άνθρωπος” και “ανθρώπου” θα πρέπει να αντιστοιχίζονται στην ίδια μέτρηση. Αφού λοιπόν ολοκληρωθούν η προεπεξεργασία και η λημματοποίηση του κειμένου, τα επιλεγμένα χαρακτηριστικά μετρώνται και κάθε κείμενο απεικονίζεται πλέον ως ένα διάνυσμα χαρακτηριστικών, ικανό να χρησιμοποιηθεί από κάποιο αλγόριθμο αναγνώρισης προτύπων.

2.2 Οργάνωση κειμένων με βάση το συγγραφέα

Κάθε κείμενο είναι γραμμένο με το δικό του ύφος, βάσει του οποίου και βάσει του περιεχομένου του διαφοροποιείται, αλλά και συσχετίζεται με άλλα κείμενα (Stamatatos, Fakotakis Kokkinakis. 2001). Ο τρόπος εκφοράς της γλώσσας σχετίζεται άμεσα με το δημιουργό, αλλά και το συγκεκριμένο είδος του κειμένου. Ως υφολογική ανάλυση ενός κειμένου ορίζεται η ανάλυση – καταγραφή ενός συνόλου από μετρήσιμες ιδιότητες που σχετίζονται στενά με τον τρόπο χρήσης της γλώσσας. Είναι δυνατόν ένα σύνολο κειμένων να διαχωριστεί ως προς το είδος ή το συγγραφέα με βάση τον τρόπο γραφής, όπως αυτός αποτυπώνεται στην υφολογική ανάλυση. Η επιλογή κατάλληλων γλωσσικών χαρακτηριστικών εστιάζει συνεπώς στον ιδιαίτερο τρόπο χρήσης της γλώσσας από ένα συγγραφέα.

Η πλειονότητα των συστημάτων που κατά καιρούς έχουν εφαρμοσθεί σε προβλήματα αναγνώρισης συγγραφέων αποτελείται συνήθως από δύο φάσεις. Στην πρώτη φάση επιλέγεται και μετράται για κάθε κείμενο ένα σύνολο προεπιλεγμένων χαρακτηριστικών. Στη δεύτερη φάση το σύστημα κατατάσσει τα κείμενα σε προεπιλεγμένες κατηγορίες εφαρμόζοντας κάποια τεχνική αναγνώρισης προτύπων. Τα συστήματα που έχουν υλοποιηθεί κατά το παρελθόν διαφοροποιούνται κυρίως με βάση το είδος των χαρακτηριστικών που αξιοποιούν, το είδος των αλγορίθμων κατηγοριοποίησης που χρησιμοποιούν, αλλά και τον τρόπο με τον οποίο εφαρμόζουν τις τεχνικές αναγνώρισης προτύπων, ώστε να επιτύχουν το καλύτερο δυνατό αποτέλεσμα.

Από τα πρώτα πειράματα που έγιναν με χρήση νευρωνικών δικτύων με σκοπό την κατηγοριοποίηση κειμένων βάσει του δημιουργού τους ήταν οι προσπάθειες των Matthews και Merriam (Matthews R. & Merriam T. 1993 και Merriam T. & Matthews R. 1994). Το δικό τους σύστημα κατηγοριοποίησης χρησιμοποιούσε τεχνητά πολυστρωματικά νευρωνικά δίκτυα (multilayer perceptrons – MLP), ενώ ως σύνολο χαρακτηριστικών επιλέγονταν λόγοι που μετρούσαν την προτίμηση σε συγκεκριμένες λέξεις που επέλεγε συστηματικά ο ένας συγγραφέας σε σχέση με τους άλλους. Το πλήθος των χαρακτηριστικών ήταν σχετικά μικρό (5 χαρακτηριστικά) και το πρόβλημα επικεντρωνόταν στην απόδοση κειμένων αγνώστου δημιουργού σε δύο υποψήφιους συγγραφείς (του Shakespeare έναντι του Fletcher και του Shakespeare έναντι του Marlowe αντίστοιχα).

Στο πρόβλημα διαχωρισμού του Shakespeare από το Fletcher δοκιμάσθηκε, για λόγους σύγκρισης, και μία άλλη προσέγγιση (Lowe D. & Matthews R. 1995) με χρήση νευρωνι-

κού δικτύου ακτινικών συναρτήσεων βάσης (RBF) σε συνδυασμό με τις συχνότητες των συχνότερων λημμάτων ως είσοδο για το RBF αντί των λόγων προτίμησης λημμάτων.

Στο αντίστοιχο πρόβλημα απόδοσης ενός άλλου συνόλου κειμένων, συγκεκριμένα των Federalist Papers (Tweedie F., Singh S. & Holmes D. 1996 και Tweedie F., Singh S. & Holmes D. 1996), μετρήθηκαν οι συχνότητες χρήσης ενός μικρού συνόλου λειτουργικών λέξεων και χρησιμοποιήθηκαν ως χαρακτηριστικά σε συνδυασμό με ένα πολυστρωματικό νευρωνικό δίκτυο (MLP). Αρκετά πιο πλούσια ήταν η αναπαράσταση των κειμένων από τους (Gurney P. & Gurney L. 1998a) οι οποίοι χρησιμοποίησαν διάφορες κατηγορίες χαρακτηριστικών (λήμματα, λειτουργικές λέξεις, καθώς και άλλες μετρήσεις σχετιζόμενες με τις εμφανίσεις λημμάτων στο κείμενο). Ως αλγόριθμο κατηγοριοποίησης χρησιμοποίησαν στατιστικές τεχνικές όπως η διακριτική ανάλυση (Linear Discriminant Analysis) και η ανάλυση ομάδων (Cluster Analysis). Στα πλαίσια της προσπάθειας αυτής (Gurney P. & Gurney L. 1998b) αποδείχθηκε ότι τα επιλεγμένα αποσπάσματα των κειμένων δεν ήταν αντιπροσωπευτικά του ύφους.

Μία νεότερη προσπάθεια (Diederich J., Kindermann J., Leopold E. & Paass G. 2003) περιελάμβανε την κατηγοριοποίηση περίπου 2006 κειμένων μίας γερμανικής εφημερίδας από 5 συγγραφείς. Το σύστημα που χρησιμοποιήθηκε βασίστηκε στον αλγόριθμο κατηγοριοποίησης SVM (Vapnik V. 1998), μία τεχνική παραπλήσια με τα νευρωνικά δίκτυα ως προς τη λειτουργία της, η εκπαίδευση της οποίας διέπεται από στατιστικές αρχές. Στο σύστημα αυτό, πέρα από το κλασσικό χαρακτηριστικό της συχνότητας λημμάτων, χρησιμοποιήθηκαν διγράμματα μερών του λόγου τα οποία συνιστούν συντακτικές δομές, καθώς επίσης και ιστογράμματα μήκους λέξεων.

Λιγότερα πειράματα αναγνώρισης συγγραφέα έχουν υλοποιηθεί για την Ελληνική γλώσσα. Πιο συγκεκριμένα, στην εργασία των (Stamatatos E., Fakotakis N., & Kokkinakis G. 2001) χρησιμοποιήθηκε ένα σώμα κειμένων προερχόμενο από την εφημερίδα «Το ΒΗΜΑ», αποτελούμενο από κείμενα που ανήκουν σε 30 περίπου συγγραφείς. Στη μελέτη αυτή δεν ελήφθησαν καθόλου υπόψιν χαρακτηριστικά σε λεξικό επίπεδο, όπως για παράδειγμα οι συχνότητες λημμάτων. Αντίθετα, προτιμήθηκαν μετρήσεις συχνοτήτων συντακτικών δομών, όπως αυτές προήλθαν από εξειδικευμένα γλωσσικά εργαλεία (chunkers, taggers). Ως αλγόριθμος διαχωρισμού επελέγη η στατιστική τεχνική γραμμική διαχωριστική ανάλυση (Linear Discriminant Analysis).

Ένα σύνολο από αντίστοιχα πειράματα είχε λάβει χώρα στο Ινστιτούτο Επεξεργασίας του Λόγου (ΙΕΛ) κατά το παρελθόν. Πιο συγκεκριμένα στο (Tambouratzis G., Markantonatou S., Hairetakis N. , Vassiliou M., Tambouratzis D. & Carayannis G. 2000) έχει αναφερθεί

ένα σύνολο από χαρακτηριστικά, βασισμένα σε γλωσσολογικές μελέτες και παραδοχές, τα οποία σε συνδυασμό με τη διακριτική ανάλυση είναι σε θέση να διαχωρίζουν επιτυχώς 5 ομιλητές της Ελληνικής Βουλής. Πρόσθετα πειράματα τα οποία σχετίζονταν με το διαχωρισμό ειδών λόγου (π.χ. πολιτικός, ιστορικός, πεζογραφία), αλλά και με την αναγνώριση ομιλητών (Tambouratzis G., Hairetakis N., Markantonatou S. & Carayannis G. 2003 και Tambouratzis G. 2006), χρησιμοποιήθηκε το μοντέλο νευρωνικών δικτύων αυτοοργανούμενου χάρτη (Self-Organising Maps – SOM (Kohonen T., 1982)).

Στα πλαίσια της παρούσας εργασίας υλοποιήθηκε ένα σύστημα αντίστοιχο με προηγούμενα συστήματα διαχωρισμού ύφους, το οποίο αναλύει κείμενα γραμμένα στην ελληνική γλώσσα. Το σύστημα αυτό υιοθετεί το σύνολο χαρακτηριστικών που είχε αναπτυχθεί στο ΙΕΛ (Tambouratzis G., Markantonatou S., Hairetakis N. , Vassiliou M., Tambouratzis D. & Carayannis G. 2000), αλλά στοχεύει στο να βελτιώσει την ακρίβεια και να παράσχει μία επιπρόσθετη λειτουργικότητα αξιολόγησης, με αυτόματο τρόπο, της συμβολής των χαρακτηριστικών στο τελικό αποτέλεσμα. Απώτερο στόχο συνιστά η βελτίωση του συστήματος σε αποτελεσματικότητα και απόδοση ως προς τις υπολογιστικές του απαιτήσεις.

2.3 Οργάνωση κειμένων με βάση το περιεχόμενο

Όπως προαναφέρθηκε, σε πολλές πρακτικές εφαρμογές εμφανίζεται συχνά η ανάγκη της διατήρησης ενός μεγάλου όγκου κειμένων οργανωμένων με βάση όχι μόνο το ύφος, αλλά και το περιεχόμενό τους. Ιδιαίτερα στο χώρο του διαδικτύου οι αναζητήσεις αφορούν σχεδόν αποκλειστικά το περιεχόμενο. Οι αντίστοιχες πρακτικές εφαρμογές ποικίλλουν, σε μεγάλο βαθμό καθοριζόμενες από τις ανάγκες του τελικού χρήστη. Για παράδειγμα, σε ορισμένες εφαρμογές είναι επιθυμητό να εμφανίζονται στον τελικό χρήστη κείμενα σχετικά με αυτά που έχει ήδη επιλέξει, ενώ σε άλλες περιπτώσεις είναι επιθυμητή η πλοήγηση με βάση μία υπάρχουσα δομή, συνήθως δενδρικής μορφής, αποτελούμενη από διαδρομές που χαρακτηρίζονται με λέξεις-κλειδιά. Μία τέτοια αναζήτηση γίνεται σε φυσική γλώσσα, ενώ μία διαφορετική προσέγγιση είναι η τοποθέτηση των κειμένων σε ένα «θεματικό» χάρτη, ο οποίος λειτουργεί ως βάση αναζήτησης κειμένων. Ειδικά η τελευταία προσέγγιση είναι αρκετά συνηθισμένη σε εφαρμογές ταξινόμησης κειμένων με τεχνητά νευρωνικά δίκτυα αυτοοργανούμενου χάρτη (SOM (Kohonen T., Kaski S., Lagus K., Salojarvi, Honkela J., Patero V. & Saarela A. 2000 και Lagus K., Kaski S. & Kohonen T. 2004)).

Οι κυριότερες εφαρμογές αυτόματης επεξεργασίας κειμένων μπορούν να διακριθούν σε 3 κατηγορίες (Freeman R.T. & Yin. H. 2005):

1. **Αναγνώριση του θέματος ενός κειμένου** (topic detection): προσδιορισμός της θεματικής κατηγορίας διαφόρων τμημάτων του κειμένου.
2. **Ταξινόμηση κειμένων με επίβλεψη** (classification): ταξινόμηση των κειμένων με βάση κάποιο κριτήριο ομοιότητας καθορισμένο από το χρήστη. Συνεπώς, ο χρήστης καθορίζει απόλυτα το είδος της αναζήτησης και τον τρόπο χρήσης της εφαρμογής.
3. **Ομαδοποίηση κειμένων χωρίς επίβλεψη** (clustering): ομαδοποίηση των κειμένων σε σύνολα με βάση τις μεταξύ τους ομοιότητες. Το αποτέλεσμα της ομαδοποίησης δεν είναι γνωστό εκ των προτέρων, ενώ η προκύπτουσα οργάνωση είναι περισσότερο ελεύθερη και γενικά όχι άμεσα συμβατή με τον ανθρώπινο τρόπο σκέψης.

Τρόποι οργάνωσης κειμένων όπως οι προαναφερθέντες εξαρτώνται σημαντικά από το είδος των κειμένων προς επεξεργασία. Είναι σαφέστατα ευκολότερο να διαχωριστούν κείμενα που εμπίπτουν σε εντελώς διαφορετικές κατηγορίες (π.χ. αθλητικά και επιστημονικά κείμενα) από ό,τι κείμενα που ανήκουν στην ίδια ευρύτερη ομάδα (π.χ. ιστορικά κείμενα από διάφορες εποχές).

Τα εν λόγω συστήματα συχνά αναπαριστούν τα κείμενα απεικονίζοντας τα σε ένα σύνολο μετρήσιμων χαρακτηριστικών. Σε αντίθεση με την υφολογική ανάλυση, όπου χρησιμοποιούνται χαρακτηριστικά προερχόμενα από γλωσσολογικές προσεγγίσεις, στα συστήματα οργάνωσης με βάση το περιεχόμενο τα κείμενα αντιμετωπίζονται ως ένα σύνολο λέξεων (bag of words) (Manning C. D. & Schütze H. 1999), δεδομένης της άποψης ότι οι λέξεις είναι οι κύριοι φορείς του νοήματος του κειμένου. Αυτή η φαινομενικά απλοποιημένη προσέγγιση, όπου σε κάθε κείμενο αντιστοιχίζονται διανύσματα που περιγράφουν το σύνολο (ή μέρος) των λέξεων που το αποτελούν, έχει συχνά αυξημένες υπολογιστικές απαιτήσεις, ιδιαίτερα σε περιπτώσεις όπου το πλήθος των κειμένων είναι πολύ μεγάλο (της τάξεως των μερικών χιλιάδων) και περιλαμβάνει μεγάλο πλήθος διαφορετικών λέξεων. Για το λόγο αυτό, ανάλογα με το κάθε σύστημα, η επιλογή της αναπαράστασης καθώς και η ανάπτυξη του αντίστοιχου συστήματος διαφέρουν σημαντικά. Στις περισσότερες περιπτώσεις αναπτύσσονται συστήματα που χρησιμοποιούν χαρακτηριστικά από μετρήσεις βασισμένες στη συχνότητα εμφάνισης των λέξεων των κειμένων (Kohonen T., Kaski S., Lagus K., Salojarvi, Honkela J., Patero V. & Saarela A. 2000, Lagus K., Kaski S. & Kohonen T. 2004 και Freeman R.T. & Yin. H. 2005).

Για τον περιορισμό του πλήθους των χαρακτηριστικών πολλές διαφορετικές προσεγγίσεις έχουν προταθεί στη βιβλιογραφία, με πιο γνωστή την τυχαία προβολή-μείωση διάστασης (Kaski S., 1998), η οποία χρησιμοποιήθηκε στο WEBSOM (Kohonen T., Kaski S., Lagus K., Salojarvi, Honkela J., Patero V. & Saarela A. 2000 και Lagus K., Kaski S. & Kohonen T. 2004). Με την τεχνική αυτή προτιμάται ένας τυχαίος πίνακας προβολής των δεδομένων σε μικρότερες διαστάσεις έναντι της ανάλυσης σε πρωτεύουσες συνιστώσες (Principal Components Analysis – PCA ή LSI –Latent Semantic Indexing για την περίπτωση των κειμένων (Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. & Hashman R. 1990)), η οποία είναι μία υπολογιστικά απαιτητικότερη διαδικασία. Μία διαφορετική προσέγγιση συνιστά η κατασκευή ενδιάμεσου χάρτη λέξεων, πάνω στον οποίο προβάλλονται τα κείμενα. Τέτοιο είναι το σύστημα των (Georgakis A., Kotropoulos C., Xafopoulos A., & Pitas I., 2004), το οποίο χρησιμοποιεί την τυχαία προβολή για τη δημιουργία του χάρτη λέξεων (Kaski S., 1998). Συχνότητες λέξεων, αλλά σε επίπεδο προτάσεων και φράσεων, χρησιμοποιήθηκαν από τον (Pullwitt D. 2002) με σκοπό τη οργάνωση των προτάσεων σε έναν πρώτο χάρτη τύπου SOM, ενώ στο επόμενο επίπεδο, και αυτό ένας χάρτης SOM, οργανώνονται τα κείμενα. Εξάλλου στο σύστημα που περιγράφεται στο (Linden K. 2004) χρησιμοποιήθηκαν πρόσθετα γλωσσολογικά χαρακτηριστικά, όπως τα μέρη του λόγου, αλλά και μετρήσεις σε συντακτικό επίπεδο.

Στις περισσότερες μελέτες επιλέγονται δίκτυα τύπου SOM (τα οποία θα παρουσιασθούν αναλυτικότερα στο αντίστοιχο κεφάλαιο). Η προτίμηση σε αυτά τα δίκτυα οφείλεται στην αποτελεσματικότητά τους σε προβλήματα πολλών διαστάσεων, όπως η οργάνωση κειμένων αναπαριστώμενων με διανύσματα λέξεων. Η μεγάλη διάσταση των διανυσμάτων χαρακτηριστικών είναι αναμενόμενη, αφού σε κάθε σώμα κειμένων απαντάται ένα πολύ μεγάλο πλήθος διαφορετικών λέξεων. Χαρακτηριστικές εργασίες που ασχολήθηκαν αποκλειστικά με την οργάνωση λέξεων χρησιμοποιώντας το μοντέλο SOM είναι οι (Kohonen T. & Somervuo P., 1998 και Kohonen T., 1985).

Από τα παραπάνω γίνεται σαφές ότι το πρόβλημα οργάνωσης κειμένων ποικίλλει αρκετά ανάλογα με το είδος της ομαδοποίησης: με βάση το ύφος ή το περιεχόμενο. Πέραν αυτής της βασικής διαφοροποίησης, επιμέρους ζητήματα προκύπτουν ανάλογα με το είδος των δεδομένων, αλλά και τις απαιτήσεις της εφαρμογής. Η επίλυση του προβλήματος παρουσιάζει εξαιρετικό ενδιαφέρον ιδίως μέσα από το πρίσμα των σύγχρονων απαιτήσεων για οργάνωση της υπάρχουσας πληροφορίας, ώστε να υποστηριχθεί η κατά το δυνατό αποτελεσματικότερη πρόσβαση και αναζήτηση σε αυτή.

ΚΕΦΑΛΑΙΟ 3ο

3.1 Σύστημα οργάνωσης κειμένων με βάση το συγγραφέα

Το σύστημα οργάνωσης κειμένων με βάση το συγγραφέα αποσκοπεί στην αυτόματη οργάνωση κειμένων ανάλογα με την προέλευσή τους, η οποία καθορίζεται βάσει του γλωσσικού ύφους κάθε συγγραφέα. Η επιλογή των κατάλληλων γλωσσικών χαρακτηριστικών μπορεί να φανερώσει τον ιδιαίτερο τρόπο με τον οποίο ένας συγγραφέας χρησιμοποιεί την γλώσσα για να εκφραστεί (Matthews R. & Merriam T. 1993, Merriam T. & Matthews R. 1994 και Lowe D. & Matthews R. 1995).

Συγκεκριμένα, η υφολογική ανάλυση περιλαμβάνει τη μέτρηση ενός συνόλου χαρακτηριστικών, τα οποία, ενδεικτικά, μπορεί να αφορούν σε α) δομικό επίπεδο (π.χ. πλήθος λέξεων ή προτάσεων, ιστογράμματα του μήκους των λέξεων ή του μήκους των προτάσεων, πλήθος των σημείων στίξης κ. ά.), β) λεξικό επίπεδο (π.χ. πλήθος λέξεων με μοναδική εμφάνιση μέσα στο κείμενο, συχνότητα εμφάνισης προεπιλεγμένων λέξεων κ. ά.) και (γ) συντακτικό επίπεδο (π.χ. μέτρηση γραμματικών χαρακτηριστικών ή μερών του λόγου, εμφανίσεις προκαθορισμένων συντακτικών δομών κ. ά.). Το σύστημα οργάνωσης κειμένων ως προς το συγγραφέα βασίζεται σε ένα σύνολο γλωσσικών χαρακτηριστικών σε συνδυασμό με ένα αλγόριθμο κατηγοριοποίησης. Η οργάνωση στηρίζεται στην προσπάθεια αναγνώρισης του ύφους των κειμένων, όπως αυτό μπορεί να μετρηθεί με αυτόματο τρόπο, βάσει αυτού του συνόλου χαρακτηριστικών.

Ειδικότερα για την ελληνική γλώσσα, η οποία παρουσιάζει εξαιρετικά πλούσια μορφολογία, προτού εξαχθούν τα κατάλληλα γλωσσικά χαρακτηριστικά, καθίσταται αναγκαία η προεπεξεργασία των κειμένων προς ανάλυση. Για το σκοπό αυτό χρησιμοποιήθηκε το εργαλείο γραμματικής επισημείωσης και λημματοποίησης (tagger-lemmatiser) του ΙΕΛ (Papageorgiou H., Prokopidis P., Giouli V. & Piperidis S. 2000). Το εργαλείο αυτό, βασιζόμενο σε ένα πλήρες μορφολογικό λεξικό, χαρακτηρίζει τις λέξεις γραμματικά και αποδίδει το λήμμα τους. Παραδείγματος χάριν, όταν ο λημματοποιητής απαντά τους τύπους «παιδιά», «παιδιών» κ.λπ., τους αντικαθιστά με το λήμμα «παιδί»· ομοίως, διαφορετικοί ρηματικοί τύποι όπως «μιλούσε», «μίλαγε» κ.λπ. αντικαθίστανται από το λήμμα «μιλώ». Συνάγεται, επομένως, ότι η προεπεξεργασία διευκολύνει σημαντικά τη διαδικασία εξα-

γωγής χαρακτηριστικών, καθώς οδηγεί στη μείωση του πλήθους των διαφορετικών λεξικών μορφών στα κείμενα.

Για την υλοποίηση του συστήματος αναπτύχθηκε αρχικά ένα εργαλείο, το οποίο παρουσιάζει με εποπτικό τρόπο ένα σύνολο χαρακτηριστικών ανά κατηγορία, ώστε ο σχεδιαστής του συστήματος να είναι σε θέση να κάνει κάποιες αρχικές επιλογές. Στη συνέχεια το σύστημα μετράει το σύνολο των επιλεγμένων χαρακτηριστικών και αποθηκεύει την πληροφορία σε ένα αρχείο. Το σύνολο των χαρακτηριστικών είναι δυναμικό και μπορεί να καθορίζεται με τη βοήθεια μίας αρκετά σύνθετης διεπαφής, μέσω της οποίας είναι δυνατό να ορίζονται και νέα χαρακτηριστικά για κάθε κατηγορία. Το σύστημα, υλοποιημένο σε γλώσσα προγραμματισμού C++, χειρίζεται μετρήσεις από τις ακόλουθες κατηγορίες:

- **Συχνότητες λημμάτων** (lemma frequency): Στην κατηγορία αυτή ανήκουν χαρακτηριστικά που αναφέρονται στον αριθμό των εμφανίσεων όλων των λημμάτων κάθε κειμένου. Ο αριθμός των εμφανίσεων είναι κανονικοποιημένος ως προς το συνολικό αριθμό των λέξεων του κειμένου. Από την κατηγορία αυτή επιλέχθηκε αλγοριθμικά η μέτρηση της συχνότητας εμφάνισης 17 συγκεκριμένων λημμάτων.
- **Γραμματικά Χαρακτηριστικά** (grammatical features): Στην κατηγορία αυτή υπάγονται χαρακτηριστικά που μετρούν τον αριθμό των εμφανίσεων σχεδόν όλων των μερών του λόγου καθώς και των ποικίλων κλιτών μορφών στο κείμενο. Για παράδειγμα, μετράται τόσο το σύνολο των άρθρων στο κείμενο όσο και ο αριθμός των διαφορετικών τύπων του άρθρου ανά πτώση (ονομαστική, γενική ή αιτιατική), ανά αριθμό (ενικό ή πληθυντικό) και ανά γένος (αρσενικό, θηλυκό και ουδέτερο). Οι μετρήσεις έχουν κανονικοποιηθεί ως προς τον αριθμό των λέξεων του κειμένου, προκειμένου οι μετρήσεις μεταξύ διαφορετικών κειμένων να είναι συγκρίσιμες, αφού το μέγεθος των κειμένων διαφέρει. Από την προκείμενη κατηγορία μετρήθηκε η συχνότητα εμφάνισης λέξεων αντιστοιχισμένων σε 19 γραμματικούς χαρακτηρισμούς.
- **Δομικά Χαρακτηριστικά** (structural features): Στην κατηγορία αυτή υπάγονται διάφορα χαρακτηριστικά όπως ο αριθμός των λέξεων με συγκεκριμένο μήκος (το οποίο κυμαίνεται από 1 έως 30 γράμματα) κανονικοποιημένος ως προς το συνολικό αριθμό λέξεων του κειμένου, ο αριθμός των προτάσεων με συγκεκριμένο μήκος (το οποίο κυμαίνεται από 1 έως 160 λέξεις) κανονικοποιημένος ως προς το συνολικό αριθμό των προτάσεων του κειμένου και ο αριθμός ορισμένων σημείων στίξης (τελεία, ερωτηματικό, άνω και κάτω στιγμή, κόμμα, παύλα, παρένθεση), χρονολογιών

και συντομογραφιών κανονικοποιημένων ως προς τον αριθμό λέξεων. Από αυτήν την κατηγορία επιλέχθηκαν 27 χαρακτηριστικά.

- **Καταλήξεις Καθαρεύουσας και Δημοτικής** (inflectional endings of Katharevousa & Demotiki): Επειδή είναι δυνατόν από ορισμένους ομιλητές να προτιμάται η χρήση ορισμένων καταλήξεων της Καθαρεύουσας σε σχέση με αυτές της Δημοτικής ή και το αντίστροφο, έγιναν μετρήσεις της εμφάνισης ορισμένων ρηματικών καταλήξεων κανονικοποιημένων ως προς το συνολικό αριθμό των λέξεων του κειμένου. Από την κατηγορία αυτή προήλθαν 14 μετρήσεις, ως συνδυασμοί των τριών προσώπων και των δύο αριθμών (ενικός και πληθυντικός), και 2 συνολικές μετρήσεις Δημοτικής και Καθαρεύουσας.
- **Χρήση λέξεων που εκφράζουν άρνηση** (negation words): Στην κατηγορία αυτή μετρήθηκε η συχνότητα εμφάνισης οκτώ λέξεων που δηλώνουν άρνηση.

Το στάδιο της επιλογής χαρακτηριστικών (feature selection) δίνει ως έξοδο ένα σύνολο χαρακτηριστικών, το οποίο εμπλουτίσθηκε με κάποιες επιπλέον μεταβλητές με βάση παλαιότερες εργασίες που έχουν γίνει πάνω στο θέμα αυτό από το ΙΕΛ (Tambouratzis G., Markantonatou S., Hairetakis N. , Vassiliou M., Tambouratzis D. & Carayannis G. 2000). Οι μεταβλητές που επιλέχθηκαν αναγράφονται στον πίνακα 3.1.

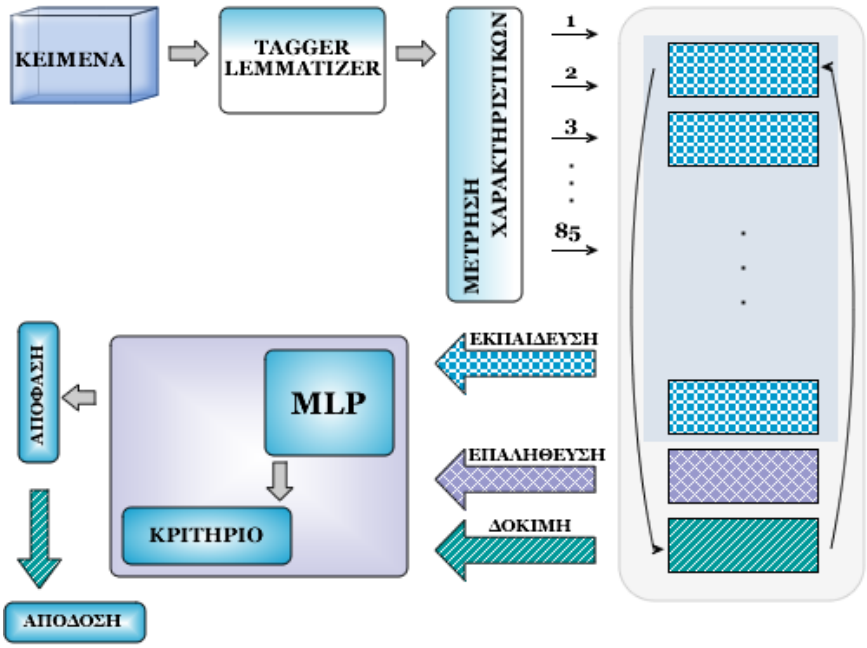
Αφού καθορισθεί το σύνολο των χαρακτηριστικών, στη συνέχεια, μέσω της διαδικασίας εξαγωγής των χαρακτηριστικών, τα κείμενα απεικονίζονται στο χώρο προτύπων ως ένα διάνυσμα με αριθμητικές συνιστώσες. Κάθε κείμενο αντιστοιχίζεται στις μετρήσεις των χαρακτηριστικών του και βάσει αυτών αποδίδεται σε κάποια ομάδα.

Ως **κατηγοριοποιητής**, δηλαδή το τμήμα του συστήματος που αποδίδει τα κείμενα σε κατηγορίες, επιλέχθηκε ένα νευρωνικό δίκτυο τύπου MLP, το οποίο εκπαιδεύτηκε με τον αλγόριθμο ανάστροφης διάδοσης Levenberg – Marquardt, που περιλαμβάνει και όρο μείωσης της τιμής των βαρών του δικτύου. Ο συγκεκριμένος κατηγοριοποιητής λειτουργεί με επίβλεψη και ως εκ τούτου, για να προσαρμοσθεί στα δεδομένα του προβλήματος, χρειάζεται ένα σύνολο δεδομένων τα οποία να έχουν χαρακτηρισθεί χειρωνακτικά ως προς τις ομάδες που ανήκουν. Μετά το πέρας της εκπαίδευσης ο κατηγοριοποιητής είναι ικανός να αποφασίζει για τις ομάδες στις οποίες ανήκουν άγνωστα σε αυτόν δεδομένα.

1	Γράμματα ανά Λέξη	30	Ρήματα	59	Προτάσεις με 6-10 Λέξεις
2	Λέξεις ανά πρόταση	31	Γενική Επιθέτων	60	Προτάσεις με 11-15 Λέξεις
3	Κόμματα ανά πρόταση	32	Γενική Ουσιαστικών	61	Προτάσεις με 16-20 Λέξεις
4	Παρενθέσεις	33	1ο Ενικό Ρημάτων	62	Προτάσεις με 21-25 Λέξεις
5	Παύλες	34	2ο Ενικό Ρημάτων	63	Προτάσεις με 26-30 Λέξεις
6	1ο Ενικό Καθαρεύουσα	35	3ο Ενικό Ρημάτων	64	Προτάσεις με 31-40 Λέξεις
7	2ο Ενικό Καθαρεύουσα	36	1ο Πληθυντικό Ρημάτων	65	Προτάσεις με 41-50 Λέξεις
8	3ο Ενικό Καθαρεύουσα	37	2ο Πληθυντικό Ρημάτων	66	Προτάσεις με 51-75 Λέξεις
9	1ο Πληθυντικό Καθαρεύουσα	38	3ο Πληθυντικό Ρημάτων	67	Προτάσεις με 76-100 Λέξεις
10	2ο Πληθυντικό Καθαρεύουσα	39	Συχνότητα του «όχι»	68	Προτάσεις με 101-150 Λέξεις
11	3ο Πληθυντικό Καθαρεύουσα	40	Συχνότητα του «ουδείς»	69	Λήμμα «αλλά»
12	1ο Ενικό Δημοτική	41	Συχνότητα του «ουδέποτε»	70	Λήμμα «αυτός»
13	2ο Ενικό Δημοτική	42	Συχνότητα του «ουδαμού»	71	Λήμμα «εγώ»
14	3ο Ενικό Δημοτική	43	Συχνότητα του «άνευ»	72	Λήμμα «Ελλάδα»
15	1ο Πληθυντικό Δημοτική	44	Συχνότητα του «δεν»	73	Λήμμα «Έλληνας»
16	2ο Πληθυντικό Δημοτική	45	Συχνότητα του «μην»	74	Λήμμα «ένας»
17	3ο Πληθυντικό Δημοτική	46	Σημεία Στίξης	75	Λήμμα «κάνω»
18	Σύνολό Καθαρεύουσας	47	Ερωτηματικά	76	Λήμμα «κυβέρνηση»
19	Σύνολο Δημοτικής	48	Συχνότητα του «οποίος – που»	77	Λήμμα «κύριος»
20	Επίθετα	49	A	78	Λήμμα «λαός»
21	Προθέσεις	50	B	79	Λήμμα «λέγω»
22	Επιρρήματα	51	ουδ*********	80	Λήμμα «μου»
23	Άρθρα	52	Λέξεις με 1-3 Γράμματα	81	Λήμμα «μπορώ»
24	Σύνδεσμοι	53	Λέξεις με 4-8 Γράμματα	82	Λήμμα «Πρωθυπουργός»
25	Ουσιαστικά	54	Λέξεις με 9-15 Γράμματα	83	Λήμμα «συνεννόηση»
26	Αριθμητικά	55	Λέξεις με 16-20 Γράμματα	84	Λήμμα «θέλω»
27	Μόρια	56	Λέξεις με 21-30 Γράμματα	85	Λήμμα «υπάρχω»
28	Αντωνυμίες	57	Προτάσεις με 1 Λέξη		
29	Υπόλοιπα	58	Προτάσεις με 2-5 Λέξεις		

Πίνακας 3.1. Γλωσσικά χαρακτηριστικά

Μία συνοπτική εικόνα του συστήματος απεικονίζεται στο σχήμα 3.1, όπου παρουσιάζονται το αρχικό στάδιο επεξεργασίας με το λημματοποιητή-γραμματικό επισημειωτή του ΙΕΛ, στη συνέχεια το στάδιο μέτρησης των χαρακτηριστικών και απεικόνισης των κειμένων στο χώρο των προτύπων, αλλά και ο κατηγοριοποιητής που περιλαμβάνει το νευρωνικό δίκτυο MLP σε συνδυασμό με το κριτήριο λήψης απόφασης. Επισημαίνεται ότι χρησιμοποιούνται διαφορετικά σύνολα δεδομένων για την προσαρμογή-εκπαίδευση του κατηγοριοποιητή MLP και για την εκτίμηση της απόδοσης του συστήματος. Προκειμένου να αποτιμηθεί η επάρκεια του συστήματος, διενεργήθηκε ένα σύνολο πειραμάτων με στόχο να εκτιμηθεί η ακρίβεια κατηγοριοποίησης του συστήματος, αλλά και να συγκριθεί αυτό με άλλα υπάρχοντα συστήματα και τεχνικές που χρησιμοποιούνται ευρέως σε τέτοιου είδους προβλήματα. Τα πειραματικά αποτελέσματα δίνονται στις ενότητες 3.4 και 3.5.



Σχήμα 3.1. Διάγραμμα περιγραφής συστήματος οργάνωσης κειμένων με βάση το συγγραφέα

3.2 Αλγόριθμοι κατηγοριοποίησης συστήματος

Στον τομέα της αναγνώρισης προτύπων υπάρχει ένας μεγάλος αριθμός από τεχνικές, ικανές να διαχωρίσουν ένα πλήθος από αριθμητικά δεδομένα. Τέτοιες τεχνικές έχουν εφαρμοσθεί κατά καιρούς σε διάφορα είδη προβλημάτων όπως η αναγνώριση εικόνας, φωνής, χαρακτήρων κ.λπ. Κύριο γνώρισμα αυτών των τεχνικών είναι ότι δέχονται ως είσοδο για κάθε αντικείμενο ένα διάνυσμα – σύνολο αριθμητικών, συνήθως, χαρακτηριστικών και δίνουν ως έξοδο την απόφασή τους για το σύνολο στο οποίο ανήκει το αντικείμενο. Παρόμοιες τεχνικές, όπως αυτές που αναφέρθηκαν στην προηγούμενη ενότητα, είναι δυνατό να χρησιμοποιηθούν και στο πρόβλημα της οργάνωσης κειμένων.

Οι **αλγόριθμοι κατηγοριοποίησης ή κατηγοριοποιητές** (classifiers) είναι μέθοδοι αναγνώρισης προτύπων που έχουν την ικανότητα να ταξινομούν σε κατηγορίες δεδομένα εκφρασμένα σε διανυσματική μορφή. Κύρια χαρακτηριστικά τους είναι η μάθηση με επίβλεψη (supervised learning) καθώς και η ικανότητά τους να εφαρμόζονται σε δεδομένα πολλών διαστάσεων. Οι κατηγοριοποιητές μπορεί να βασίζονται και σε κανόνες (rule-based), όμως ανεξαρτήτως του είδους τους απαιτούν τη χρήση διαθέσιμων δεδομένων εκπαίδευσης. Τα δεδομένα εκπαίδευσης αποτελούν ζεύγη από μετρήσεις και επιθυμητά αποτελέσματα, στα οποία προσαρμόζονται τα συστήματα. Η συλλογή δεδομένων εκπαίδευσης είναι κοπιαστική εργασία και έχει μεγάλο κόστος. Τα πολυστρωματικά νευρωνικά δίκτυα τύπου MLP (multilayer perceptron) που εφαρμόζονται σε συστήματα οργάνωσης κειμένων ανήκουν στην οικογένεια των κατηγοριοποιητών. Αρκετά αποδοτικές σε προβλήματα κατηγοριοποίησης είναι και οι **Μηχανές Διανύσματος Υποστήριξης** (Support Vector Machines - SVM), οι οποίες λειτουργούν διαφορετικά από τα πολυστρωματικά νευρωνικά δίκτυα και παρουσιάζονται στην ενότητα 3.2.5.

3.2.1 Δομή πολυστρωματικών δικτύων (Multi-Layer Perceptron)

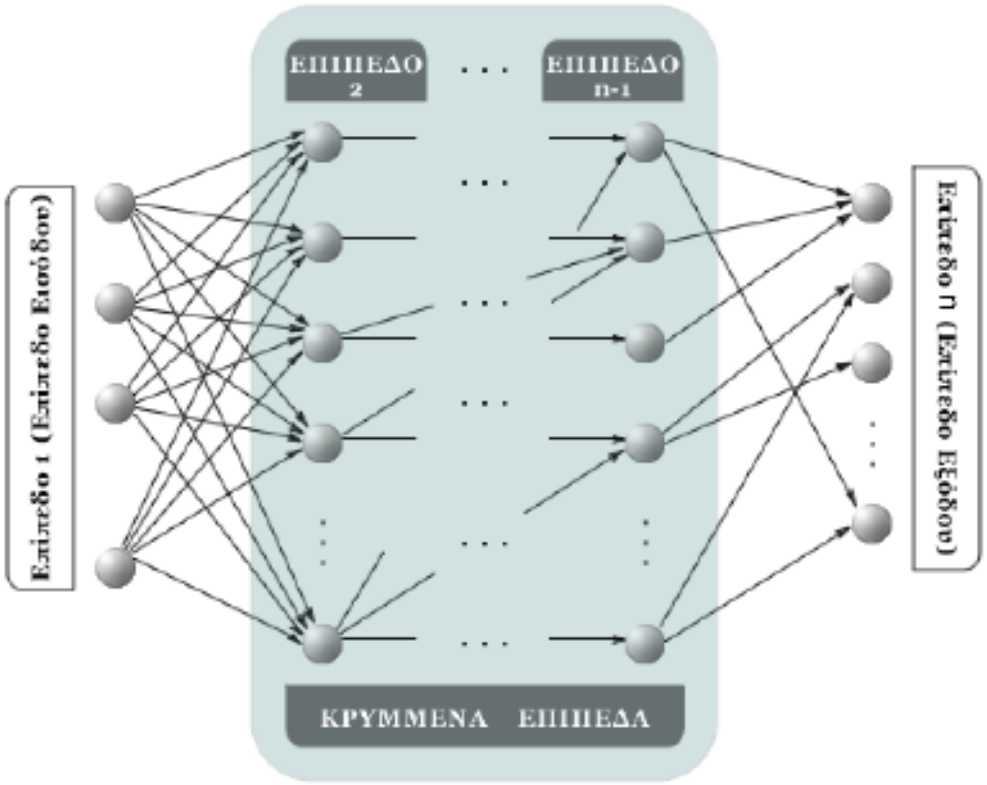
Η δομή ενός πολυστρωματικού δικτύου είναι η ακόλουθη (Haykin S. 1999):

1) Το επίπεδο **εισόδου**, όπου και εφαρμόζονται οι είσοδοι στο σύστημα.

2) Τα **ενδιάμεσα κρυφά** επίπεδα.

3) Το επίπεδο **εξόδου**.

Οι νευρώνες κάθε επιπέδου συνδέονται με τους νευρώνες των επόμενων επιπέδων με τα βάρη w^m_{ji} , όπου w^m_{ji} είναι το βάρος που συνδέει το νευρώνα *j* του *m* επιπέδου με το νευρώνα *i* του *m+1* επιπέδου. Στα δίκτυα αυτά, κάθε νευρώνας συνδέεται μόνο με νευρώνες επόμενων επιπέδων, συνεπώς ανήκουν στην κατηγορία των δικτύων προσωτροφοδότησης (feed-forward networks). Γενικά, συνηθίζεται κάθε νευρώνας να συνδέεται με όλους ή κάποιους από τους νευρώνες μόνο του αμέσως επόμενου επιπέδου, οπότε και το νευρωνικό δίκτυο ονομάζεται **δομημένο** και έχει τη μορφή του ακόλουθου σχήματος (σχήμα 3.2).



Σχήμα 3.2. Αρχιτεκτονική δικτύου MLP με *n* επίπεδα

Η έξοδος κάθε νευρώνα είναι ίση με το άθροισμα του γινομένου των εισόδων του επί τα αντίστοιχα βάρη τους. Το άθροισμα αυτό αυξάνεται κατά μία σταθερά, δηλαδή ένα κατώφλι (bias), και μετασχηματίζεται από μία συνάρτηση ενεργοποίησης *f*. Η έξοδος ενός νευρώνα εφαρμόζεται ως είσοδος στους νευρώνες των επόμενων επιπέδων με τους οποίους συνδέεται μέχρι το επίπεδο εξόδου. Όπως αναφέρθηκε, το βάρος που συνδέει το νευρώνα *j* του *m* επιπέδου με το νευρώνα *i* του *m+1*, συμβολίζεται με w_{ji}^m, ενώ με b_i^m συμβολίζεται το κατώφλι του νευρώνα και με y_i^m η έξοδός του. Συνεπώς, η αντίστοιχη έξοδος του *i* νευρώνα στο επίπεδο *m* είναι ίση με:

$$y_i^m = \begin{cases} f((\sum_j w_{ji}^{m-1} y_i^{m-1}) + b_i^m) & m \geq 2 \\ x_i & m = 1 \end{cases} \quad (3\text{-}1)$$

όπου *f* η συνάρτηση ενεργοποίησης και x_i οι *i* είσοδοι του δικτύου. Παρατηρούμε ότι για το πρώτο επίπεδο, δηλαδή για *m=1*, η έξοδος του κάθε νευρώνα είναι ίδια με την είσοδο.

Γενικά, είναι επιθυμητό η συνάρτηση ενεργοποίησης *f* να είναι συνεχής και παραγωγίσιμη, κάτι που απλοποιεί σε μεγάλο βαθμό τους αλγόριθμους εκπαίδευσης. Τις περισσότερες φορές η συγκεκριμένη συνάρτηση προσεγγίζει τη διπολική διακριτική συνάρτηση, χωρίς φυσικά να αποκλείονται και άλλες μορφές ανάλογα με τις ανάγκες του προβλήματος. Οι πιο συχνά χρησιμοποιούμενες συναρτήσεις είναι οι ακόλουθες:

- Η **λογιστική σιγμοειδής** συνάρτηση,

$$log\,sig(x) = \frac{1}{1+e^{-ax}} \quad (3\text{-}2)$$

όπου *a* είναι μία παράμετρος που ορίζει τη μορφή και πιο συγκεκριμένα την κλίση της καμπύλης της συνάρτησης (όσο μεγαλύτερη είναι η τιμή του *a* τόσο πιο απότομη είναι η καμπύλη).

- Η **υπερβολική εφαπτομένη**,

$$tan\,sig(x) = \frac{1-e^{-ax}}{1+e^{-ax}} \quad (3\text{-}3)$$

όπου *a* είναι η παράμετρος που καθορίζει τη μορφή της καμπύλης.

- Η **απλή γραμμική** συνάρτηση,

$$f(x) = ax \quad (3\text{-}4)$$

η οποία ορίζεται για οποιοδήποτε εύρος του *x* και χρησιμοποιείται κυρίως για προβλήματα προσέγγισης συνάρτησης (regression).

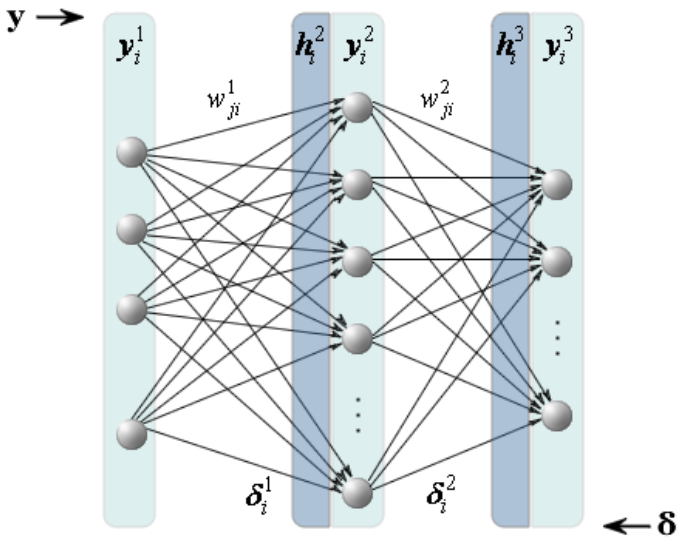
Στα περισσότερα προβλήματα αναγνώρισης προτύπων απαιτούνται δύο καταστάσεις για τις εξόδους (π.χ. κάτι ανήκει ή δεν ανήκει σε μία κατηγορία), οπότε χρησιμοποιούνται διπολικές συναρτήσεις, σιγμοειδούς συνήθως μορφής, που περιορίζουν το εύρος των τιμών στο [-1,1] ή [0,1]. Όταν όμως επιδιώκεται η προσέγγιση μίας συνάρτησης, της οποίας οι τιμές δεν περιορίζονται σε κάποιο εύρος, τότε δεν είναι αποδοτική η χρήση διπολικών συναρτήσεων και προτιμώνται απλές γραμμικές συναρτήσεις στην έξοδο, οι οποίες δεν έχουν περιορισμένο πεδίο τιμών.

Στα περισσότερα πολυστρωματικά νευρωνικά δίκτυα συνηθίζεται η χρήση ενός μόνο κρυφού επιπέδου, καθώς έχει αποδειχθεί ότι ένα δίκτυο με μόνο ένα κρυφό επίπεδο έχει τις ίδιες δυνατότητες, σε ό,τι αφορά την απεικόνιση των δεδομένων, με ένα δίκτυο περισσότερων επιπέδων, βεβαίως με την προϋπόθεση ότι διαθέτει επαρκή αριθμό νευρώνων στο κρυφό επίπεδο. Είναι, δηλαδή, δυνατό, αυξάνοντας τον αριθμό των νευρώνων στο μοναδικό κρυφό επίπεδο, να επιτυγχάνεται η ίδια απόδοση με εκείνη ενός δικτύου με περισσότερα επίπεδα.

Πιο αναλυτικά, το θεώρημα Kolmogorov αναφέρει τα εξής: «Για κάθε *ε>0* και κάθε συνεχή συνάρτηση *F* υπάρχει ένα πολυστρωματικό νευρωνικό δίκτυο με ένα μοναδικό κρυφό στρώμα, τέτοιο ώστε να ισχύει για την έξοδο του δικτύου y: $\|F(x)-y(x)\|<\varepsilon$ για κάθε είσοδο *x*». Το θεώρημα αυτό εξασφαλίζει ότι ένα κρυφό επίπεδο είναι αρκετό για την προσέγγιση οποιασδήποτε συνάρτησης από το δίκτυο (Τζαφέστας Σ. 2002). Στα περισσότερα πολυστρωματικά νευρωνικά δίκτυα όλοι οι νευρώνες έχουν την ίδια συνάρτηση ενεργοποίησης *f*, πλην του επιπέδου εξόδου σε προβλήματα προσέγγισης συνάρτησης, όπου προτιμάται πάντα η γραμμική συνάρτηση (3-4).

3.2.2 Αλγόριθμος ανάστροφης διάδοσης MLP (Back Propagation)

Μία πολύ σημαντική ικανότητα των νευρωνικών δικτύων είναι η δυνατότητα να «μαθαίνουν», δηλαδή να προσαρμόζουν τα βάρη τους (με αλλαγή των τιμών τους) με τέτοιο τρόπο, ώστε να παράγουν περισσότερο «επιθυμητές» εξόδους. Σε ένα δίκτυο MLP παρέχεται, εκτός από την είσοδο, και η επιθυμητή έξοδος. Αυτός ο τρόπος εκπαίδευσης ονομάζεται εκπαίδευση με επίβλεψη (supervised). Για τα πολυστρωματικά νευρωνικά δίκτυα ο πιο διαδεδομένος αλγόριθμος εκπαίδευσης είναι ο αλγόριθμος ανάστροφης διάδοσης (Back Propagation) (βλ. σχήμα 3.3), που σκοπεύει στη μείωση της τιμής της συνάρτησης σφάλματος (Haykin S. 1999, Rumelhart D.E., Hinton† G.E. & Williams R.J. 1986a και Rumelhart D.E., Hinton† G.E. & Williams R.J. 1986b).



Σχήμα 3.3. Εφαρμογή αλγόριθμου ανάστροφης διάδοσης σε MLP με ένα κρυφό επίπεδο

Ο αλγόριθμος ανάστροφης διάδοσης λειτουργεί ως εξής: Αρχικά παρουσιάζεται κάποιο παράδειγμα στην είσοδο του δικτύου και υπολογίζεται η έξοδος του δικτύου, υπολογίζοντας πρώτα τις ενδιάμεσες αποκρίσεις του δικτύου, με τον τύπο (3-1) με φορά από την είσοδο προς την έξοδο, για *m=1, 2, 3*.

Από την έξοδο του δικτύου υπολογίζεται η τιμή του σφάλματος. Η συνάρτηση σφάλματος που χρησιμοποιείται συνήθως είναι το τετραγωνικό σφάλμα:

$$E_p(k) = \frac{1}{2}\sum_i (e_i(k))^2 \tag{3-5}$$

με

$$e_i(k) = d_i(k) - y_i^3(k) \tag{3-6}$$

όπου $d_i(k)$ είναι η επιθυμητή έξοδος για την έξοδο *i* του επιπέδου *3* για το διάνυσμα εισόδου *k* (οι έξοδοι του τρίτου επιπέδου είναι και αποτελούν τις εξόδους του δικτύου, αφού το νευρωνικό δίκτυο έχει μόνο ένα κρυφό επίπεδο).

Στη συνέχεια υπολογίζονται με φορά από την έξοδο προς την είσοδο (δηλαδή με ανάστροφη φορά) τα σήματα σφάλματος *δ*. Πιο αναλυτικά, με βάση τον κανόνα της κατάβασης στην επιφάνεια σφάλματος (Steepest Gradient Descent) προκύπτει ο τύπος (3-7) που παρέχει έναν τρόπο αλλαγής των βαρών, ώστε να μειωθεί το σφάλμα του δικτύου:

$$\Delta w_{ji}^{m} = -\gamma \frac{dE_p(k)}{dw_{ji}^{m}} \tag{3-7}$$

όπου *γ* είναι μια θετική παράμετρος μικρότερη της μονάδας, η οποία καθορίζει το ρυθμό μάθησης.

Συμβολίζοντας με h_i^m το άθροισμα των σημάτων διέγερσης της συνάρτησης ενεργοποίησης *f*, προκύπτει ότι:

$$h_i^m = \sum_j y_j^{m-1} \cdot w_{ji}^{m-1} + b_i^m \tag{3-8}$$

Η παράγωγος της επιφάνειας του σφάλματος ως προς τα βάρη, με χρήση του κανόνα της αλυσίδας, είναι:

$$\frac{dE_p(k)}{dw_{ji}^{m}} = \frac{dE_p(k)}{dy_i^{m+1}} \frac{dy_i^{m+1}}{dh_i^{m+1}} \frac{dh_i^{m+1}}{dw_{ji}^{m}} \tag{3-9}$$

Θέτοντας

$$\delta_i^m = -\frac{dE_p(k)}{dy_i^{m+1}} \frac{dy_i^{m+1}}{dh_i^{m+1}} \tag{3-10}$$

προκύπτει ότι

$$\frac{dE_p(k)}{dw_{ji}^{m}} = -\delta_i^m y_j^m \tag{3-11}$$

Συνδυάζοντας τις σχέσεις (3-7), (3-10) και (3-11) προκύπτει ο τύπος (3-12):

$$\Delta w_{ji}^{m} = \gamma \delta_i^m y_j^m \tag{3-12}$$

Από την παραπάνω σχέση φαίνεται ότι ο αλγόριθμος μάθησης ανάστροφης διάδοσης υπακούει στον κανόνα του Hebb (Haykin S. 1999), καθώς η αλλαγή στα βάρη εξαρτάται τόσο από την είσοδο (y_j^m) που δέχεται ο νευρώνας *j του επιπέδου m*, όσο και από τα σφάλματα δ_i^m που του αποδίδονται (Haykin S. 1999).

Αν χρησιμοποιηθεί η σιγμοειδής συνάρτηση ενεργοποίησης (3-2) με *a=1*, η παράγωγος είναι:

$$\frac{d\,log\,sig(x)}{dx} = \ldots = log\,sig(x)(1 - log\,sig(x)) \tag{3-13}$$

Ενώ, εάν χρησιμοποιηθεί η υπερβολική εφαπτομένη (3-3), η παράγωγος είναι:

$$\frac{d\,tan\,sig(x)}{dx} = \ldots = (1 - tan\,sig(x))^2 \tag{3-14}$$

Στη συνέχεια, υπολογίζονται αναλυτικά τα σήματα σφάλματος για κάθε ένα από τα τρία επίπεδα ενός τυπικού δικτύου MLP, προκειμένου να υπολογισθούν οι διορθώσεις που πρέπει να γίνουν στα βάρη του δικτύου για κάθε επίπεδο. Πιο συγκεκριμένα, για το επίπεδο εξόδου τα σήματα σφάλματος δίνονται από τον τύπο (3-15):

$$\delta_i^2 = -\frac{dE_p(k)}{dy_i^3}\frac{dy_i^3}{dh_i^3} = \ldots = e_i\,\frac{df(h_i^3)}{dh_i^3} \tag{3-15}$$

Κατά συνέπεια, τα βάρη ενημερώνονται με βάση την εξίσωση (3-12). Αντίστοιχα, τα κατώφλια b_i^m μπορεί και αυτά να θεωρηθούν ως βάρη με σταθερή είσοδο ίση με 1.

Προφανώς για το κρυφό επίπεδο δεν υπάρχει η γνώση της επιθυμητής εξόδου, οπότε αυτή θα πρέπει να εκτιμηθεί από τα σήματα σφάλματος *δ* των νευρώνων εξόδου, δηλαδή από το άθροισμα των σφαλμάτων όλων των νευρώνων, με τα οποία συνδέεται το κρυφό επίπεδο, ανάλογα με την τιμή του βάρους μέσω του οποίου γίνεται η σύνδεση. Πιο αναλυτικά, με εφαρμογή του κανόνα της αλυσίδας για τα βάρη εισόδου του κρυφού επιπέδου (για *k=1*) προκύπτει:

$$\delta_i^1 = -\frac{dE_p(k)}{dy_i^2}\frac{dy_i^2}{dh_i^2} = \ldots = -\frac{df(h_i^2)}{h_i^2}\sum_j \frac{dE_p(k)}{dy_j^3}\frac{dy_j^3}{dh_j^3}\frac{dh_j^3}{dy_i^2} \Rightarrow \delta_i^1 = -\frac{df(h_i^2)}{h_i^2}\sum_j w_{ji}^2\delta_j^2 \tag{3-16}$$

Η εκπαίδευση ενός πολυστρωματικού νευρωνικού δικτύου τύπου MLP συνιστά μία επαναληπτική διαδικασία (iterative process), κατά την οποία το δίκτυο τροφοδοτείται με ένα σύνολο δεδομένων, τα δεδομένα εκπαίδευσης, για κάθε ένα από τα οποία «επιβάλλονται» στο δίκτυο οι επιθυμητές έξοδοι. Κάθε παρουσίαση των δεδομένων εκπαίδευσης στο δίκτυο ονομάζεται εποχή (epoch). Πολλοί αλγόριθμοι εκπαίδευσης, αντί να τροποποιούν τα βάρη με κάθε παρουσιαζόμενο παράδειγμα, προβαίνουν σε συνολική αλλαγή των βαρών μετά το πέρας όλου του συνόλου εκπαίδευσης (batch training), δηλαδή στο τέλος κάθε εποχής.

Όταν τα βάρη ανανεώνονται με το πέρασμα κάθε παραδείγματος του συνόλου εκπαίδευσης, ο αλγόριθμος εκπαίδευσης δεν εγγυάται τη σύγκλιση, αφού η κλίση που υπολογίζεται για την ενημέρωση των βαρών αντιστοιχεί μόνο σε ένα παράδειγμα και όχι στο σύνολο εκπαίδευσης. Για να εξασφαλισθεί η σύγκλιση, ο αλγόριθμος θα πρέπει να βασίζεται σε όλο το σύνολο εκπαίδευσης. Ο βελτιωμένος αυτός αλγόριθμος ανανεώνει το σύνολο των βαρών μόνο μία φορά στο τέλος κάθε εποχής (batch training) με βάση τη μέση τιμή των σημάτων *δ* για κάθε δεδομένο εκπαίδευσης (Haykin S. 1999). Σημειωτέον, ωστόσο, ότι ο συγκεκριμένος αλγόριθμος, αν και βελτιωμένος ως προς τη σύγκλιση, έχει πολύ μεγαλύτερες απαιτήσεις σε μνήμη. Η εν λόγω αλλαγή στον αλγόριθμο είναι ισοδύναμη με την τροποποίηση της συνάρτησης σφάλματος (3-5) ως εξής:

$$E_b = \sum_{k=1}^{N} E_p(k) = \sum_{k=1}^{N} \sum_{i} (e_i(k))^2 \tag{3-17}$$

3.2.3 Αρχικοποίηση δικτύου MLP

Καθοριστική για την εκπαίδευση του δικτύου είναι η σωστή αρχικοποίησή του, δεδομένου ότι οι αρχικές τιμές των βαρών διαδραματίζουν πολύ σημαντικό ρόλο στην πορεία της εκπαίδευσης. Συνήθως τα βάρη αρχικοποιούνται με μικρές τιμές ομοιόμορφα κατανεμημένες σε ένα διάστημα. Μία συχνά χρησιμοποιούμενη μέθοδος αρχικοποίησης βαρών, η οποία επελέγη και στην παρούσα υλοποίηση, είναι η μέθοδος Nguyen-Widrow (Haykin S. 1999 και Nguyen D. & Widrow B. 1990), ένα από τα κύρια πλεονεκτήματα της οποίας είναι το ότι δεν αχρηστεύονται νευρώνες, αφού τα βάρη αρχικοποιούνται στον ενεργό χώρο εισόδου των σιγμοειδών συναρτήσεων, με αποτέλεσμα τα γινόμενα των εισόδων να κινούνται περίπου στο ίδιο εύρος τιμών με τα βάρη και η εκπαίδευση να προχωράει γρηγορότερα. Αποφεύγεται, συνεπώς, η ενεργοποίηση των σιγμοειδών συναρτήσεων στην περιοχή κορεσμού. Όλα τα βάρη στη μέθοδο Nguyen-Widrow αρχικοποιούνται με βάση τον τύπο (3-18):

$$w = 0{,}7 \cdot m^{\frac{1}{p}} \frac{1}{R} U(-0{,}5;0{,}5) \tag{3-18}$$

όπου *m* ο αριθμός των νευρώνων στο επίπεδο, *R* το εύρος των τιμών των εισόδων, *p* ο αριθμός των βαρών σε κάθε νευρώνα (χωρίς τα κατώφλια) και *U(-0,5, 0,5)* η ομοιόμορφη κατανομή στο διάστημα [*-0,5 0,5*].

Σημειώνεται ότι αρχικά το δίκτυο θεωρείται πλήρως διασυνδεδεμένο, οπότε όλοι οι νευρώνες κάθε επιπέδου έχουν τον ίδιο αριθμό βαρών.

Στη συνέχεια, για τα κατώφλια θα θεωρήσουμε *m* τιμές β_i ισοκατανεμημένες στο διάστημα [0 0.7m$^{1/p}$], οπότε,

$$b_i = sign(w_{1i}) \cdot \beta_i \tag{3-19}$$

όπου w_{1i} το βάρος που συνδέει την $1^η$ είσοδό του με το νευρώνα *i*.

Στη μέχρι τώρα παρουσίαση ο τρόπος ενημέρωσης των παραμέτρων ενός νευρωνικού δικτύου βασίζεται στον απλό κανόνα της μεθόδου βαθμωτής κατάβασης (Gradient Descent). Εφαρμόζοντας, ωστόσο, διαφορετικούς κανόνες κίνησης στην επιφάνεια σφάλματος, κυρίως από το χώρο της αριθμητικής ανάλυσης και βελτιστοποίησης, προκύπτουν παραλλαγές του αλγόριθμου ανάστροφης διάδοσης (Back Propagation).

3.2.4 Παραλλαγές του αλγόριθμου ανάστροφης διάδοσης και αποφυγή γενίκευσης MLP

Ο αλγόριθμος Levenberg-Marquardt (LM) (Hagan, M.T & Menhaj, M.B. 1994) είναι ένας γενικός αλγόριθμος βελτιστοποίησης που ελαχιστοποιεί συναρτήσεις τετραγωνικής μορφής, όπως η συνάρτηση σφάλματος (3-5), και εφαρμόζεται γενικότερα και σε άλλα προβλήματα πέραν των νευρωνικών δικτύων. Η ενημέρωση των βαρών στο συγκεκριμένο αλγόριθμο γίνεται με τον ακόλουθο τύπο:

$$W' = W - [J^T J + \mu I]^{-1} J^T \vec{e} \tag{3-20}$$

όπου με *W* συμβολίζεται ο πίνακας-στήλη ο οποίος περιέχει όλα τα βάρη και κατώφλια του δικτύου με κάποια διάταξη και έχει διάσταση *Sx1*, όπου *S* είναι ο αριθμός των παραμέτρων του δικτύου. Με *J* συμβολίζεται ο πίνακας που περιλαμβάνει τις πρώτες παραγώγους των όρων e_i (που εκφράζονται από την εξίσωση (3-6)) ως προς τα βάρη και τα κατώφλια του δικτύου[1].

Ο πίνακας *J* ονομάζεται Ιακωβιανός και έχει διάσταση *QxV*, όπου το *V* ορίζεται από τον αριθμό των εξόδων. Με *I* συμβολίζεται ο μοναδιαίος πίνακας διάστασης *V*, ενώ το διάνυσμα $\vec{e}$ αποτελείται από τους *V* όρους e_i, όπως ορίσθηκαν στη σχέση (3-6).

Το μέγεθος *V* είναι ίσο με το πλήθος των εξόδων του δικτύου, εάν χρησιμοποιείται εκπαίδευση ανά παράδειγμα, ενώ στην περίπτωση της συνολικής εκπαίδευσης (batch training) ισούται με το πλήθος των εξόδων επί τον αριθμό των παραδειγμάτων *N* (καθώς

[1] Σημειώνεται ότι οι παράγωγοι ($\frac{de_i}{dw_j}$) υπολογίζονται με τον αλγόριθμο ανάστροφης διάδοσης.

θα περιλαμβάνει *N* φορές το σφάλμα e_i κάθε εξόδου, σε κυκλική διάταξη). Τέλος, η παράμετρος *μ* καθορίζει την εξέλιξη του αλγόριθμου.

Ο αλγόριθμος Levenberg-Marquardt αποτελεί ένα συνδυασμό της μεθόδου Gauss - Newton (για την επίλυση προβλημάτων βελτιστοποίησης) (Haykin S. 1999, Hagan, M.T & Menhaj, M.B. 1994 και Foresee F. & Hagan, T. 1997) και της μεθόδου βαθμωτής κατάβασης. Όταν η παράμετρος *μ* αυξάνει, τότε η μέθοδος προσεγγίζει περισσότερο τη μέθοδο της βαθμωτής κατάβασης, η οποία ωστόσο είναι λιγότερο αποτελεσματική κοντά σε περιοχές της συνάρτησης σφάλματος που έχουν τοπικά ελάχιστα. Από την άλλη μεριά, η μέθοδος Gauss - Newton είναι πιο γρήγορη στις περιοχές αυτές. Συνεπώς, η παράμετρος *μ* αυξάνεται, όταν χειροτερεύει η απόδοση του δικτύου, ενώ αντίστοιχα μειώνεται, όταν η απόδοση καλυτερεύει, προκειμένου να αναζητηθεί γρηγορότερα το ελάχιστο.

Πρέπει να σημειωθεί ότι ο αλγόριθμος εκπαίδευσης δεν εξασφαλίζει ότι το σύστημα θα σταθεροποιηθεί στην κατάσταση που ελαχιστοποιεί απόλυτα τη συνάρτηση σφάλματος. Συμβαίνει συχνά το δίκτυο που εκπαιδεύεται με τον αλγόριθμο της ανάστροφης διάδοσης να εγκλωβίζεται σε κάποιο τοπικό ελάχιστο.

Η σημαντικότερη ιδιότητα των νευρωνικών δικτύων MLP είναι ότι μπορούν να παράγουν «λογικά» αποτελέσματα για δεδομένα, στα οποία δεν έχουν μεν εκπαιδευτεί, είναι ωστόσο σχετικά με τα δεδομένα εκπαίδευσής τους. Με βάση δηλαδή το σύνολο εκπαίδευσής τους έχουν τη δυνατότητα να εξάγουν νέα συμπεράσματα. Αυτή η ιδιότητά τους είναι και η πλέον σημαντική και ονομάζεται **γενίκευση**.

Σε περίπτωση που ένα δεδομένο δίκτυο είναι αρκετά πιο περίπλοκο από ό,τι απαιτείται για το δεδομένο πρόβλημα, τότε είναι πολύ πιθανό να παρατηρηθεί το φαινόμενο της υπερεκπαίδευσης (overtraining) (Haykin S. 1999), το οποίο και μειώνει δραστικά την ικανότητα γενίκευσης του δικτύου. Το δίκτυο μαθαίνει πάρα πολύ καλά μόνο τα δεδομένα εκπαίδευσης (overfitting), δίνοντας πολύ χαμηλές τιμές στη συνάρτηση σφάλματος (3-5) ή (3-17), αλλά δεν αποδίδει καθόλου καλά σε δεδομένα στα οποία δεν έχει εκπαιδευτεί.

Για να αποφευχθεί το αρνητικό αυτό ενδεχόμενο, ακολουθούνται δύο κυρίως τεχνικές:

1. **Έγκαιρος τερματισμός της εκπαίδευσης** (early stopping): Ένα μέρος των διαθέσιμων δεδομένων, το οποίο ονομάζεται σύνολο επαλήθευσης (validation set), δεν παρουσιάζεται στο δίκτυο κατά τη φάση της εκπαίδευσης, αλλά χρησιμοποιείται για να εκτιμηθεί η απόδοση του δικτύου σε άγνωστα δεδομένα μετά το τέλος κάθε εποχής. Όταν η απόδοση του δικτύου στο σύνολο επαλήθευσης αρχίζει να μειώνεται, τότε η εκπαίδευση τερματίζεται.

2. **Απλοποίηση του μοντέλου**: Η μείωση του αριθμού των νευρώνων στο κρυφό επίπεδο αποτελεί την πλέον ενδεδειγμένη λύση, αφού ο αριθμός των νευρώνων στην είσοδο και την έξοδο εξαρτάται συνήθως από τις παραμέτρους του προβλήματος. Πιο αναλυτικά, το πλήθος των εισόδων εξαρτάται άμεσα από τα διαθέσιμα χαρακτηριστικά, ενώ το πλήθος των εξόδων από τις δυνατές αποφάσεις που χρειάζεται να λάβει το σύστημα. Η απλοποίηση του μοντέλου γίνεται επιβάλλοντας στην εκπαίδευση όχι μόνο να ελαχιστοποιεί το σφάλμα, αλλά και να μειώνει την τιμή των χρησιμοποιούμενων βαρών. Αυτό επιτυγχάνεται με την τροποποίηση του κριτηρίου σφάλματος (3-17) ως εξής:

$$\widehat{E}_b = a \cdot E_b + \beta \cdot E_w \tag{3-21}$$

όπου $\widehat{E}_b$ η νέα συνάρτηση εκτίμησης σφάλματος, ενώ ο όρος E_w εξαρτάται από τα βάρη και συνήθως είναι:

$$E_w = \frac{1}{2}\sum_i (w_i)^2 \tag{3-22}$$

Όπως μπορούμε να παρατηρήσουμε, η $\widehat{E}_b$ «τιμωρεί» δίκτυα με πολλά βάρη και με μεγάλες τιμές βαρών.

Χρησιμοποιώντας τη συνάρτηση σφάλματος (3-21) αντί της (3-17) ή της (3-5), η σχέση (3-20) στον αλγόριθμο Levenberg-Marquardt τροποποιείται ακολούθως:

$$W' = W - [aJ^T J + (\mu + \beta)I]^{-1}(J^T \vec{e} + \beta W) \tag{3-23}$$

όπου ο πίνακας *J* παραμένει ίδιος με αυτόν της συνάρτησης (3-20), ενώ για τον υπολογισμό των παραγώγων δεν χρησιμοποιείται η νέα συνάρτηση εκτίμησης $\widehat{E}_b$ (3-21).

Η προσθήκη των νέων όρων σε σχέση με την (3-20) οφείλεται στο γεγονός ότι κατά τον υπολογισμό των παραγώγων της συνάρτησης σφάλματος ως προς τα βάρη προστίθεται ένας όρος λόγω του δεύτερου όρου της (3-21). Για παράδειγμα, αν η E_w είναι αυτή που δίνεται στον τύπο (3-22), τότε:

$$\frac{d\widehat{E}_b}{dw_i} = a \cdot \frac{dE_b}{dw_i} + \beta \cdot w_i \tag{3-24}$$

Με χρήση συλλογιστικής κατά Bayes (Mackay, D. 1992 και Foresee F. & Hagan, T. 1997) οι παράμετροι *α* και *β* εκτιμώνται και ανανεώνονται με το πέρας κάθε εποχής με βάση τους τύπους (3-25) και (3-26) αντιστοίχως:

$$a' := \frac{S - \{L - \beta \cdot tr[(aJ^T J + \beta I)^{-1}]\}}{2E_b} \tag{3-25}$$

$$\beta' := \frac{L - \beta \cdot tr[(aJ^T J + \beta I)^{-1}]}{2E_w} \tag{3-26}$$

όπου a' και β' είναι οι νέες τιμές των *α* και *β*, *S* είναι ο συνολικός αριθμός εξόδων του δικτύου για όλα τα δεδομένα εκπαίδευσης, δηλαδή είναι ίσος με τις εξόδους του δικτύου επί τον αριθμό των δεδομένων εκπαίδευσης, και *L* είναι ο συνολικός αριθμός παραμέτρων (βάρη και κατώφλια) του δικτύου.

Η ποσότητα *F* του (3-27)

$$F = L - \beta \cdot tr\left[(aJ^T J + \beta I)^{-1}\right] \tag{3-27}$$

εκφράζει τον αριθμό των χρήσιμων μεταβλητών του δικτύου ή των βαθμών ελευθερίας ή, αλλιώς, τον αριθμό των ενεργών παραμέτρων. Είναι χαρακτηριστικό ότι, εάν το δίκτυο είναι το ελάχιστο για τα δεδομένα εκπαίδευσης, τότε $F = L \Rightarrow \beta = 0$ και η ελαχιστοποίηση λαμβάνει χώρα μόνο με βάση το σφάλμα στα δεδομένα. Ο συνδυασμός της Levenberg-Marquardt με την Bayesian συλλογιστική (BR – Bayesian Regulation) και το κριτήριο (3-21) συμβολίζεται ως LMBR.

3.2.5 Μηχανή Διανύσματος Υποστήριξης SVM

Το SVM είναι ένας αρκετά αποτελεσματικός αλγόριθμος κατηγοριοποίησης και βασίζεται σε αρχές στατιστικής μάθησης (Vapnik V. 1998, Haykin S. 1999 και Καραγιάννης Γ., Σταϊνχάουερ Γ. 2001). Βασίζεται στην υπόθεση ότι, ακόμη και αν τα δεδομένα δεν είναι γραμμικά διαχωρίσιμα, είναι εφικτό να αποκτήσουν αυτή την ιδιότητα με το μετασχηματισμό του χώρου εισόδου (input space). Ο μετασχηματισμός αυτός πραγματοποιείται με τη χρήση μίας μη γραμμικής συνάρτησης απεικόνισης, γνωστής ως συνάρτηση *Φ*. Στη συνέχεια, αναζητείται το βέλτιστο υπερεπίπεδο ($\vec{w}$),ικανό να διαχωρίσει δεδομένα δύο τάξεων (που η ετικέτα τους αντιστοιχίζεται σε $c = \pm 1$), αφού έχουν μετασχηματισθεί στο νέο χώρο ($\Phi(\vec{x})$):

$$\begin{aligned} &\Phi(\vec{x}_i)\cdot\vec{w}+b \geq +1, \ \acute{o}\tau\alpha\nu \quad c_i = +1 \\ &\Phi(\vec{x}_i)\cdot\vec{w}+b \leq -1, \ \acute{o}\tau\alpha\nu \quad c_i = -1 \end{aligned} \tag{3-28}$$

όπου $\vec{x}_i$ είναι το *i*-οστό παράδειγμα που εμφανίζεται στην είσοδο, ενώ το υπερεπίπεδο ($\vec{w}$) πρέπει να ελαχιστοποιεί την ακόλουθη συνάρτηση (3-29):

$$min \ \frac{1}{2}\|\vec{w}\|^2 \tag{3-29}$$

Η παραπάνω έκφραση του προβλήματος βελτιστοποίησης με περιορισμούς έχει τροποποιηθεί, ώστε να δίνεται η δυνατότητα στο μοντέλο να αντεπεξέρχεται σε περιπτώσεις όπου τα δεδομένα δεν μπορούν να διαχωριστούν γραμμικά. Αυτή η διαφοροποίηση υλοποιείται με την εισαγωγή των βοηθητικών μεταβλητών ξ, με τις οποίες το πρόβλημα τροποποιείται ως εξής:

$$\begin{aligned} &\Phi(\vec{x}_i)\cdot\vec{w}+b \geq +1-\xi_i, \ \acute{o}\tau\alpha\nu \quad c_i = +1 \\ &\Phi(\vec{x}_i)\cdot\vec{w}+b \leq -1+\xi_i, \ \acute{o}\tau\alpha\nu \quad c_i = -1 \end{aligned} \tag{3-30}$$

Αντίστοιχα, και η αντικειμενική συνάρτηση (3-29) εκφράζεται ως:

$$min \ \frac{1}{2}\|\vec{w}\|^2 + C\sum_i \xi_i \tag{3-31}$$

όπου η παράμετρος *C* ρυθμίζει το μέγιστο ποσοστό των παραδειγμάτων που επιτρέπεται να κατηγοριοποιηθούν με λανθασμένο τρόπο.

Ο πυρήνας (kernel) του SVM ορίζεται ως μία συνάρτηση ίση με το εσωτερικό γινόμενο δύο μετασχηματισμένων εισόδων:

$$K(\vec{x}_i,\vec{x}_j) = <\Phi(\vec{x}_i),\Phi(\vec{x}_j)> \tag{3-32}$$

Χρησιμοποιώντας τον ορισμό του πυρήνα, η δυαδική μορφή (dual form) του προβλήματος εκφράζεται χωρίς περιορισμούς ως:

$$max \sum_{i=1}^{R} a_i - \frac{1}{2}\sum_{i=1}^{R}\sum_{j=1}^{R} a_i a_j c_i c_j K(\vec{x}_i,\vec{x}_j) \tag{3-33}$$

όπου *R* είναι η διάσταση της εισόδου και *a* οι μεταβλητές ως προς τις οποίες επιλύεται το πρόβλημα βελτιστοποίησης και οι οποίες αντανακλούν τη συμβολή του κάθε παραδείγματος εκπαίδευσης στο σχηματισμό του βέλτιστου υπερεπίπεδου (support vectors).

Χρησιμοποιώντας την συνάρτηση πυρήνα το βέλτιστο υπερεπίπεδο ορίζεται ως:

$$D(\vec{x}) = sign\left(\sum_{i=1}^{R} c_i a_i K(\vec{x}_i, \vec{x}) + b\right) \tag{3-34}$$

Οι μοναδικές παράμετροι που πρέπει να ορίζονται κατά το σχεδιασμό ενός SVM είναι το είδος του πυρήνα (ο οποίος συνηθέστερα είναι της μορφής της συνάρτησης Gauss) και η τιμή της παραμέτρου *C*, που εκφράζει την ανοχή στα σφάλματα κατά τη διαδικασία της εκπαίδευσης.

Το SVM, όπως παρουσιάσθηκε, φαίνεται ότι είναι προσαρμοσμένο σε προβλήματα μόνο δύο τάξεων. Για να χρησιμοποιηθεί σε προβλήματα πολλαπλών τάξεων, απαιτούνται ορισμένες τροποποιήσεις. Πιο συγκεκριμένα, απαιτείται η εκπαίδευση *m* πλήθους SVM, όπου *m* ισούται με το πλήθος των τάξεων και κάθε μοντέλο από αυτά αποφασίζει αν ένα πρότυπο αντιστοιχίζεται στη δεδομένη τάξη ή όχι. Το μοντέλο που θα παρουσιάσει τη μεγαλύτερη θετική απόκριση είναι αυτό στο οποίο τελικά θα αποδοθεί το δεδομένο. Η μεθοδολογία αυτή είναι γνωστή γενικότερα ως **μηχανή επιτροπής** (committee machine), διάφορες παραλλαγές της οποίας έχουν προταθεί στη βιβλιογραφία (Duda R.O., Hart P.E. & Stork D.G. 2001 και Καραγιάννης Γ., Σταϊνχάουερ Γ. 2001).

3.3 Περιγραφή δεδομένων

Για τα πειράματα αξιολόγησης του συστήματος χρησιμοποιήθηκε ένα σύνολο 1005 κειμένων από τα πρακτικά συνεδριάσεων του ελληνικού κοινοβουλίου. Τα κείμενα αυτά, τα οποία περιλαμβάνουν ομιλίες πέντε μελών της Ελληνικής Βουλής από διαφορετικά πολιτικά κόμματα κατά την περίοδο Οκτώβριος 1996 – Μάρτιος 2000, έχουν υποστεί προεπεξεργασία, προκειμένου να αφαιρεθούν σύντομοι διάλογοι και ολιγόλογες παρεμβάσεις. Επιπλέον διορθώθηκαν τα όποια ορθογραφικά λάθη. Πιο αναλυτικά, το σώμα κειμένων αξιολόγησης έχει την ακόλουθη δομή:

Ομιλητές	**1997**	**1998**	**1999**	**2000**	**Αριθμός ομιλιών**	**Αριθμός λέξεων**
Ομιλητής Α	161	107	125	25	418	463.680
Ομιλητής Β	36	28	18	3	85	177.853
Ομιλητής Γ	81	71	67	26	245	241.882
Ομιλητής Δ	72	49	30	3	154	217.305
Ομιλητής Ε	16	42	38	7	103	190.601
Σύνολο	**366**	**297**	**278**	**64**	**1005**	**1.291.321**

Πίνακας 3.2. Σώμα κειμένων που χρησιμοποιήθηκε για τα πειράματα αξιολόγησης του συστήματος

3.4 Πειράματα Συστήματος

Για την εκτίμηση της ακρίβειας κατηγοριοποίησης του συστήματος ακολουθήθηκε μία επαναληπτική διαδικασία αξιολόγησης (leave-one-out), κατά την οποία αρχικά διαιρείται το σύνολο των δεδομένων σε δέκα μικρότερα σύνολα, ξένα μεταξύ τους. Από τα σύνολα αυτά, τα οκτώ χρησιμοποιούνται για την εκπαίδευση του νευρωνικού δικτύου, το ένα για τον έγκαιρο τερματισμό της εκπαίδευσης και το τελευταίο ως σύνολο επαλήθευσης. Η ακρίβεια κατηγοριοποίησης μετράται μόνο πάνω στο άγνωστο, κατά την εκπαίδευση, σύνολο δοκιμής. Τα πειράματα κατηγοριοποίησης επαναλαμβάνονται για όλους τους δυνατούς συνδυασμούς συνόλων, ενώ η απόδοση του μοντέλου δίνεται ως ο μέσος όρος των συνδυασμών αυτών (συνολικά 90 συνδυασμοί).

Ένα από τα βασικότερα ζητήματα στην εφαρμογή αλγορίθμων νευρωνικών δικτύων είναι ο κατάλληλος προσδιορισμός της αρχιτεκτονικής τους. Είναι δεδομένο ότι οι συνθήκες του προβλήματος καθορίζουν το πλήθος των εισόδων (ή των νευρώνων στο επίπεδο εισόδου), το οποίο πρέπει να είναι ίσο με τη διάσταση του διανύσματος εισόδου και, τις περισσότερες φορές, με το πλήθος των χαρακτηριστικών, το οποίο στην παρούσα εφαρμογή είναι 85.

Στο επίπεδο εξόδου, το πλήθος των νευρώνων καθορίζεται από το είδος του προβλήματος. Σε προβλήματα κατηγοριοποίησης ο αριθμός των εξόδων είναι συνήθως ίσος με το πλήθος των επιθυμητών τάξεων. Αν και είναι δυνατό με κατάλληλη δυαδική κωδικοποίηση (συνδυασμοί εξόδων) να χρησιμοποιηθούν λιγότερες έξοδοι ($\lceil log_2 N \rceil$, όπου *N* ο αριθμός των επιθυμητών τάξεων), κάτι τέτοιο δεν είναι επιθυμητό, δεδομένου ότι εισάγει κάποια προκατάληψη (bias) στο σύστημα, αφού οι αποστάσεις μεταξύ όλων των ζευγαριών των τάξεων δεν είναι ίδιες και κάποιες τάξεις εμφανίζονται αυθαίρετα να είναι πιο κοντά από κάποιες άλλες (είτε βάσει της Ευκλείδειας απόστασης είτε βάσει της απόστασης Hamming) (MacKay D. 2003).

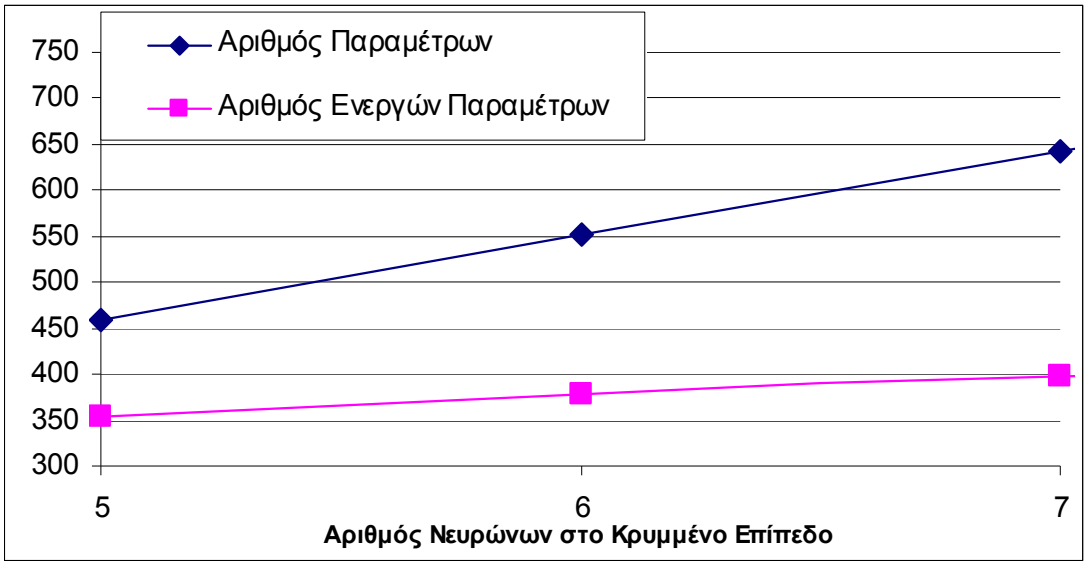
Στα πλαίσια ενός συστήματος κατηγοριοποίησης, αποδίδεται σε κάθε τάξη μία έξοδος, η οποία θα πρέπει να ενεργοποιείται, μόνο όταν το πρότυπο ανήκει στην αντίστοιχη τάξη. Στην προτεινόμενη αρχιτεκτονική, όπου υπάρχουν πέντε τάξεις προς επεξεργασία, απαιτούνται πέντε νευρώνες εξόδου. Σε συνθήκες ενός πραγματικού προβλήματος, όπου οι έξοδοι δεν είναι απόλυτα κβαντισμένες στο επιλεγμένο διάστημα ([0,1] ή [-1,1]), συνήθως επιλέγεται η τάξη που έχει τη μέγιστη έξοδο.

Πιο συγκεκριμένα, έχει αποδειχθεί ότι τα πολυστρωματικά δίκτυα (MLP) με μόνο ένα κρυφό επίπεδο έχουν τη δυνατότητα να λύσουν οποιοδήποτε πρόβλημα κατακερματισμού του χώρου προτύπων, όσο δύσκολο και αν είναι αυτό, με μοναδική προϋπόθεση το κρυφό επίπεδο να διαθέτει ικανό πλήθος νευρώνων. Από την άλλη, το δίκτυο δε θα πρέπει να έχει πάρα πολλούς νευρώνες (πολύ περισσότερους από ό,τι χρειάζεται), γιατί τότε προκύπτει εντονότερα το φαινόμενο της υπερεκπαίδευσης, αφού το δίκτυο γίνεται πολυπλοκότερο (δηλαδή αποτελείται από περισσότερες παραμέτρους οι οποίες πρέπει να προσδιορισθούν), ενώ το πλήθος των δεδομένων γίνεται μικρότερο συγκριτικά με το πλήθος των παραμέτρων, με συνέπεια την αδυναμία του δικτύου να γενικεύσει σωστά σε άγνωστα δεδομένα.

Για να αντιμετωπισθεί το πρόβλημα αυτό, αρχικά χρησιμοποιήθηκε ο αλγόριθμος εκπαίδευσης Levenberg-Marquardt σε συνδυασμό με το κριτήριο μείωσης των τιμών των βαρών (LMBR). Το δίκτυο που επιλέχθηκε διέθετε αρχικά ένα αρκετά μεγάλο αριθμό νευρώνων στο κρυφό επίπεδο (12 νευρώνες), ενώ εκπαιδεύτηκε με όλα τα διαθέσιμα δεδομένα, προκειμένου να εκτιμηθεί η επάρκειά του στην αντιμετώπιση του συγκεκριμένου προβλήματος. Ο αλγόριθμος LMBR παρέχει ένα στατιστικό κριτήριο εκτίμησης του πλήθους των ενεργών παραμέτρων του δικτύου. Εάν υποθέσουμε ότι το δίκτυο διαθέτει h νευρώνες στο κρυφό επίπεδο, τότε ο συνολικός αριθμός παραμέτρων (βαρών και κατωφλίων) είναι:

$$S_p = (85+1)\cdot h + (h+1)\cdot 5 = 91\cdot h + 5 \tag{3-35}$$

Εάν η εξίσωση (3-35) επιλυθεί ως προς h, χρησιμοποιώντας το πλήθος των ενεργών παραμέτρων όπως προέκυψε από το πείραμα, προσδιορίζεται το μέγεθος του κρυφού επιπέδου. Για το δεδομένο πρόβλημα, το h βρέθηκε ίσο με 4,56. Επειδή όμως οι παράμετροι του δικτύου δεν προσαρμόζονται τελείως ανεξάρτητα, είναι αναμενόμενο ότι απαιτούνται περισσότεροι νευρώνες. Συνεπώς, η τιμή αυτή απλώς αποτέλεσε ένα αρχικό σημείο για τον προσδιορισμό των παραμέτρων του δικτύου. Για το λόγο αυτό, εκτιμήθηκε η απόδοση σε δίκτυα με πέντε, έξι και εφτά κρυφούς νευρώνες. Η απόδοση των συστημάτων με έξι και εφτά κρυφούς νευρώνες είναι 88,7% και με τους πέντε νευρώνες είναι 88,4%. Ο αριθμός των ενεργών παραμέτρων, όπως εκτιμώνται από τον αλγόριθμο εκπαίδευσης για τα διάφορα μεγέθη των δικτύων, φαίνεται στο σχήμα 3.4.



Σχήμα 3.4. Αριθμός πραγματικών παραμέτρων σε σχέση με τις ενεργές παραμέτρους όπως τις εκτιμάει ο αλγόριθμος LMBR

Είναι σαφές ότι, όσο μεγαλώνει το μοντέλο, τόσο περισσότερο διαφοροποιούνται μεταξύ τους οι πραγματικές παράμετροι από τις ενεργές παραμέτρους που εκτιμάει ο αλγόριθμος. Συνεπώς, μεγαλύτερο πλήθος παραμέτρων μένει ανενεργό και δε συμβάλλει στη βελτίωση του αποτελέσματος. Αντιθέτως, εισάγει μεγαλύτερη πολυπλοκότητα στο δίκτυο και επιβαρύνει άσκοπα τον αλγόριθμο εκπαίδευσης.

Το κριτήριο σφάλματος που επιλέχθηκε σε συνδυασμό με τον αλγόριθμο ανάστροφης διάδοσης (LMBR) βελτιστοποιεί τόσο την αρχιτεκτονική του δικτύου, οδηγώντας το σε μείωση της τιμής των βαρών του, όσο και την ακρίβεια κατηγοριοποίησης του συνόλου εκπαίδευσης. Με τη μεταβολή των παραμέτρων *a*, *b* είναι εφικτό να κατευθυνθεί η εκπαίδευση του δικτύου περισσότερο στη μείωση των τιμών των βαρών ή στη μείωση του μέσου τετραγωνικού σφάλματος των δεδομένων εκπαίδευσης. Ο αλγόριθμος ανάστροφης διάδοσης που παρουσιάσθηκε έχει την τάση να προσαρμόζει αυτόματα τις παραμέτρους *a, b* της συνάρτησης σφάλματος με χρήση Bayesian συλλογιστικής ανάλογα με τη διαφοροποίηση που παρατηρείται κατά τη φάση της εκπαίδευσης. Πιο συγκεκριμένα, εάν υποθέσουμε ότι το σφάλμα μειώνεται κατά πολύ, τότε το κριτήριο βελτιστοποίησης ενισχύει τη συμμετοχή της πολυπλοκότητας του μοντέλου. Αντιθέτως, εάν οι παράμετροι είναι ακριβώς όσες απαιτούνται, τότε στο κριτήριο βελτιστοποίησης ενισχύεται ο όρος του σφάλματος.

Για την καλύτερη εκτίμηση των παραμέτρων (*a, b*), οι οποίες επηρεάζουν σε πολύ μεγάλο βαθμό την εκπαίδευση και ως εκ τούτου τη γενίκευση του δικτύου σε άγνωστα δεδομένα, εκτός από τον αλγόριθμο ανάστροφης διάδοσης (LMBR), ο οποίος προσδιορίζει αυτόματα σε κάθε βήμα τις τιμές αυτές, χρησιμοποιήθηκε για λόγους σύγκρισης μία απλούστερη διαδικασία γραμμικής αναζήτησης (linear search). Πιο συγκεκριμένα, τέθηκε αρχικά ο περιορισμός να αθροίζουν οι τιμές των *a* και *b* στη μονάδα. Επιλέχθηκε ένα σύνολο από ισοκατανεμημένες στο διάστημα [0,1] τιμές για την παράμετρο *a* και με μέγεθος βήματος 0,2. Με βάση τον περιορισμό του αθροίσματος προσδιορίζονται και οι τιμές της παραμέτρου *b*. Όλες οι τιμές των παραμέτρων (συνολικά έξι ζευγάρια) χρησιμοποιήθηκαν για να εκπαιδεύσουν διαφορετικά δίκτυα και τελικά επιλέχθηκε εκείνο το δίκτυο με το μικρότερο σφάλμα στο σύνολο επαλήθευσης (validation set), το οποίο ήταν εν μέρει άγνωστο στον αλγόριθμο εκπαίδευσης. Τα αποτελέσματα για την ακρίβεια κατηγοριοποίησης του αλγόριθμου γραμμικής αναζήτησης παρουσιάζονται στον πίνακα 3.3.

	MLP 5	MLP 6	MLP 7
LMBR	88,7%	88,7%	88,5%
Γραμμική Αναζήτηση (Linear Search)	89,3%	89,5%	89,2%

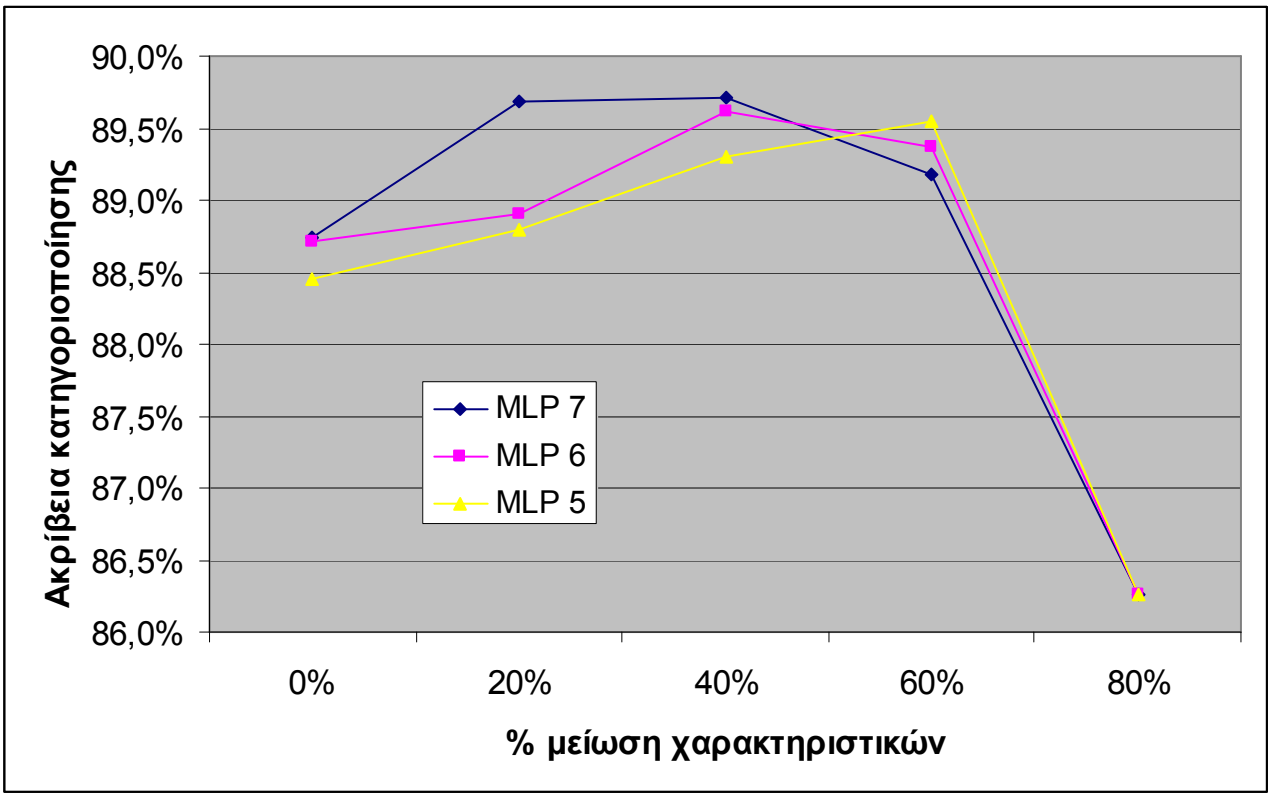
Πίνακας 3.3. Συγκριτικός πίνακας της απόδοσης του MLP εκπαιδευόμενου με την LMBR σε σχέση με την απλούστερη μέθοδο γραμμικής αναζήτησης

Όπως φαίνεται από τον πίνακα 3.3, ο αλγόριθμος γραμμικής αναζήτησης είναι καλύτερος σε ποσοστό που κυμαίνεται από 0,5% έως 0,8%. Πρέπει να ληφθεί υπόψιν ότι ο αλγόριθμος γραμμικής αναζήτησης είναι πολύ πιο απαιτητικός, αφού προϋποθέτει την εκπαίδευση ενός συνόλου από δίκτυα (όσος είναι ο διαφορετικός αριθμός των τιμών του *a* ή *b*). Δεδομένων του μεγάλου υπολογιστικού κόστους, αλλά και της μικρής βελτίωσης της απόδοσης η χρήση της LMBR κρίνεται προτιμητέα.

Το κριτήριο σφάλματος εκπαίδευσης του δικτύου (3-21) οδηγεί το δίκτυο σε μείωση των τιμών των παραμέτρων του. Με αυτό το δεδομένο, μπορεί να εκτιμηθεί η συμβολή του κάθε χαρακτηριστικού στην προκύπτουσα ακρίβεια κατηγοριοποίησης ανάλογα με την ένταση (το πλάτος ή την απόλυτη τιμή) των συνδέσεων που έχει κάθε είσοδος του δικτύου με το κρυφό επίπεδο. Ως ένα τέτοιο μέτρο εκτίμησης της σημασίας κάθε παραμέτρου χρησιμοποιήθηκε το κριτήριο S_j:

$$S_j = E\{\sum_{i=1}^{h} |w_{ij}|\} \tag{3-36}$$

όπου ο δείκτης *j* ορίζει το εκάστοτε χαρακτηριστικό, w_{ij} είναι το βάρος που συνδέει το χαρακτηριστικό *j* με το νευρώνα *i* του κρυφού επιπέδου, *h* είναι ο αριθμός των νευρώνων στο κρυφό επίπεδο και το *E* συμβολίζει τη μέση τιμή για όλα τα πειράματα που πραγματοποιήθηκαν για όλους τους δυνατούς συνδυασμούς συνόλων εκπαίδευσης και δοκιμής. Τα 85 χαρακτηριστικά ιεραρχήθηκαν ανάλογα με την τιμή του κριτηρίου (3-5). Σταδιακά αφαιρέθηκε από το σύστημα το 20% (17 σε απόλυτη τιμή) των λιγότερο σημαντικών χαρακτηριστικών από το διάνυσμα εισόδου. Κάθε μειωμένο διάνυσμα εισόδου χρησιμοποιήθηκε για την εκπαίδευση και στη συνέχεια για τον έλεγχο της απόδοσης ενός νέου δικτύου. Τα αποτελέσματα της ακρίβειας κατηγοριοποίησης σε συνδυασμό με το ποσοστό μείωσης του διανύσματος χαρακτηριστικών για τα διάφορα μεγέθη δικτύων δίνονται στο σχήμα 3.5.

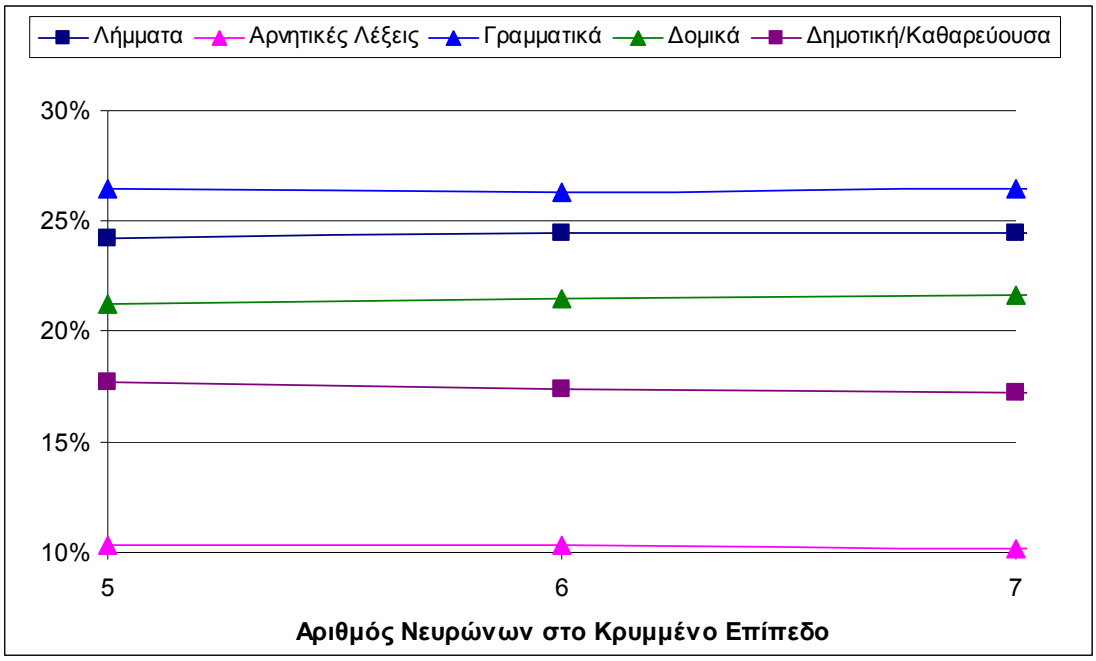


Σχήμα 3.5. Μεταβολή της απόδοσης του συστήματος ως συνάρτηση της μείωσης των χαρακτηριστικών για διάφορα μεγέθη δικτύου

Από το σχήμα 3.5 γίνεται σαφές ότι η μείωση των χαρακτηριστικών αρχικά επιφέρει βελτίωση της απόδοσης σχεδόν για όλα τα μοντέλα. Αυτό μπορεί να αναχθεί στο γεγονός ότι κάποια χαρακτηριστικά ίσως να μη λειτουργούν ικανοποιητικά και να εισάγουν περισσότερο θόρυβο παρά να προσφέρουν χρήσιμη για το πρόβλημα πληροφορία. Επιπλέον, η αφαίρεση χαρακτηριστικών οδηγεί σε μείωση παραμέτρων των δικτύων και ως εκ τούτου σε απλούστερα μοντέλα που κατά κανόνα γενικεύουν καλύτερα σε άγνωστα δεδομένα. Οι αλγόριθμοι εκπαίδευσης, που συνήθως καταλήγουν σε τοπικά ελάχιστα, συμπεριφέρονται καλύτερα με λιγότερες παραμέτρους (Curse Of Dimensionality). Βέβαια η μείωση των παραμέτρων κατά 80% οδηγεί σε δραστική μείωση της απόδοσης του συστήματος. Ενδεικτικά αναφέρεται ότι, όταν το διάνυσμα εισόδου μειώνεται κατά 40% (δηλαδή όταν υπάρχουν μόλις 51 χαρακτηριστικά), η απόδοση που παρατηρείται για τα δίκτυα με 5, 6 και 7 νευρώνες στο κρυφό επίπεδο είναι αντίστοιχα 89,3%, 89,6% και 89,7%. Επιπλέον, εάν διατηρηθούν μόνο 34 χαρακτηριστικά (60% μείωση), τότε η απόδοση μειώνεται ελάχιστα στις τιμές 89,5%, 89,4% και 89,2% αντίστοιχα.

Είναι προφανές ότι η μείωση των παραμέτρων μπορεί να επιδράσει θετικά τόσο στη βελτίωση της απόδοσης, αλλά και στην απλοποίηση του συστήματος ακόμα και κατά τη διαδικασία της προεπεξεργασίας και της μέτρησης χαρακτηριστικών. Ειδικότερα η μείωση του πλήθους των χαρακτηριστικών κατά τη φάση του σχεδιασμού του συστήματος επιφέρει σημαντικό υπολογιστικό κέρδος, όταν το σύστημα καλείται να εφαρμοσθεί σε νέα δεδομένα, μιας και το διάνυσμα αναπαράστασης θα απαιτεί τη μέτρηση λιγότερων χαρακτηριστικών.

Κατά τη διαδικασία μείωσης των παραμέτρων και ανάλογα με τις τιμές του κριτηρίου αξιολόγησης της σπουδαιότητας (3-36) παρουσιάζεται η συμβολή κάθε ομάδας χαρακτηριστικών στο τελικό αποτέλεσμα για όλα τα δίκτυα που δοκιμάσθηκαν. Τα αποτελέσματα που παρουσιάζονται στο ακόλουθο σχήμα (3.6) δείχνουν τη συμβολή της κάθε ομάδας, κανονικοποιημένης ως προς τον αριθμό των χαρακτηριστικών που ανήκουν σε κάθε ομάδα (αφού οι ομάδες δεν περιέχουν ίσο αριθμό χαρακτηριστικών).

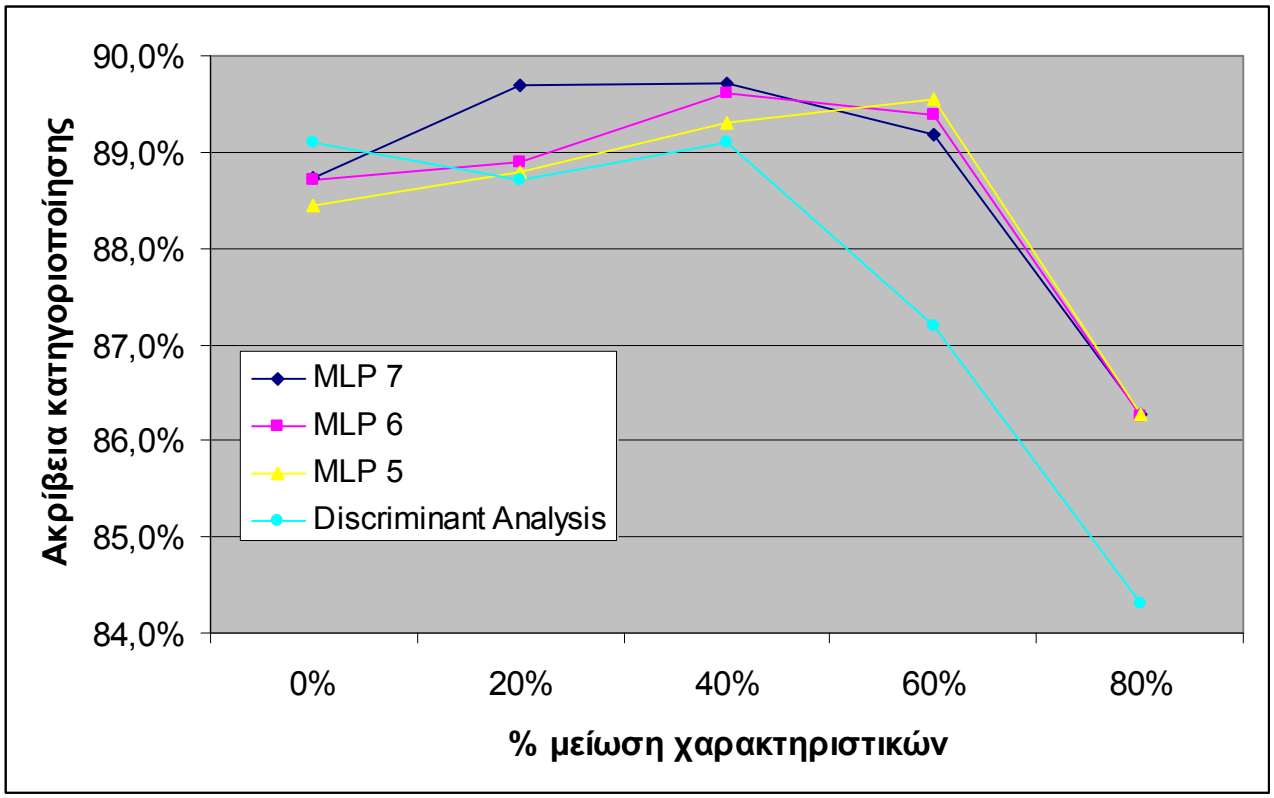


Σχήμα 3.6. Σχετική συμβολή κάθε ομάδας χαρακτηριστικών στο τελικό αποτέλεσμα της κατηγοριοποίησης για διάφορα μεγέθη MLP

Από το σχήμα 3.6 φαίνεται ότι η πιο σημαντική ομάδα χαρακτηριστικών είναι αυτή των μερών του λόγου, ακολουθούμενη από εκείνες των λημμάτων και των δομικών χαρακτηριστικών, ενώ η ομάδα των λέξεων που εκφράζουν άρνηση συμβάλλει ελάχιστα στο τελικό αποτέλεσμα. Αξίζει να σημειωθεί ότι η συμβολή των χαρακτηριστικών είναι πρακτικά αμετάβλητη για τα όλα μεγέθη νευρωνικών δικτύων που δοκιμάσθηκαν.

Ένα αντίστοιχο σύνολο πειραμάτων, στηριζόμενο όμως σε στατιστικές μεθόδους, είχε λάβει χώρα στο ΙΕΛ κατά το παρελθόν και αφορούσε στο ίδιο σύνολο δεδομένων. Συγκεκριμένα, είχαν χρησιμοποιηθεί τα ίδια χαρακτηριστικά (85 τον αριθμό) σε συνδυασμό με το στατιστικό αλγόριθμο κατηγοριοποίησης "γραμμικού διαχωρισμού" (Linear Discriminant Analysis – LDA) (Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D. & Carayannis, G. 2004). Η ακρίβεια κατηγοριοποίησης αυτού του μοντέλου για όλες τις παραμέτρους ήταν αρχικά 89,1%, ακρίβεια ελάχιστα μεγαλύτερη από την ακρίβεια των αντίστοιχων μοντέλων (με όλες τις παραμέτρους) των νευρωνικών δικτύων MLP, ωστόσο όχι υψηλότερη από την ακρίβεια των μοντέλων με λιγότερες παραμέτρους (π.χ. για 40% μείωση και 7 νευρώνες επιτυγχάνεται ακρίβεια 89,7%).

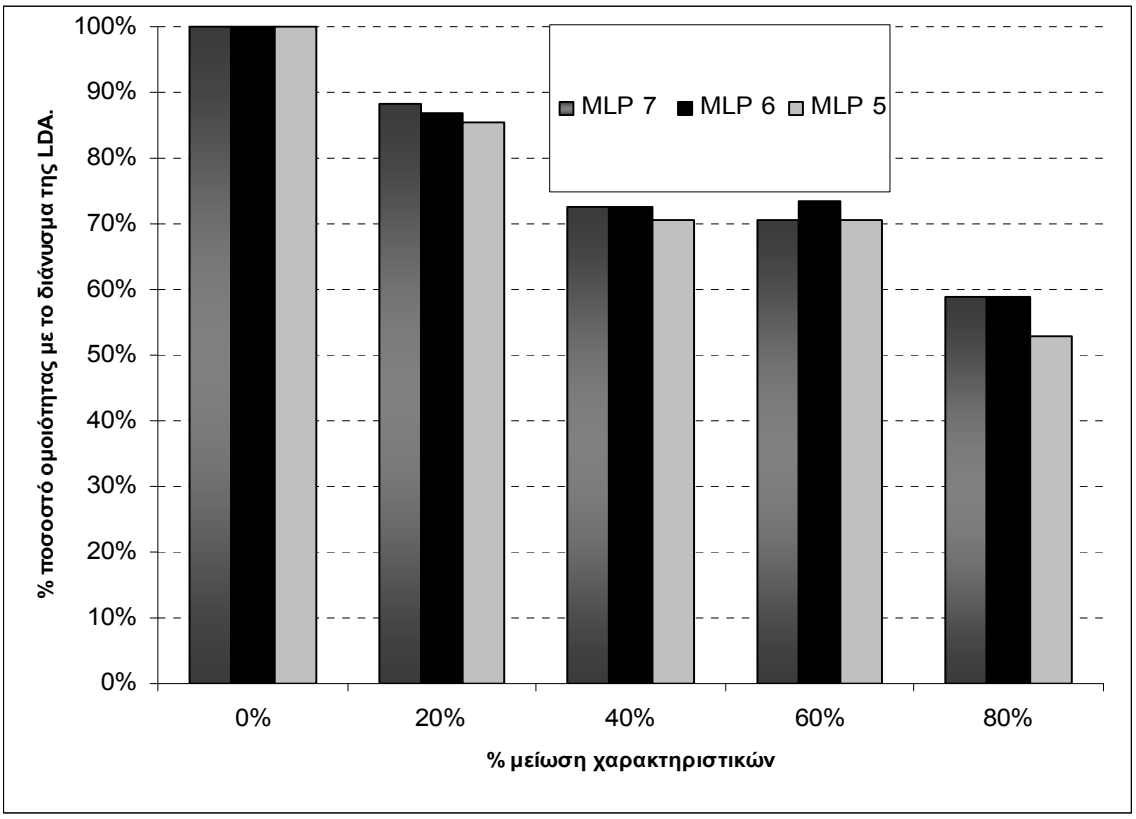
Για την καλύτερη αξιολόγηση των δύο μεθόδων έγινε μία συγκριτική δοκιμή για την περίπτωση κατά την οποία μειώνεται ο αριθμός των παραμέτρων – χαρακτηριστικών. Για την αποτίμηση της σημασίας της κάθε παραμέτρου στην περίπτωση της LDA χρησιμοποιήθηκε η τιμή της παραμέτρου *F*, η οποία δηλώνει την πληροφορία που προσφέρει στο μοντέλο η κάθε μεταβλητή. Μεταβάλλοντας την τιμή της παραμέτρου *F*, είναι δυνατόν το μοντέλο που παράγει η LDA να διατηρεί λιγότερα χαρακτηριστικά. Με τον τρόπο αυτό μπορεί να εκτιμηθεί η απόδοση του συστήματος, όταν μειώνεται κατά κάποιο ποσοστό η διάσταση του διανύσματος εισόδου, όπως αντίστοιχα στην περίπτωση του MLP με τη χρήση του κριτηρίου (3-36). Στο παρακάτω διάγραμμα του σχήματος 3.7 συγκρίνεται η ακρίβειας της μεθόδου LDA σε σχέση με την ακρίβεια που επιτυγχάνουν τα δίκτυα MLP.



Σχήμα 3.7. Συγκριτική απόδοση του MLP με τη μέθοδο LDA για διάφορα μεγέθη διανύσματος εισόδου

Από το σχήμα 3.7 γίνεται σαφές ότι για το πλήρες σύνολο χαρακτηριστικών η LDA συμπεριφέρεται καλύτερα, καθώς κατηγοριοποιεί τα δεδομένα με υψηλότερη ακρίβεια. Όταν όμως οι παράμετροι μειώνονται, το μοντέλο LDA δε βελτιώνεται, ενώ αντίθετα το μοντέλο MLP παρουσιάζει βελτίωση και υπερβαίνει την ακρίβεια του αντίστοιχου μοντέλου LDA. Η καλύτερη απόδοσή του μάλιστα, δηλαδή 89,7%, παρατηρείται για το μειωμένο κατά 40% μοντέλο που διαθέτει εφτά νευρώνες.

Για να αποτιμηθεί σαφέστερα η συμπεριφορά αυτών των δύο αλγορίθμων κατηγοριοποίησης, μελετήθηκαν περαιτέρω εκείνα τα χαρακτηριστικά, τα οποία τελικά διαδραματίζουν το σημαντικότερο ρόλο σε κάθε περίπτωση. Αναλυτικότερα, εκτιμήθηκε η ομοιότητα του μειωμένου διανύσματος εισόδου ως προς τα χαρακτηριστικά που διατηρούνται, όπως προκύπτει από τα MLP με πέντε, έξι και εφτά κρυφούς νευρώνες, με το διάνυσμα ίδιας διάστασης που προκύπτει από τη μέθοδο LDA. Πιο συγκεκριμένα, μετρήθηκε το ποσοστό των κοινών χαρακτηριστικών για τις δύο μεθόδους (σχήμα 3.8). (Είναι σαφές ότι για 0% μείωση όλες οι μέθοδοι διατηρούν το ίδιο σύνολο χαρακτηριστικών).



Σχήμα 3.8. Σύγκριση της ομοιότητας των MLP και LDA ως προς τα χαρακτηριστικά που διατηρούνται

Παρατηρείται ότι, όταν υπάρχει μείωση των χαρακτηριστικών κατά 20%, πάνω από 80% (85%-90%) των χαρακτηριστικών είναι κοινό και στις δύο μεθόδους. Συνεπώς, σε αυτό το στάδιο και τα δύο μοντέλα αποτιμούν με τον ίδιο τρόπο τη σπουδαιότητα των χαρακτηριστικών. Αντίθετα, όταν το αφαιρούμενο ποσοστό υπερβαίνει το 20%, τότε η ομοιότητα μειώνεται περίπου στο 75% ή ακόμα περισσότερο. Το γεγονός αυτό, σε συνδυασμό με τη βελτιωμένη απόδοση των μοντέλων MLP για λιγότερες παραμέτρους, οδηγεί στο συμπέρασμα ότι η μέθοδος MLP προσφέρει ένα καλύτερο τρόπο απλοποίησης του μοντέλου. Σημειώνεται ότι για λόγους πληρότητας έγινε και η σύγκριση των παραμέτρων μεταξύ των μοντέλων MLP για τα διάφορα στάδια μείωσης χαρακτηριστικών (πίνακας 3.4). Τα μοντέλα MLP με διαφορετική αρχιτεκτονική διατηρούν περίπου το ίδιο σύνολο χαρακτηριστικών, αφού γενικά παρουσιάζουν ομοιότητα, η οποία υπερβαίνει το 90%.

	MLP 5 – MLP 6	MLP 5 – MLP 7	MLP 6 – MLP 7
0%	100%	100%	**100%**
20%	99%	97%	**99%**
40%	96%	94%	**98%**
60%	88%	85%	**94%**
80%	**94%**	**94%**	**100%**

Πίνακας 3.4. Σύγκριση της ομοιότητας των χαρακτηριστικών που διατηρούνται στα μοντέλα MLP με διαφορετικό πλήθος κρυφών νευρώνων

Συμπερασματικά, όταν επιλέγεται ένα υποσύνολο χαρακτηριστικών, τα δίκτυα MLP διαφέρουν από την LDA ως προς τα χαρακτηριστικά που επιλέγουν (βλ. σχήμα 3.8). Όπως προκύπτει, ωστόσο, από τα πειραματικά αποτελέσματα, το MLP επιτυγχάνει μεγαλύτερη απόδοση στο δεδομένο πρόβλημα, όταν μειώνονται τα χαρακτηριστικά, άρα φαίνεται να έχει τη δυνατότητα να επιλέγει καλύτερο υποσύνολο χαρακτηριστικών.

3.5 Σύγκριση προτεινόμενου συστήματος με άλλα συστήματα

Για το διαχωρισμό των κειμένων η συνηθέστερη πρακτική είναι η χρήση χαρακτηριστικών TF-IDF (Drucker, H., Wu, D. & Vapnik, V. 1999) σε συνδυασμό με τον αλγόριθμο κατηγοριοποίησης SVM (Support Vector Machines – Μηχανές Διανύσματος Υποστήριξης).

Ο όρος **TF-IDF** (Term Frequency – Inverse Document Frequency) χρησιμοποιείται ευρέως σε εφαρμογές εξόρυξης πληροφορίας από κείμενα (information retrieval – text mining (Salton, G. & Buckley, C. 1988)). Πρόκειται για ένα σύνθετο όρο , ο οποίος εκφράζει το γινόμενο των όρων TF και IDF. Ειδικότερα, ο όρος TF (συχνότητα όρου) απεικονίζει τον αριθμό των εμφανίσεων ενός όρου, συνήθως λήμματος, στο κείμενο, κανονικοποιημένου ως προς το συνολικό αριθμό όρων στο δεδομένο κείμενο. Με αυτό τον τρόπο κανονικοποίησης μετριάζεται η διαφοροποίηση ανάμεσα σε διαφορετικού μεγέθους κείμενα. Ο όρος IDF (αντίστροφη συχνότητα κειμένου) μετράει το ποσοστό των κειμένων στα οποία εμφανίζεται ένας δεδομένος όρος. Υπολογίζεται διαιρώντας τον αριθμό των κειμένων που περιέχουν τον όρο ως προς το σύνολο των κειμένων και λογαριθμίζοντας το προκύπτον αποτέλεσμα. Αυτό το είδος χαρακτηριστικών είναι δυνατόν να χρησιμοποιηθεί εναλλακτικά, αντί για τα 85 γλωσσικά χαρακτηριστικά που αναφέρθηκαν στην προηγούμενη ενότητα.

Οι μηχανές κατηγοριοποίησης SVM σε συνδυασμό με τα χαρακτηριστικά TF-IDF έχουν χρησιμοποιηθεί με ιδιαίτερη επιτυχία σε εφαρμογές επεξεργασίας κειμένων (Drucker, H., Wu, D. & Vapnik, V. 1999) και θεωρούνται ως ένα κοινά αποδεκτό μέτρο σύγκρισης (state-of-the-art). Οι μηχανές SVM (Vapnik V. 1998) παρουσιάζουν κάποιες ομοιότητες ως προς τον τρόπο λειτουργίας τους με τα νευρωνικά δίκτυα MLP. Αρχικά εφαρμόζονται οι είσοδοι στο σύστημα και στη συνέχεια το διάνυσμα εισόδου μετασχηματίζεται σε ένα άλλο χώρο, μεγαλύτερης συνήθως διάστασης. Ο μετασχηματισμός αυτός γίνεται μέσω κάποιας μη γραμμικής συνάρτησης, της συνάρτησης *Φ*. Αφού μετασχηματιστεί το διάνυσμα εισόδου στην αυξημένη διάσταση, θεωρείται ότι τα δεδομένα είναι γραμμικά διαχωρίσιμα ανά τάξη. Μ' αυτόν τον τρόπο, στη συνέχεια υπάρχει ένα γραμμικό επίπεδο εξόδου, αντίστοιχο με ένα νευρώνα MLP με γραμμική συνάρτηση ενεργοποίησης. Το πρόσημο της εξόδου (θετικό ή αρνητικό) δείχνει σε ποια πλευρά της διαχωριστικής ευθείας ανήκει το πρότυπο, το οποίο ακολούθως αποδίδεται στην αντίστοιχη τάξη. Η εκπαίδευση των SVM δεν εκτελείται ακολουθιακά, όπως ο αλγόριθμος ανάστροφης διάδοσης, αλλά με βάση ένα κριτήριο βελτιστοποίησης υπό περιορισμούς και συγκλίνει σε κάποιο συγκεκριμένο σημείο με ντετερμινιστικό τρόπο.

Το μοντέλο SVM είναι σχεδιασμένο για προβλήματα δύο τάξεων. Για την αντιμετώπιση προβλημάτων περισσότερων τάξεων (multiclass), απαιτείται η χρήση συνδυαστικών τεχνικών μηχανών επιτροπής (committee machines). Δύο προτεινόμενες τεχνικές είναι (α) η κατηγοριοποίηση μίας τάξης έναντι όλων των άλλων (one against all – **OAA**) ή (β) ο διαχωρισμός των τάξεων ανά ζευγάρι (one against one - **OAO**).

Σύμφωνα με την πρώτη επιλογή (OAA) χρησιμοποιούνται *M* μοντέλα SVM, καθένα από τα οποία αναγνωρίζει δεδομένα που ανήκουν σε μία μόνο τάξη σε σχέση με όλες τις άλλες. Ο αριθμός *M* πρέπει να ταυτίζεται με τον αριθμό των τάξεων του προβλήματος, ενώ το κάθε πρότυπο αποδίδεται σε εκείνη την τάξη, της οποίας το αντίστοιχο SVM δίνει την υψηλότερη καταφατική έξοδο.

Η δεύτερη επιλογή (OAO) χρησιμοποιεί ένα δυαδικό κατηγοριοποιητή για κάθε ζεύγος τάξεων, γεγονός που μεταφράζεται σε συνολικά $M \bullet (M-1)/2$ μοντέλα SVM. Κάθε πρότυπο αποδίδεται τελικά σε εκείνη την τάξη την οποία επιλέγει η πλειονότητα των μοντέλων. Όπως είναι εύκολα αντιληπτό, η δεύτερη επιλογή (OAO) απαιτεί περισσότερα μοντέλα SVM, ενώ η πρώτη παρουσιάζει ασυμμετρία ως προς τα δεδομένα εκπαίδευσης, μιας και τα θετικά παραδείγματα για κάθε κατηγορία είναι πολύ λιγότερα από τα αρνητικά.

Για τα πειράματα που παρουσιάζονται στη συνέχεια χρησιμοποιήθηκε το μοντέλο SVM – C με Γκαουσσιανό πυρήνα από το πακέτο LIBSVM (Chang, C,-C. & Lin, C.-J. 2001). Για να συγκριθεί η απόδοση του προτεινόμενου συστήματος με αυτή του μοντέλου TF-IDF SVM, επιλέχθηκε ένας αριθμός όρων (terms) για τα TF-IDF ίσος με το πλήθος των γλωσσικών χαρακτηριστικών του συστήματος (85), και συγκεκριμένα τα *N* (*N*=85) πιο συχνά λήμματα, όπως προέκυψαν από τις μετρήσεις που έγιναν μετά την εφαρμογή του λημματοποιητή στα δεδομένα κείμενα (Papageorgiou H., Prokopidis P., Giouli V. & Piperidis S. 2000). Σε αυτά τα συγκριτικά πειράματα για την απόδοση των μοντέλων SVM – MLP χρησιμοποιήθηκαν και τα δύο είδη μετρήσεων, αυτών με βάση τα αρχικά γλωσσικά χαρακτηριστικά (OM – Original Measurements) του πίνακα 3.1 και αυτών με βάση τα TF-IDF. Το SVM σε συνδυασμό με τις 85 TF-IDF μεταβλητές, θα ήταν αναμενόμενο να δώσει υψηλότερα αποτελέσματα από το MLP με τις 85 μεταβλητές γλωσσικές μεταβλητές, αφού όπως φαίνεται από τα πειράματα που παρουσιάσθηκαν στο σχήμα 3.6 τα χαρακτηριστικά που βασίζονται σε λήμματα (όπως είναι και οι μεταβλητές TF-IDF) είναι ιδιαίτερα αποτελεσματικά συγκρινόμενα με τις υπόλοιπες ομάδες χαρακτηριστικών.

Η ακρίβεια κατηγοριοποίησης για διάφορους συνδυασμούς μοντέλων (SVM και MLP) και χαρακτηριστικών (OM και TF-IDF) παρουσιάζεται στον πίνακα 3.5:

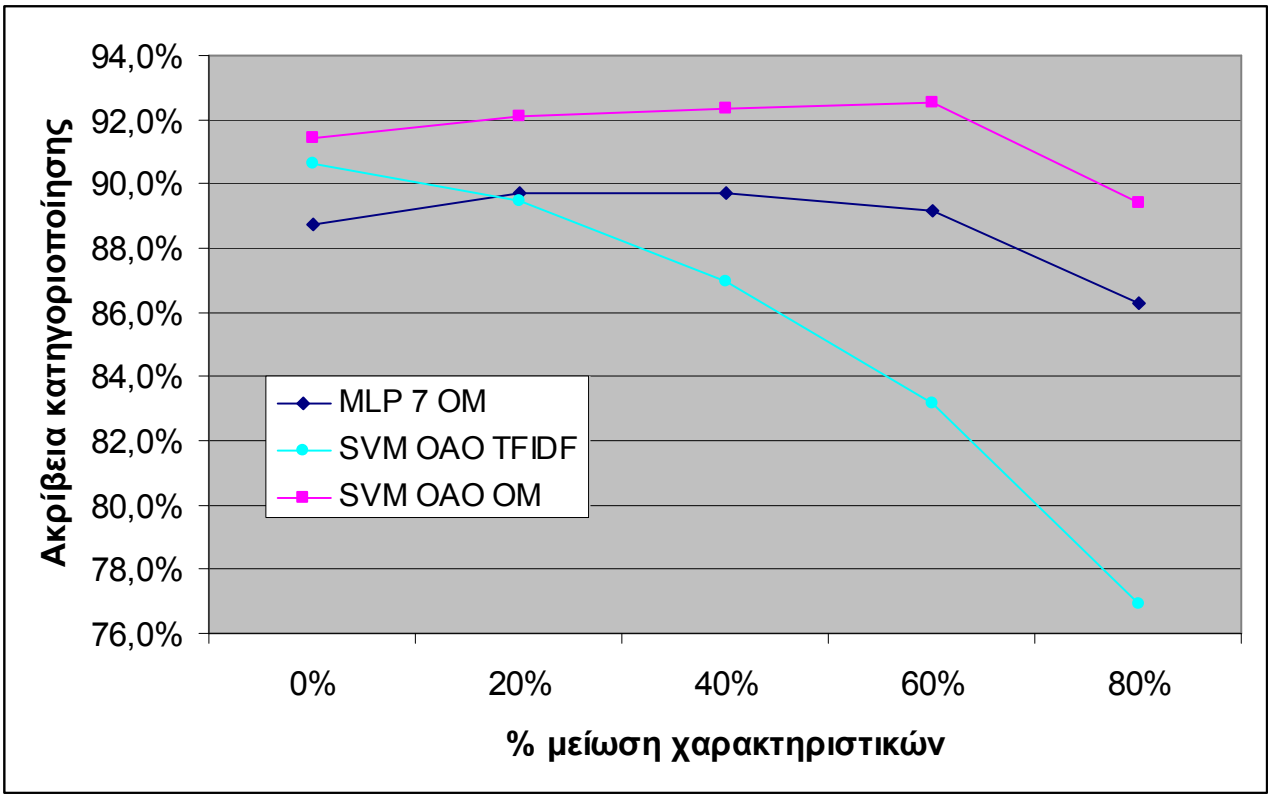
Μοντέλο	Σύνολο Χαρακτηριστικών	Ακρίβεια Κατηγοριοποίησης
MLP 5	ΟΜ	88,45%
MLP 6	ΟΜ	88,71%
MLP 7	ΟΜ	88,73%
SVM-OAA	ΟΜ	86,66%
SVM-OAO	ΟΜ	91,43%
SVM-OAA	TFIDF	86,55%
SVM-OAO	TFIDF	90,64%

Πίνακας 3.5. Ακρίβεια κατηγοριοποίησης για διάφορους συνδυασμούς μοντέλων (MLP και SVM) και χαρακτηριστικών

Σύμφωνα με τον πίνακα 3.5, η βέλτιστη ακρίβεια κατηγοριοποίησης παρατηρήθηκε για το μοντέλο SVM-OAO με το αρχικό σύνολο χαρακτηριστικών (ΟΜ), το οποίο παρουσιάζει μεγαλύτερη ακρίβεια από το μοντέλο MLP στις ίδιες μετρήσεις, αλλά και από το SVM με τις TF-IDF μετρήσεις. Επιπλέον, όπως αναμενόταν, εξαιτίας και του μεγαλύτερου αριθμού δυαδικών κατηγοριοποιητών, για τα μοντέλα SVM η OAO επιλογή συνδυασμού τους οδηγεί πάντα σε υψηλότερη ακρίβεια κατηγοριοποίησης από την OAA. Εύκολα συμπεραίνεται, λοιπόν, ότι οι κατηγοριοποιητές SVM-OAO αποδίδουν γενικά καλύτερα από τα MLP. Το μοντέλο TF-IDF SVM είναι καλύτερο από το προτεινόμενο MLP για το πλήρες σύνολο των χαρακτηριστικών (85). Πιο συγκεκριμένα, η ακρίβεια κατηγοριοποίησης για το TF-IDF SVM είναι 90.6%, δηλαδή περίπου 1,9% υψηλότερη από το καλύτερο MLP. Εξάλλου, το αρχικό σύνολο χαρακτηριστικών (ΟΜ) είναι καλύτερο από το γενικότερο σύνολο TF-IDF, αφού, όπως φαίνεται και στον πίνακα 3.5, ο συνδυασμός SVM-ΟΜ αποδίδει καλύτερα από το συνδυασμό SVM-TFIDF κατά 0,79%.

Η ακρίβεια κατηγοριοποίησης εξετάσθηκε και στην περίπτωση μείωσης της διάστασης του διανύσματος εισόδου, αφού, όπως προέκυψε, η μείωση της διάστασης μπορεί να επιφέρει βελτίωση των αποτελεσμάτων. Πιο συγκεκριμένα, ακολουθώντας αντίστοιχη διαδικασία και χρησιμοποιώντας το κριτήριο μέτρησης της αξίας των χαρακτηριστικών (3-36), μειώθηκε ο αριθμός αυτών για τις αρχικές μετρήσεις, ενώ για τις μετρήσεις TF-IDF απλά επιλέχθηκαν λιγότεροι όροι, ώστε τα αποτελέσματα να είναι συγκρίσιμα. Στο διάγραμμα του σχήματος 3.9 απεικονίζονται τα αποτελέσματα της σύγκρισης ως προς την ακρίβεια κατηγοριοποίησης, όταν μειώνεται το πλήθος των χαρακτηριστικών για το μοντέλο MLP με τις αρχικές μετρήσεις και εφτά νευρώνες στο κρυφό επίπεδο, καθώς επί-

σης και για τα καλύτερα μοντέλα SVM-OAO τόσο με τις αρχικές μετρήσεις όσο και με τις TF-IDF.



Σχήμα 3.9. Μεταβολή ακρίβειας κατηγοριοποίησης για τα SVM και τα MLP, καθώς μειώνεται το πλήθος των χαρακτηριστικών

Από το διάγραμμα 3.9 προκύπτει ότι το μοντέλο SVM, όταν επιλέγεται σε συνδυασμό με τις αρχικές μετρήσεις (OM), είναι πάντα καλύτερο από το MLP για κάθε βήμα μείωσης του διανύσματος εισόδου. Όταν το σύνολο των 85 χαρακτηριστικών μειωθεί κατά 20% ή και περισσότερο, το προτεινόμενο σύστημα MLP-OM είναι καλύτερο από το μέτρο σύγκρισης (state of the art) TF-IDF SVM. Το αποτέλεσμα αυτό μπορεί να σχετίζεται με το γεγονός ότι το αρχικό σύνολο των χαρακτηριστικών, που συγκροτήθηκε κατά το παρελθόν βάσει γλωσσολογικών επιλογών αλλά και της διεθνούς βιβλιογραφίας (Lowe D. & Matthews R. 1995, Tweedie F., Singh S. & Holmes D. 1996 και Tambouratzis G., Hairetakis N., Markantonatou S. & Carayannis G. 2003), ενδεχομένως εμπεριέχει μη πληροφοριακά στοιχεία. Καθώς το κριτήριο αξιολόγησης μειώνει τα στοιχεία αυτά, το προτεινόμενο μοντέλο παρουσιάζει μεγαλύτερη ακρίβεια κατηγοριοποίησης και ξεπερνά το SVM TF-IDF.

Ένα πολύ ενδιαφέρον συμπέρασμα είναι ότι με το συνδυασμό του κριτηρίου αξιολόγησης των χαρακτηριστικών για τη μέθοδο MLP (τύπος (3-21)) και του μοντέλου SVM μπορεί να βελτιωθεί η απόδοση του συστήματος. Ο αλγόριθμος εκπαίδευσης που χρησιμοποιήθηκε

(LMBR) είναι χρήσιμο εργαλείο για την αξιολόγηση των χαρακτηριστικών και ακολούθως για τη βελτίωση της απόδοσης του SVM.

Πιο συγκεκριμένα, όταν το σύνολο των χαρακτηριστικών μειώνεται κατά 60%, η ακρίβεια κατηγοριοποίησης του SVM βελτιώνεται στο 92,5% από το 91,4% για το πλήρες σύνολο. Η ακρίβεια 92,5% με μόλις 34 παραμέτρους είναι η μεγαλύτερη ακρίβεια κατηγοριοποίησης που έχει παρατηρηθεί για το συγκεκριμένο πρόβλημα στο σύνολο δεδομένων συγκριτικά προς οποιοδήποτε άλλο συνδυασμό μοντέλου χαρακτηριστικών.

Στην παρούσα ενότητα παρουσιάσθηκε ένα σύστημα κατηγοριοποίησης κειμένων ανάλογα με το ύφος του δημιουργού τους. Το συγκεκριμένο σύστημα επιτυγχάνει πολύ υψηλό βαθμό ακρίβειας, ενώ παρουσιάζει και μερικά επιπλέον ενδιαφέροντα στοιχεία. Το σύστημα έχει τη δυνατότητα να αποτιμά επιτυχώς το βαθμό συμβολής κάθε παραμέτρου στο τελικό αποτέλεσμα, δυνάμενο κατά συνέπεια να μειώσει τη διάσταση του προβλήματος, βελτιώνοντας ταυτοχρόνως την ακρίβεια κατηγοριοποίησης. Η χρήση του συστήματος σε συνδυασμό με πιο κλασσικές προσεγγίσεις μπορεί να οδηγήσει σε αρκετά βελτιωμένη απόδοση στο δεδομένο πρόβλημα.

ΚΕΦΑΛΑΙΟ 4ο

4.1 Σύστημα οργάνωσης κειμένων με βάση το περιεχόμενο

Το σύστημα οργάνωσης κειμένων ομαδοποιεί τα κείμενα σε θεματικές κατηγορίες ανάλογα με το περιεχόμενό τους. Το πρόβλημα της θεματικής οργάνωσης κειμένων είναι ως προς τη φύση του αρκετά διαφορετικό από το πρόβλημα αναγνώρισης ομιλητών, καθώς το δεύτερο πρόβλημα βασίζεται στην αναγνώριση του ύφους των κειμένων, ενώ το πρώτο εξαρτάται από το περιεχόμενό τους. Το περιεχόμενο των κειμένων αντικατοπτρίζεται κυρίως στις λέξεις που περιέχει και όχι τόσο σε δομικά, μορφολογικά ή γραμματικά χαρακτηριστικά. Κατά συνέπεια, η σχεδίαση ενός τέτοιου συστήματος βασίζεται πρωτίστως στις λέξεις.

Καθώς το πλήθος των διαφορετικών κλιτών μορφών στην ελληνική γλώσσα είναι αρκετά μεγάλο, λόγω της πλούσιας μορφολογίας, είναι πολύ σημαντικό το σύστημα να επεξεργάζεται κείμενα που έχουν προηγουμένως λημματοποιηθεί. Η διαδικασία της λημματοποίησης εκτελείται χρησιμοποιώντας το εργαλείο Tagger-Lemmatiser του ΙΕΛ (Papageorgiou H., Prokopidis P., Giouli V. & Piperidis S. 2000), το οποίο μειώνει δραστικά το πλήθος των διαφορετικών κλιτών μορφών σε ένα σώμα κειμένων, καθώς τις ανάγει στην αντίστοιχη λημματική τους μορφή. Με αυτό τον τρόπο, το σύστημα αναγνώρισης περιεχομένου μπορεί να βασισθεί σε χαρακτηριστικά μέτρησης εμφανίσεων λημμάτων.

Το σύστημα που αναπτύχθηκε αποτελείται από δύο βασικά στάδια:

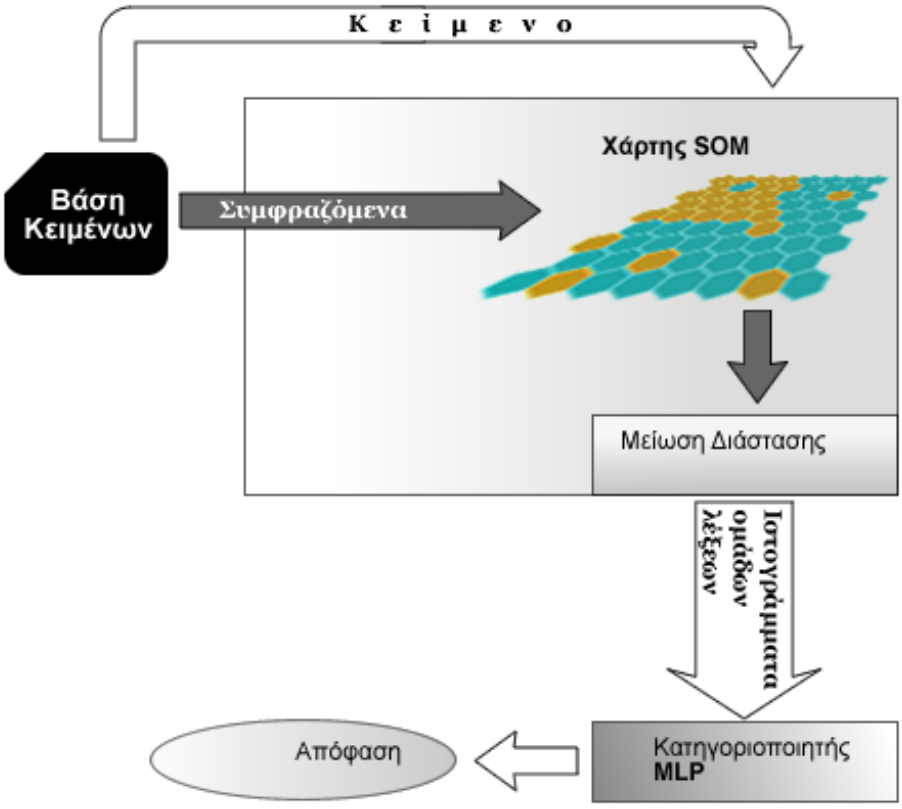
1. Το πρώτο στάδιο αφορά στη δημιουργία ενός **χάρτη λέξεων**, παρόμοιου με εκείνον που έχει προταθεί από τους (Pullwitt D., 2002) (Kohonen T., 1982), ο οποίος συνιστά το σύνολο των χαρακτηριστικών για την απεικόνιση των κειμένων στο χώρο προτύπων.
2. Στο δεύτερο στάδιο λαμβάνεται η **απόφαση-συμπέρασμα** για τη θεματική κατηγορία του κάθε προτύπου.

Πιο αναλυτικά, το σύστημα περιλαμβάνει τρία επιμέρους υποσυστήματα:

- **Το υποσύστημα του αρχικού χάρτη λέξεων** (word map): Με τη χρήση ενός χάρτη κατασκευασμένου από το νευρωνικό δίκτυο SOM δημιουργείται χωρίς επίβλεψη ένας χάρτης λέξεων.

- **Το υποσύστημα μείωσης διάστασης** (dimensionality reduction module): Οι ομάδες των λέξεων, ανάλογα και με την τοπολογία του χάρτη, συμπυκνώνονται περισσότερο με κάποιο κοινό αλγόριθμο ομαδοποίησης, ώστε να μειωθεί ο αριθμός των ομάδων και να σχηματισθούν κατά συνέπεια μεγαλύτερες ομάδες (Vesanto J., 2000).
- **Το υποσύστημα κατηγοριοποίησης**: Χρησιμοποιείται κάποιος κατηγοριοποιητής που εκπαιδεύεται με επίβλεψη. Στη συγκεκριμένη περίπτωση, ως κατηγοριοποιητής επιλέχθηκε ένα νευρωνικό δίκτυο MLP.

Τα δύο πρώτα υποσυστήματα υπάγονται στο πρώτο στάδιο της διαδικασίας, που περιλαμβάνει την εξαγωγή χαρακτηριστικών και την αναπαράσταση των κειμένων στο χώρο προτύπων. Το τελευταίο στάδιο της διαδικασίας, δηλαδή η λήψη απόφασης, υλοποιείται με το υποσύστημα του κατηγοριοποιητή (σχήμα 4.1).



Σχήμα 4.1. Συνοπτικό διάγραμμα συστήματος οργάνωσης κειμένων με βάση το περιεχόμενο

Για την απεικόνιση των κειμένων στο χώρο προτύπων, απαραίτητη προϋπόθεση συνιστά η χρήση ενός συνόλου μετρήσιμων χαρακτηριστικών. Σε παλαιότερες εφαρμογές του SOM σε προβλήματα οργάνωσης κειμένων όπως το WEBSOM (Lagus K., Kaski S. & Kohonen T., 2004, και K Kohonen T., Kaski S., Lagus K., Salojarvi, Honkela J., Patero V. & Saarela A., 2000) κάθε κείμενο αναπαρίσταται με ένα διάνυσμα, κάθε συνιστώσα του οποίου ισούται με τη συχνότητα κάποιου λήμματος. Εξαιτίας του πολύ μεγάλου αριθμού λημμάτων που υπάρχουν σε ένα σώμα κειμένων, η διάσταση του διανύσματος είναι εξίσου μεγάλη. Δυστυχώς, η αυξημένη αυτή διάσταση δημιουργεί μεγάλες υπολογιστικές απαιτήσεις αλλά και απαιτήσεις μνήμης, καθιστώντας την υλοποίησή του συστήματος προβληματική. Για το λόγο αυτό, συχνότερα επιλέγονται ως τρόποι μείωσης της διάστασης αυτής είτε η διαγωνοποίηση (SVD) είτε κάποια υπολογιστικά απλούστερη μέθοδος τυχαίας προβολής σε άξονες (Random projection method (Kaski S., 1998)).

4.1.1 Δημιουργία χάρτη λέξεων

Στην παρούσα ενότητα περιγράφεται η δημιουργία κατηγοριών με λέξεις-λήμματα ανάλογα με τη θεματική τους κατηγορία. Πρέπει εδώ να τονιστεί ότι δεν υπάρχει πρότυπη οργάνωση και ότι ενδεχομένως να υπάρχουν ποικίλες απόψεις σχετικά με την απόδοση των λέξεων σε κάποια κατηγορία. Μία λέξη μπορεί να ανήκει σε διαφορετικές κατηγορίες ανάλογα με τα συμφραζόμενα (Pullwitt D., 2002 και Kohonen T., 1982) ή να ανήκει γενικά σε περισσότερες της μίας κατηγορίες. Κατά συνέπεια, δεν είναι απλή η αξιολόγηση των αποτελεσμάτων, αφού δεν είναι εύκολο να δοθεί μία αριθμητική τιμή σε ένα αποτέλεσμα που – τις περισσότερες φορές – κρίνεται υποκειμενικά. Από την άλλη μεριά, υπάρχουν και λέξεις για τις οποίες είναι σαφές ότι σε οποιαδήποτε συμφραζόμενα δε σχετίζονται νοηματικά.

Όπως έχει ήδη αναφερθεί, καθίσταται αναγκαία η αναπαράσταση των λέξεων, που συνιστούν συμβολικά δεδομένα, με ένα διάνυσμα χαρακτηριστικών στο χώρο προτύπων. Τα χαρακτηριστικά αυτά εκφράζουν το περιβάλλον εμφάνισης της εκάστοτε λέξης, θεωρώντας ως περιβάλλον εμφάνισης τις άλλες λέξεις με τις οποίες συνεμφανίζεται. Τέλος, θα πρέπει να τονισθεί ξανά ότι στην παρούσα υλοποίηση αντικείμενο επεξεργασίας αποτέλεσαν μόνο λήμματα, συνεπώς όπου αναφέρεται ο όρος "λέξη" εννοείται το λήμμα ενός συνόλου λεξικών μορφών.

4.1.1.1 Ανάλυση ABC – Επιλογή συμφραζομένων

Αρχικά, τα λήμματα οργανώνονται σε ομάδες πάνω σε ένα χάρτη, τον οποίο δημιουργεί ένα δίκτυο τύπου SOM. Η διάσταση των διανυσμάτων των χαρακτηριστικών που επιλέγονται για να εκφράσουν κάθε λήμμα είναι πολύ μικρότερη από το πλήθος των λημμάτων στο σύνολο των δεδομένων. Ιδανικά, θα ήταν επιθυμητό για κάθε λήμμα να μετράται ο αριθμός των συνεμφανίσεών του με όλα τα άλλα τα λήμματα. Κάτι τέτοιο όμως θα αύξανε σημαντικά τη διάσταση του διανύσματος εισόδου και θα καθιστούσε το σύστημα υπολογιστικά δυσκίνητο και την υλοποίησή του ανέφικτη. Για το λόγο αυτό επιλέχθηκε ένα υποσύνολο των λημμάτων (λήμματα-χαρακτηριστικά), ώστε κάθε λήμμα αναπαραστάθηκε βάσει των συνεμφανίσεών του με το συγκεκριμένο υποσύνολο. Η επιλογή των λημμάτων έγινε χρησιμοποιώντας έναν εμπειρικό κανόνα προσαρμοσμένο στα φαινόμενα που διέπουν τη γλώσσα (Bell T. C., Cleary J. G., & Witten I. H., 1990, MacKay D. 2003 και Manning C. D. & Schütze H. 1999).

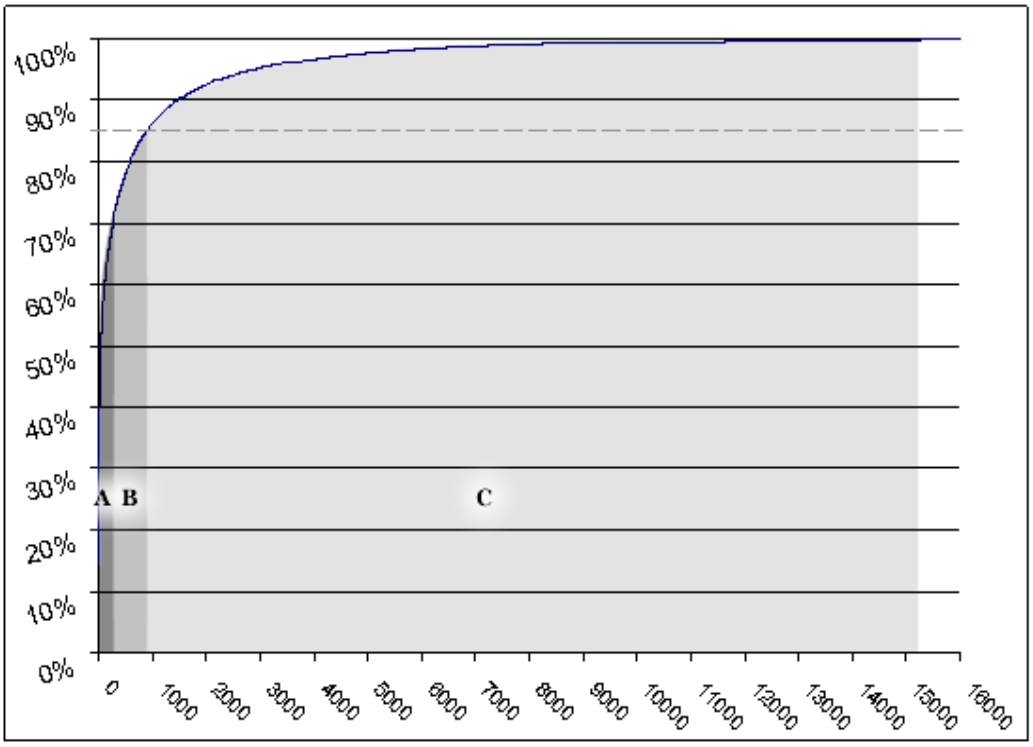
Πιο αναλυτικά, η συνεισφορά σε επίπεδο εμφανίσεων για κάθε λήμμα σε ένα σύνολο κειμένων ακολουθεί την κατανομή του Zipf, (MacKay D. 2003 και Manning C. D. & Schütze H. 1999) δηλαδή, μόνο ένα πολύ μικρό σύνολο λημμάτων εμφανίζεται πάρα πολύ συχνά, ενώ τα περισσότερα λήμματα είναι πιο σπάνια. Σε συμφωνία με την κατανομή του Zipf βρίσκεται η αρχή του Παρέτο (Tanwari A., Lakhiar, Q.A., & Shaikh G.Y. 2000 και Cam, R. and Gilles, L. 2002), η οποία είναι γνωστή και ως κανόνας του 80-20 και υποστηρίζει το εξής: «το 20% των αιτίων είναι υπεύθυνο για το 80% των αποτελεσμάτων». Η αρχή του Παρέτο, γνωστή και ως ανάλυση ABC, έχει εφαρμοσθεί με επιτυχία στη διοίκηση επιχειρήσεων και στον ποιοτικό έλεγχο (Tanwari A., Lakhiar, Q.A., & Shaikh G.Y. 2000 και Cam, R. and Gilles, L. 2002) και σύμφωνα με αυτή ένα ποσοστό των αιτίων χαρακτηρίζεται ως A, κάτι που σημαίνει ότι έχουν αρκετά μεγάλο αντίκτυπο, ενώ τα υπόλοιπα αίτια χαρακτηρίζονται ως B και C κατά φθίνουσα σημασία.

Στο προκείμενο πρόβλημα, ως A χαρακτηρίζονται τα ιδιαιτέρως συχνά εμφανιζόμενα λήμματα, τα οποία συνήθως είναι λειτουργικές λέξεις όπως άρθρα, βοηθητικά ρήματα και σύνδεσμοι, τα οποία δε συμβάλλουν ιδιαίτερα στο περιεχόμενο. Σε αυτό το σημείο θα πρέπει να γίνει σαφές ότι η ανάλυση ABC ταξινομεί αρχικά τις λέξεις ανάλογα με το πλήθος των εμφανίσεων τους και όχι ανάλογα με τη σημαντικότητά τους για το πρόβλημα της κατηγοριοποίησης. Συνεπώς, οι σημαντικές ως προς τις εμφανίσεις τους λέξεις A δεν είναι απαραίτητα και οι σημαντικότερες ως χαρακτηριστικά, λέξεις του τελικού συστήματος. Στην κατηγορία B υπάγονται λήμματα, τα οποία χρησιμοποιούνται αρκετά συχνά, ενώ τα πιο σπάνια λήμματα χαρακτηρίζονται ως C.

Από τις τρεις αυτές ομάδες φαίνεται ότι καταλληλότερη ως σύνολο χαρακτηριστικών είναι η κατηγορία Β, αφού τα λήμματά της δεν ανήκουν στην κατηγορία των λειτουργικών λέξεων (η οποία εμπίπτει, όπως αναμένεται άλλωστε, περισσότερο στην κατηγορία Α), ενώ είναι αρκετά συχνά ώστε να έχουν αρκετές συνεμφανίσεις με τα περισσότερα λήμματα του σώματος κειμένων και να μπορούν να τα περιγράψουν χωρίς να προκύπτουν πολλές μηδενικές μετρήσεις. Τα όρια των εμφανίσεων για την ανάλυση ABC τίθενται ως εξής (βλέπε και σχήμα 4.2):

- Στην **κατηγορία Α** ανήκουν τα πιο συχνά λήμματα, τα οποία καλύπτουν το 70% του συνολικού αριθμού εμφανίσεων όλων των λημμάτων. (Ενδεικτικά αναφέρεται, όπως θα ειπωθεί αναλυτικότερα στην ενότητα των πειραμάτων, ότι στην κατηγορία αυτή ανήκουν 257 λήμματα από το σώμα κειμένων της Βουλής.)
- Στην **κατηγορία Β** ανήκουν τα λήμματα, τα οποία είναι λιγότερο συχνά και αντιστοιχούν στο επόμενο 15% των εμφανίσεων από 70% έως 85%. (Στην προκείμενη κατηγορία προκύπτουν 634 λήμματα.)
- Στην **κατηγορία C** ανήκουν και τα περισσότερα λήμματα, τα οποία δεν είναι συχνά και αντιστοιχούν στο υπόλοιπο 15% των εμφανίσεων. (Στην κατηγορία αυτή ανήκουν 16.197 λήμματα.)

Από την κατηγορία C αφαιρέθηκαν τα λήμματα με λιγότερες από 3 εμφανίσεις στο σώμα κειμένων, αφού λόγω της σπανιότητάς τους απλά επιβάρυναν τους αλγόριθμους εκπαίδευσης, χωρίς να προσφέρουν σημαντικό βαθμό κάλυψης για το σύνολο των δεδομένων. Επιπλέον, με τη διαδικασία αυτή εξασφαλίζεται ότι τυχαία ορθογραφικά λάθη δε λαμβάνονται υπόψιν ως ξεχωριστά λήμματα. Αποτέλεσμα αυτής της διαδικασίας συνιστά σημαντική μείωση των λημμάτων στην κατηγορία αυτή (ενδεικτικός αριθμός: 6.987 λήμματα). Το συνολικό πλήθος των λημμάτων που θα απεικονισθεί στο χάρτη είναι αυτά των κατηγοριών Α και C. Οι συχνότητες συνεμφανίσεων κάθε λήμματος με τα λήμματα της κατηγορίας Β χρησιμοποιούνται ως το διάνυσμα χαρακτηριστικών για την οργάνωση των λέξεων.



Σχήμα 4.2. Διάγραμμα ανάλυσης ABC για τα λήμματα στο σώμα κειμένων της Βουλής

Πρέπει να επισημανθεί ότι για τη μέτρηση των χαρακτηριστικών για κάθε λήμμα λαμβάνονται υπόψιν οι συνεμφανίσεις σε επίπεδο περιόδου, ενώ κάθε εμφάνιση λήμματος ανά περίοδο, ακόμα και αν επαναλαμβάνεται, μετριέται μόνο μία φορά. Η επιλογή αυτή βασίζεται στο γεγονός ότι η τελεία χωρίζει δύο περιόδους και είναι το σημείο εκείνο που μία ιδέα ή μια άποψη τερματίζεται με σαφήνεια, αφού είναι πιθανό μία επόμενη περίοδος να πραγματεύεται κάτι αρκετά διαφορετικό νοηματικά, ιδιαίτερα όταν αυτή μπορεί να ανήκει σε διαφορετική παράγραφο.

Η ανάλυση ABC διεξάγεται ως εξής: αρχικά, μετράται ο αριθμός των εμφανίσεων κάθε λήμματος στο σύνολο των δεδομένων. Οι συχνότητες αυτές κανονικοποιούνται, ώστε να αθροίζουν στη μονάδα, και τα λήμματα ταξινομούνται σε φθίνουσα σειρά συχνότητας. Στη συνέχεια, επιλέγεται το πρώτο λήμμα χωρίς αντικατάσταση σε κάθε βήμα ξεκινώντας από το πιο συχνό. Αρχικά τα λήμματα εντάσσονται στην κατηγορία A, έως ότου το άθροισμα των συχνοτήτων να υπερβεί το κατώφλι που έχει ορισθεί για την κατηγορία αυτή. Όταν αυτό το άθροισμα κυμαίνεται στο εύρος μεταξύ των κατωφλίων A και B, τα αντίστοιχα λήμματα αποδίδονται στην κατηγορία B, ενώ τα υπόλοιπα αποδίδονται στην κατηγορία C.

Οι συνιστώσες των διανυσμάτων συνεμφανίσεων υπολογίζονται με τον ακόλουθο τύπο (4-1):

$$s_f(w_i) = \frac{N(w_i, f)}{N(w_i)} \quad (4\text{-}1)$$

όπου $s_f(w_i)$ είναι η συνιστώσα για το λήμμα-χαρακτηριστικό *f* του λήμματος w_i, $N(w_i,f)$ είναι το πλήθος των ταυτόχρονων εμφανίσεων του λήμματος w_i και του χαρακτηριστικού *f* και $N(w_i)$ είναι το πλήθος των προτάσεων που περιλαμβάνουν το λήμμα w_i. Κάθε διάνυσμα κανονικοποιείται, έτσι ώστε οι συνιστώσες του να αθροίζουν στη μονάδα και βάσει αυτού ο τύπος (4-1) τροποποιείται ως εξής (4-2):

$$\hat{s}_f(w_i) = \frac{s_f(w_i)}{\sum_{forall\_f} s_f(w_i)} \quad (4\text{-}2)$$

όπου με $\hat{s}_f(w_i)$ συμβολίζεται η κανονικοποιημένη παραλλαγή.

Μετά τη διανυσματική αναπαράσταση όλων των λημμάτων, εφαρμόζεται ο ιδιαίτερα αποτελεσματικός αλγόριθμος ομαδοποίησης SOM, ο οποίος έχει κατά το παρελθόν εφαρμοσθεί με επιτυχία σε πειράματα οργάνωσης λέξεων για άλλου είδους εφαρμογές (Pullwitt D., 2002, Kohonen T., 1982 και Freeman R.T. & Yin. H. 2005).

4.1.1.2 Αλγόριθμος Ομαδοποίησης SOM

Ένα αποδοτικό μοντέλο ομαδοποίησης και οπτικοποίησης δεδομένων με αρκετά καλή συμπεριφορά σε μεγάλο αριθμό διαστάσεων είναι το νευρωνικό δίκτυο τύπου SOM (Kohonen T., 1997). Τα δίκτυα αυτοοργανούμενου τύπου (Self Organising Map – SOM) αποτελούνται από ένα και μόνο επίπεδο, στου οποίου τους νευρώνες εφαρμόζονται απευθείας οι είσοδοι χωρίς τη μεσολάβηση κάποιου βάρους. Βάρη υπάρχουν και σε αυτό το μοντέλο αλλά λειτουργούν διαφορετικά από ό,τι στο MLP. Ειδικότερα δε, κάθε νευρώνας έχει τόσα βάρη όσες είναι οι είσοδοι, δηλαδή κάθε είσοδος συνδέεται με κάθε νευρώνα μέσω ενός βάρους. Βάσει των παραχθέντων διανυσμάτων κάθε λήμμα αποδίδεται στην ομάδα εκείνη που έχει πιο «όμοιο» κέντρο. Από τη διαδικασία της εκπαίδευσης έχουν αφαιρεθεί τα λήμματα της Β κατηγορίας εξαιτίας του γεγονότος ότι αυτά «συνεμφανίζονται» με πολύ μεγάλη συχνότητα με τον εαυτό τους, γεγονός που θα μπορούσε να επηρεάσει δυσανάλογα την εκπαίδευση και τη δημιουργία του χάρτη.

Τα νευρωνικά δίκτυα τύπου SOM διαφέρουν σημαντικά από τα δίκτυα τύπου MLP ως προς τον τρόπο εκπαίδευσης και λειτουργίας. Στα δίκτυα SOM ενεργοποιείται κατά την εκπαίδευση ο νευρώνας, οι τιμές των βαρών του οποίου βρίσκονται κοντύτερα στις τιμές της εισόδου (NN-Rule) με βάση κάποια απόσταση όπως είναι η Ευκλείδεια ή το εσωτερικό γινόμενο. Όταν το ζητούμενο είναι η επίλυση ενός προβλήματος κατηγοριοποίησης, το άγνωστο πρότυπο αποδίδεται στην κατηγορία που αντιπροσωπεύει ο νικητής νευρώνας, δηλαδή ο νευρώνας, οι τιμές των βαρών του οποίου βρίσκονται κοντύτερα στο διάνυσμα εισόδου.

Πιο αναλυτικά, όταν παρουσιάζονται δεδομένα στην είσοδο, αρχικά εντοπίζεται ο νευρώνας που μοιάζει περισσότερο σε αυτά, όπως προαναφέρθηκε. Στη συνέχεια ανανεώνονται οι τιμές-παράμετροι του νικητή νευρώνα, ώστε να μοιάσουν ακόμη περισσότερο στο πρότυπο εισόδου. Αυτό το είδος της εκπαίδευσης χαρακτηρίζεται ως ανταγωνιστική μάθηση (competitive learning) χωρίς επίβλεψη (unsupervised). Εκτός όμως από το νικητή νευρώνα ανανεώνονται και οι τιμές των γειτονικών του νευρώνων σε φθίνοντα βαθμό, ώστε με την αύξηση της απόστασης από το νικητή, οι ανανεώσεις να είναι μικρότερες. Το τελευταίο αυτό στοιχείο εισάγει την έννοια της **γειτονίας**, σημαντικότατη καινοτομία του μοντέλου SOM.

Σε ένα τέτοιο δίκτυο υπάρχουν διάφορα είδη γειτονίας, με πιο συνηθισμένα το τετραγωνικό και το εξαγωνικό πλέγμα. Πιο αναλυτικά, το δίκτυο απαρτίζεται από ένα σύνολο νευρώνων με αρχικές τιμές c_i και διάσταση ίση με τη διάσταση των δεδομένων εκπαίδευσης, οι οποίοι τοποθετούνται στις κορυφές κάποιου προκαθορισμένου πλέγματος. Η αρχική τοποθέτηση των νευρώνων μπορεί να γίνει με διάφορους τρόπους, όπως θα περιγραφεί στη συνέχεια. Εάν το *i* διάνυσμα εκπαίδευσης συμβολιστεί με $\boldsymbol{x}_i$, τότε ο κοντινότερος νευρώνας στο χώρο, δηλαδή ο κοντινότερος στο δεδομένο εκπαίδευσης γείτονας βάσει της Ευκλείδειας απόστασης πλησιάζει ακόμα περισσότερο το δεδομένο εκπαίδευσης σύμφωνα με τον τύπο (4-3):

$$\left\| x - w_c \right\| = \min_i \left\| x - w_i \right\| \tag{4-3}$$

όπου w_i είναι το διάνυσμα του *i* νευρώνα διάστασης ίσης με την είσοδο $\boldsymbol{x}$, ενώ το *c* συμβολίζει το νικητή νευρώνα.

Σε αυτή την περίπτωση, ο κανόνας εκπαίδευσης εκφράζεται από την ακόλουθη σχέση:

$$\Delta w_i = \eta(n) h_{i,c(x)} (x - w_i) \tag{4-4}$$

Εδώ θα πρέπει να τονιστεί το σημείο διαφοροποίησης του SOM από τους κλασσικούς αλγόριθμους ανταγωνιστικής μάθησης, δηλαδή το ότι δεν «εκπαιδεύεται» μόνο ο νικητής, αλλά και οι κοντινοί σε αυτόν βάσει της θέση τους στο πλέγμα, όπου $h_{i,c(x)}$ είναι μία συνάρτηση (**συνάρτηση γειτονίας**) που δείχνει πώς ανανεώνονται οι τιμές σε όλο το δίκτυο με κέντρο το νικητή και *n* είναι ο αύξων αριθμός της τρέχουσας εποχής. Η συνάρτηση αυτή διαμορφώνεται κάθε φορά με κέντρο το νικητή, ο οποίος με τη σειρά του εξαρτάται από την είσοδο. Συνήθως η συνάρτηση έχει τη μορφή συνάρτησης Gauss (4-5):

$$h_{i,c(x)} = exp(-\frac{d_{i,c(x)}^{2}}{2\sigma^{2}}) \quad (4\text{-}5)$$

όπου $d_{i,c(x)}$ η απόσταση των νευρώνων $c(x)$ και i στο πλέγμα, η οποία δεν αλλάζει και υπολογίζεται από τη γεωμετρία του πλέγματος. Με $c(x)$ συμβολίζεται ο νικητής που αντιστοιχεί στην είσοδο x.

Συνήθως ο ρυθμός εκπαίδευσης $\eta(n)$ επιλέγεται να μεταβάλλεται μειούμενος με την πάροδο των εποχών και ο αλγόριθμος τελικά να συγκλίνει. Οπότε:

$$\eta(n) = \eta_o \, exp(-\frac{n}{\tau_1}) \quad (4\text{-}6)$$

όπου η_o η αρχική τιμή του ρυθμού εκπαίδευσης και τ_1 η παράμετρος που καθορίζει το ρυθμό μείωσης του ρυθμού εκπαίδευσης.

Επίσης συνηθίζεται, καθώς εξελίσσεται η διαδικασία της εκπαίδευσης, να μειώνεται το εύρος της γειτονίας και η ακτίνα σ, οπότε προκύπτει αντίστοιχα:

$$\sigma(n) = \sigma_o \, exp(-\frac{n}{\tau_1}) \quad (4\text{-}7)$$

Αξίζει να σημειωθεί ότι η σύγκλιση του μοντέλου έχει αποδειχθεί μόνο για πλέγμα μίας διάστασης (Kohonen T., 1997).

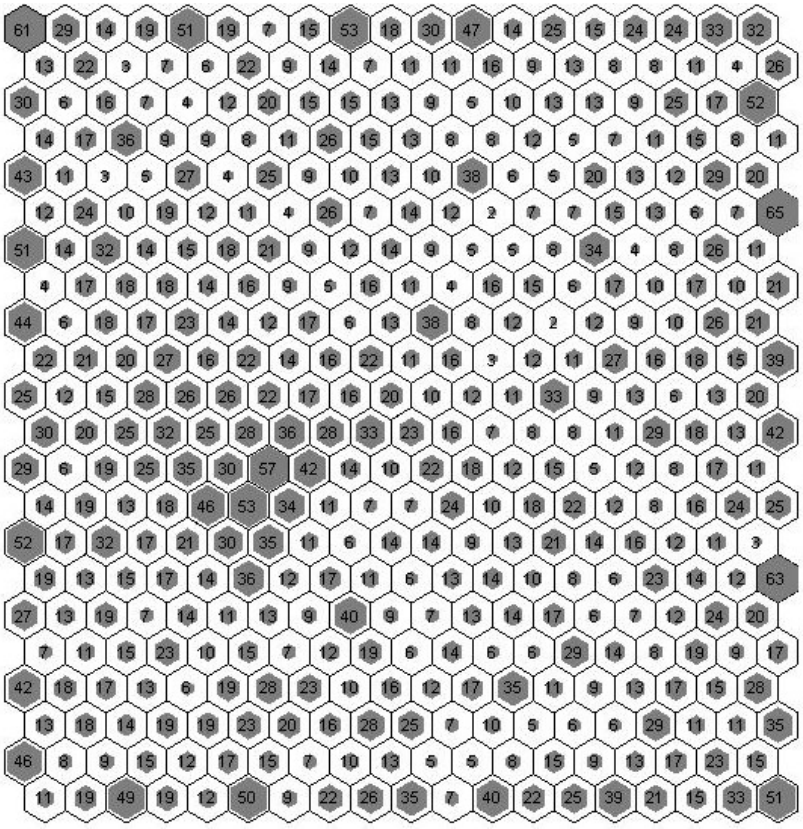
Η αρχικοποίηση του SOM μπορεί να γίνει με δύο τρόπους: είτε βάσει κάποιας κατανομής (π.χ. κανονική κατανομή) είτε με «γραμμική» αρχικοποίηση βασισμένη στους δύο κυριότερους άξονες, όπως αυτοί προκύπτουν με χρήση Ανάλυσης Πρωτευουσών Συνιστωσών (Principal Component Analysis – PCA). Σύμφωνα με το δεύτερο τρόπο, από τον πίνακα αυτοσυσχέτισης επιλέγονται εκείνα τα ιδιοδιανύσματα (eigenvectors) στα οποία αντιστοιχούν οι δύο μεγαλύτερες ιδιοτιμές (εκείνες δηλαδή που «περιέχουν» το μεγαλύτερο μέρος της πληροφορίας), και οι δύο διαστάσεις αρχικοποιούνται γραμμικά, ώστε από δεξιά προς αριστερά να αυξάνεται από την ελάχιστη στη μέγιστη η τιμή του πρώτου ιδιοδιανύσματος και αντίστοιχα από πάνω προς τα κάτω του δεύτερου.

Μελετώντας την κλασσική εξίσωση μάθησης του SOM (4-4) είναι εύκολο να παρατηρήσει κανείς ότι σε κάθε παρουσίαση ενός δεδομένου εκπαίδευσης το δίκτυο SOM τροποποιείται. Συμπεραίνουμε, λοιπόν, ότι η σειρά εμφάνισης των δεδομένων εκπαίδευσης διαδραματίζει σημαντικό ρόλο στην τελική κατάσταση του δικτύου. Μία συχνή παραλλαγή του αλγόριθμου τύπου SOM είναι η ανανέωση των βαρών όχι με το πέρασμα καθενός από τα δεδομένα εκπαίδευσης, αλλά η συνολική μεταβολή αυτών μετά την παρουσίαση του συνόλου εκπαίδευσης (batch training). Αυτή η διαδικασία, αν και πιο απαιτητική σε μνήμη, εξασφαλίζει ότι η εκπαίδευση δεν επηρεάζεται από τη σειρά παρουσίασης των δεδομένων στο δίκτυο. Πιο συγκεκριμένα τα κέντρα του χάρτη ενημερώνονται από τον παρακάτω τύπο (4-8):

$$w_i = \frac{\sum_j h_{i,c(x_j)} \cdot x_j}{\sum_j h_{i,c(x_j)}} \qquad (4\text{-}8)$$

Με βάση τον αλγόριθμο τύπου SOM έχουν υλοποιηθεί μία σειρά από παραλλαγές, οι οποίες αποσκοπούν σε βελτιωμένη απόδοση σε σχέση με τον κλασσικό αλγόριθμο SOM είτε σε επίπεδο ακρίβειας, είτε ταχύτητας εκτέλεσης, είτε εστίασης σε κάποιο άλλο στόχο της εκάστοτε εφαρμογής (Georgakis A., Kotropoulos C., Xafopoulos A., & Pitas I., 2004 και Rauber A., Merkl D., Dittenbach M., 2002).

Σύμφωνα με τις ιδιότητες του SOM αναμένεται ότι οι λέξεις που αντιστοιχίζονται στις ίδιες ή και σε γειτονικές ομάδες θα έχουν αντίστοιχα συμφραζόμενα και πιθανότατα θα συνδέονται νοηματικά. Μία ενδεικτική κατανομή λέξεων στο χάρτη απεικονίζεται στο σχήμα (4.3). Δεν περιέχουν όλες οι ομάδες του SOM το ίδιο πλήθος λέξεων και γενικότερα υπάρχει μία ανομοιομορφία ως προς την κατανομή των λέξεων στο χάρτη. Στο χάρτη του σχήματος (4.3) υπάρχουν ορισμένες κατηγορίες που περιέχουν μεγάλο πλήθος λέξεων, οι οποίες συγκεντρώνονται κυρίως στις άκρες του χάρτη, καθώς επίσης και σε μία περιοχή στο εσωτερικό του χάρτη και συγκεκριμένα κάτω αριστερά από το κέντρο του.



Σχήμα 4.3. Παράδειγμα κατανομής λέξεων στο χάρτη SOM

Πρέπει να σημειωθεί ότι το αντίστοιχο σύστημα δημιουργίας χαρτών που παρουσιάσθηκε στο (Georgakis A., Kotropoulos C., Xafopoulos A., & Pitas I., 2004) διαφέρει σημαντικά στον τρόπο εξαγωγής των χαρακτηριστικών των λέξεων. Ειδικότερα, το συγκεκριμένο σύστημα αποδίδει σημασία στη σειρά συνεμφάνισης των λέξεων και χρησιμοποιεί ένα μικρό παράθυρο λέξεων, μίας προηγούμενης λέξης και μίας επόμενης από τη λέξη που χρειάζεται να περιγραφεί. Η προσέγγιση αυτή δίνει μεγάλο βάρος στη διάταξη των όρων, στοιχείο που είναι επιθυμητό για γλώσσες όπως η Αγγλική αλλά ενδεχομένως να είναι περιττό δεδομένης της ελεύθερης σειράς των όρων στη Νέα Ελληνική. Επιπλέον η εν λόγω προσέγγιση απαιτεί διανύσματα διάστασης *2N*, όπου *N* είναι το σύνολο των λημμάτων στα οποία εκπαιδεύεται ο αλγόριθμος και είναι της τάξεως δεκάδων χιλιάδων, σε αντίθεση με το προτεινόμενο σύστημα που έχει διάσταση ίση με το πλήθος των λημμάτων της Β κατηγορίας, που ανέρχεται μόλις σε λίγες εκατοντάδες. Κατά συνέπεια, η προτεινόμενη μέθοδος έχει υπολογιστικά μικρότερες απαιτήσεις και μπορεί να εφαρμοστεί χωρίς τροποποιήσεις σε μεγαλύτερες συλλογές κειμένων. Επιπλέον, είναι περισσότερο προσαρμοσμένη στην Ελληνική γλώσσα και είναι σε θέση να αντεπεξέλθει στον πολύ μεγάλο αριθμό λημμάτων που εμφανίζονται στις συλλογές κειμένων.

Το μοντέλο SOM έχει τη δυνατότητα να μειώνει δραστικά τη διάσταση των δεδομένων και να τα προβάλλει σε δισδιάστατο πλέγμα (Kohonen T., 1997). Προκειμένου, ωστόσο, να προβεί σε τόσο δραστική συμπύκνωση πληροφορίας, θα πρέπει να διαθέτει αρκετούς νευρώνες. Συνηθισμένη πρακτική σε εφαρμογές ομαδοποίησης SOM αποτελεί η χρήση πολύ μεγαλύτερου αριθμού ομάδων από τις αναμενόμενες. Επιπλέον, αφού ο αλγόριθμος του SOM λειτουργεί χωρίς επίβλεψη, είναι πολύ πιθανό οι γειτονικές ομάδες του χάρτη να αναφέρονται ουσιαστικά στην ίδια επιθυμητή ομάδα. Η αύξηση των νευρώνων στο SOM δεν οδηγεί στα ίδια αρνητικά αποτελέσματα όπως το MLP, απλά μπορεί να δημιουργήσει κάποιες αραιές και άδειες ομάδες.

Επιπροσθέτως, η αύξηση των ομάδων μπορεί να οδηγήσει σε αύξηση της πολυπλοκότητας του υπολοίπου συστήματος, αφού αυξάνει τα χαρακτηριστικά των κειμένων. Για το λόγο αυτό έγινε αρχικά η εξής επιλογή: Χρησιμοποιήθηκε ένα SOM με επαρκή αριθμό ομάδων (ενδεικτικά αναφέρεται η διάσταση 22x19, 10x15 κ.ο.κ., όπως θα παρουσιασθεί και εκτενέστερα στην ενότητα των πειραμάτων) και στη συνέχεια οι ομάδες αυτές συμπυκνώθηκαν με ένα απλό αλγόριθμο ομαδοποίησης *k*-μέσων, που συνιστά το υποσύστημα μείωσης διάστασης. Η περαιτέρω ομαδοποίηση του χάρτη SOM έχει προταθεί κατά το παρελθόν στη διεθνή βιβλιογραφία (Vesanto J. & Alhoniemi E., 2000).

4.1.1.3 Μείωση διάστασης – Αλγόριθμος k-means

Το υποσύστημα μείωσης διάστασης χρησιμοποιεί την παραλλαγή του αλγόριθμου k-means, η οποία ενημερώνει τα υποψήφια κέντρα μόνο μετά το πέρας ολόκληρου του συνόλου εκπαίδευσης (batch k-means) (Duda R.O., Hart P.E. & Stork D.G. 2001)), μιας και η προσέγγιση αυτή είναι πιο σταθερή και παρουσιάζει καλύτερα αποτελέσματα. Φυσικά θα μπορούσε να επιλεγεί και σε αυτό το στάδιο ένα μοντέλο SOM, αλλά κάτι τέτοιο θα είχε υπολογιστικά μεγαλύτερες απαιτήσεις, αφού θα απαιτούνταν αρκετά βήματα για την εκπαίδευση του μοντέλου, το οποίο συνήθως χρειάζεται περισσότερες από τις αυστηρά απαιτούμενες ομάδες για να αποδώσει καλά. Εξάλλου, το k-means προσφέρει τη δυνατότητα του ακριβούς καθορισμού του αριθμού των απαιτούμενων ομάδων, ενώ δύναται να εκμεταλλευτεί την τοπολογία και τη γειτνίαση της ομαδοποίησης που έχει πρωτύτερα δημιουργήσει το SOM.

Στον αλγόριθμο k-means αρχικά καθορίζεται το πλήθος των επιθυμητών ομάδων διαχωρισμού των δεδομένων. Ο αλγόριθμος διατηρεί *k* διανύσματα διάστασης ίσης με τη διάσταση των διανυσμάτων εισόδου. κάθε διάνυσμα αντιστοιχεί στο κέντρο μίας ομάδας και η τιμή του αρχικοποιείται τυχαία. Σε κάθε βήμα του αλγόριθμου κάθε δεδομένο αποδίδεται στην ομάδα, της οποίας το κέντρο βρίσκεται «κοντύτερα» στο δεδομένο αυτό. Αυτός ο κανόνας απόδοσης ονομάζεται **κανόνας του κοντινότερου γείτονα** (Nearest Neighbor Rule - NNR). Η «εγγύτητα» εκφράζεται με βάση κάποια συνάρτηση υπολογισμού απόστασης, η οποία συνήθως είναι η Ευκλείδεια απόσταση (4-9):

$$d(\vec{x},\vec{y}) = \left\| (\vec{x} - \vec{y}) \right\| = \sqrt{(\vec{x} - \vec{y})^2} \tag{4-9}$$

Σε κάθε βήμα του αλγόριθμου τα κέντρα ενημερώνονται ανάλογα με τα δεδομένα που αποδόθηκαν σε αυτά, ώστε να αποτελούν το κέντρο των δεδομένων κάθε ομάδας. Πολύ συχνά, για λόγους ευστάθειας του αλγόριθμου εκπαίδευσης, οι ενημερώσεις δε γίνονται κατά την παρουσίαση κάθε δεδομένου εκπαίδευσης, αλλά μετά το πέρας της παρουσίασης του συνόλου των δεδομένων (batch k-means) (4-10):

$$\vec{c}_i = \frac{\sum_{j=1}^{N} \vec{x}_j}{N_i} \tag{4-10}$$

όπου c_i είναι το κέντρο *i* και N_i το πλήθος των στοιχείων που αποδόθηκαν στο κέντρο *i* βάσει του κριτηρίου (4-11):

$$d(\vec{x}_j, \vec{c}_{i^*}) = \min_i \left(d(\vec{x}_j, \vec{c}_i) \right) \tag{4-11}$$

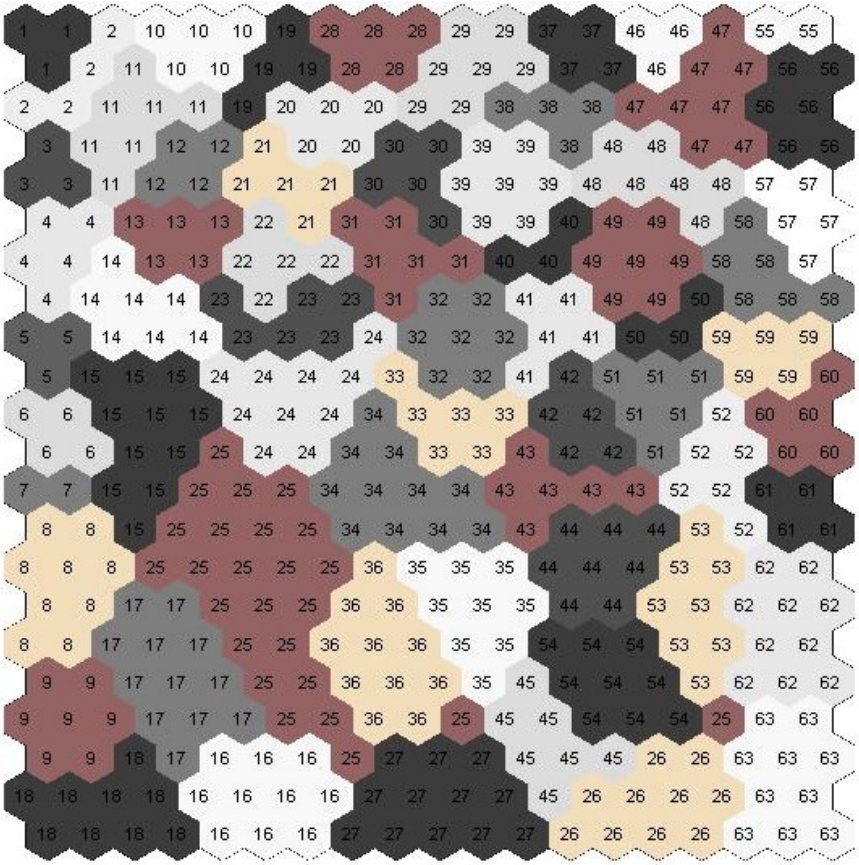
Με τον τρόπο αυτό εξασφαλίζεται το ότι η εξέλιξη του αλγόριθμου μένει ανεπηρέαστη από τη σειρά εμφάνισης των δεδομένων. Ο αλγόριθμος k-means τερματίζει, όταν για διαδοχικά περάσματα του συνόλου εκπαίδευσης δεν έχει αλλάξει σημαντικά η τοποθεσία των κέντρων.

Πρόκειται για έναν απλό στην υλοποίησή του αλγόριθμο και σχετικά γρήγορο στην εκτέλεσή του, καθώς συγκλίνει μετά από λίγες επαναλήψεις. Ωστόσο, συνήθως δεν καταλήγει στη βέλτιστη ομαδοποίηση και τείνει να δημιουργεί πάντα σφαιρικές ομάδες (κυρίως εξαιτίας της χρήσης της Ευκλείδειας απόστασης). Επιπλέον, εάν χρησιμοποιείται τυχαία αρχικοποίηση, τα αποτελέσματα που προκύπτουν μπορεί να είναι πολύ διαφορετικά κάθε φορά. Αυτή η ασυνέπεια του αλγόριθμου αποτελεί το κυριότερο μειονέκτημά του, ενώ έχουν γίνει πολλές προσπάθειες βελτίωσης της συμπεριφοράς του (Likas A., Vlassis N. & Verbeek J., 2003).

Ως δεδομένα εισόδου για την εκπαίδευση του k-means χρησιμοποιούνται τα διανύσματα των κέντρων του SOM και ως εκ τούτου τα κέντρα-πρότυπα έχουν ίδια διάσταση με αυτά. Αξιοποιώντας την τοπολογία του χάρτη SOM που έχει προκύψει από το προηγούμενο στάδιο, αρχικοποιείται ο αλγόριθμος k-means, ώστε τα αρχικά κέντρα να είναι ομοιόμορφα κατανεμημένα στο δισδιάστατο πλέγμα. Επιπλέον, κατά μήκος της καθεμίας από τις δύο διαστάσεις του πλέγματος ο αριθμός των κέντρων ορίζεται ανάλογος προς την αντίστοιχη διάσταση του πλέγματος. Με τον τρόπο αυτό κάθε κέντρο του k-means γειτνιάζει με ίσο περίπου αριθμό ομάδων, δηλαδή η Voronoi (Duda R.O., Hart P.E. & Stork D.G. 2001) περιοχή του καταλαμβάνει αρχικά περίπου το ίδιο εμβαδόν στο χάρτη.

Τα αρχικά κέντρα των k-means οπτικά αποτελούν τις κορυφές ενός ορθογώνιου παραλληλόγραμμου πλέγματος και αρχικά εξισώνονται με τα διανύσματα των αντίστοιχων ισοκατανεμημένων νευρώνων του SOM. Ενδέχεται ο αριθμός *k* να μην είναι ακριβώς ίσος με τον αριθμό των ομάδων που θα προκύψουν, εξαιτίας του περιορισμού της ομοιόμορφης κατανομής στο χάρτη. Για παράδειγμα, εάν επιλεγεί *k*=75 για χάρτη διάστασης 22x19, τότε πρέπει να υπάρχουν $\left[\sqrt{75\frac{22}{19}}\right]=[9{,}32]=9$ κέντρα κατά μήκος της διάστασης των 22 νευρώνων και $\left[\sqrt{75\frac{19}{22}}\right]=[8{,}05]=8$ κατά μήκος της διάστασης των 19, κάτι που έχει ως τελικό αποτέλεσμα 72 κέντρα αντί για 75. Η αρχικοποίηση του k-means είναι συνεπώς πολύ σημαντική, μιας και αυτός ο αλγόριθμος εξαρτάται κατά πολύ από τις αρχικές συνθήκες (Likas A., Vlassis N. & Verbeek J., 2003).

Η τοπολογία του SOM λαμβάνεται υπόψιν μόνο κατά το στάδιο της αρχικοποίησης του αλγόριθμου k-means. Ένα τυπικό παράδειγμα ομαδοποίησης k-means, μετά τη σύγκλιση, με 63 ομάδες (*k=63*) πάνω σε ένα χάρτη SOM δίνεται στο σχήμα (4.4).



Σχήμα 4.4. Παράδειγμα ομαδοποίησης του χάρτη SOM διάστασης 22x19 με χρήση του k-means, όπου σε κάθε νευρώνα αναφέρεται η ομάδα στην οποία αποδόθηκε μετά το k-means.

Τα χρώματα αντιστοιχούν στις σχηματιζόμενες ομάδες, εκ των οποίων οι περισσότερες είναι συνεκτικές, με εξαίρεση την ομάδα 25, τμήματα της οποίας είναι διασκορπισμένα σε περισσότερα του ενός (τρία) σημεία του χάρτη. Αυτό οφείλεται στο τοπογραφικό σφάλμα του SOM λόγω της δραματικής μείωσης της διάστασης της πληροφορίας από τις 600 περίπου αρχικές διαστάσεις του διανύσματος εισόδου στις 2 διαστάσεις του χάρτη. Ωστόσο, το φαινόμενο αυτό δεν είναι ιδιαίτερα εκτεταμένο, ενώ είναι αξιοσημείωτο ότι εμφανίζεται μόνο σε μία εκ των 63 ομάδων νευρώνων του χάρτη και ότι παρόμοια συμπεριφορά παρατηρείται σε όλους σχεδόν τους χάρτες των πειραμάτων μας.

Η ομαδοποίηση των λημμάτων, όπως προκύπτει από τα προηγούμενα στάδια (SOM & k-means) και ειδικότερα από το στάδιο k-means χρησιμεύει στην εξαγωγή χαρακτηριστικών από τα κείμενα. Σε κάθε κείμενο μετρώνται τα λήμματα που ανήκουν σε κάθε κατηγορία. Το σύνολο των μετρήσεων αποτελεί τις συνιστώσες του διανύσματος χαρακτηριστικών στο χώρο προτύπων, έχοντας διάσταση *k* ίση με το πλήθος των ομάδων του k-means. Ουσιαστικά, το διάνυσμα αυτό αντιστοιχεί στο ιστόγραμμα συμμετοχής των λημμάτων ανά κατηγορία. Στη συνέχεια οι συνιστώσες του διανύσματος κανονικοποιούνται, ώστε να αθροίζουν στη μονάδα.

4.1.2 Λήψη απόφασης για τις θεματικές κατηγορίες

Για το δεύτερο στάδιο της διαδικασίας επιλέχθηκαν αλγόριθμοι κατηγοριοποίησης (classifier algorithms) και όχι ομαδοποίησης (clustering algorithms), επειδή οι πρώτοι μπορούν να δώσουν άμεσα κάποιο μετρήσιμο αποτέλεσμα, σε σχέση με τις επιθυμητές προκαθορισμένες ομαδοποιήσεις, ενώ, αντίθετα, η ακρίβεια των μεθόδων ομαδοποίησης εξαρτάται σημαντικά από το σύστημα αναζήτησης που θα βασισθεί πάνω σε αυτές. Εξάλλου, οι αλγόριθμοι κατηγοριοποίησης προσαρμόζονται καλύτερα στις εκάστοτε απαιτήσεις του χρήστη για την έξοδο ενός συστήματος, ενώ είναι σαφώς πιο εναρμονισμένοι με τα όποια αναμενόμενα αποτελέσματα.

Το μοντέλο κατηγοριοποιητή που επιλέχθηκε είναι ένα πολυστρωματικό νευρωνικό δίκτυο MLP, αντίστοιχο με αυτό που χρησιμοποιήθηκε στα προβλήματα αναγνώρισης ύφους. Το επιλεγμένο δίκτυο αποτελείται από τρία επίπεδα, όπου το επίπεδο εισόδου έχει *k* εισόδους (ίσους σε πλήθος με το *k* του *k-means*), στο επίπεδο εξόδου ο αριθμός των εξόδων ισούται με το πλήθος των επιθυμητών τάξεων των δεδομένων, ενώ για το κρυφό επίπεδο δοκιμάστηκαν διάφορα μεγέθη. Για την εκπαίδευση του δικτύου προτιμήθηκε η μέθοδος **ευπροσάρμοστης ανάστροφης διάδοσης** (Resilient Back Propagation – RPROP), παραλλαγή του αλγόριθμου ανάστροφης διάδοσης, καθώς παρουσίασε οριακά καλύτερη απόδοση από αυτή του αλγόριθμου LMBR (ενότητα 3.2.4), ο οποίος χρησιμοποιήθηκε στα πειράματα ύφους. Επιπλέον ο αλγόριθμος RPROP έχει τη δυνατότητα αναπροσαρμογής των βαρών χωρίς να αντιστρέφει πίνακες παραγώγων (Ιακωβιανούς), με αποτέλεσμα να είναι πολύ πιο γρήγορος ανά βήμα εκπαίδευσης και λιγότερο απαιτητικός σε μνήμη συστήματος. Πιο συγκεκριμένα, ο αλγόριθμος RPROP ενημερώνει τα βάρη σε προαποφασισμένα βήματα, τα οποία είναι διαφορετικά για κάθε βάρος και δεν εξαρτώνται από τις τιμές των παραγώγων παρά μόνο από τη φορά του αλγόριθμου εκπαίδευσης στην επιφάνεια του σφάλματος (δηλαδή το πρόσημο των παραγώγων).

Ο αλγόριθμος RPROP αποσκοπεί στη βελτιστοποίηση του μέσου τετραγωνικού σφάλματος και το επιτυγχάνει μεταβάλλοντας το βήμα διόρθωσης των βαρών με βάση την πληροφορία για την κλίση της συνάρτησης σφάλματος (Riedmiller M. & Braun H., 1993 και Igel C. & Huesken. M., 2003). Ο αλγόριθμος αυτός αξιοποιεί για τα βάρη και την πληροφορία για τις προηγούμενες τιμές της κλίσης της συνάρτησης σφάλματος. Χρησιμοποιεί για κάθε βάρος κάποια τιμή, η οποία ονομάζεται τιμή ενημέρωσης βάρους (update weight value) και ισούται με:

$$\Delta_{ij}(t+1) = \begin{cases} n^{+} \cdot \Delta_{ij}(t), & \frac{dE(t)}{dw_{ij}} \cdot \frac{dE(t-1)}{dw_{ij}} > 0 \\ n^{-} \cdot \Delta_{ij}(t), & \frac{dE(t)}{dw_{ij}} \cdot \frac{dE(t-1)}{dw_{ij}} < 0 \\ 0, & \frac{dE(t)}{dw_{ij}} \cdot \frac{dE(t-1)}{dw_{ij}} = 0 \end{cases} \quad (4\text{-}12)$$

όπου $0<\eta^{-}<1<\eta^{+}$. Αν και η τιμή του η^{+} δεν περιορίζεται από τον αλγόριθμο, ωστόσο είναι δυνατό, εάν γίνει πολύ μεγάλη, να οδηγήσει την εκπαίδευση σε άλματα και να την κάνει ασταθή. Ενδεικτικές τιμές για τις παραμέτρους η είναι $\eta^{+}=1,2$ και $\eta^{-}=0,5$.

Στη συνέχεια, τα βάρη ενημερώνονται σύμφωνα με τον τύπο (4-13):

$$\Delta w_{ij}(t+1) = \begin{cases} -\Delta_{ij}(t), & \frac{dE(t)}{dw_{ij}} > 0 \\ +\Delta_{ij}(t), & \frac{dE(t)}{dw_{ij}} < 0 \\ 0, & \frac{dE(t)}{dw_{ij}} = 0 \end{cases} \quad (4\text{-}13)$$

Στο πρώτο βήμα γίνεται έλεγχος για το εάν στις δύο τελευταίες επαναλήψεις έχει αλλάξει πρόσημο η παράγωγος και, στην περίπτωση που κάτι τέτοιο έχει συμβεί, αναιρείται η τελευταία αλλαγή του βάρους. Η λογική του συγκεκριμένου αλγόριθμου υπαγορεύει ότι, εάν σε δύο συνεχόμενες επαναλήψεις (δηλαδή εποχές, αφού οι ενημερώσεις των βαρών γίνονται μετά το πέρας όλου του συνόλου εκπαίδευσης) η κλίση της συνάρτησης σφάλματος διατηρήσει το πρόσημό της, θα πρέπει να αυξηθεί ο ρυθμός μεταβολής των βαρών προς την κατεύθυνση που ήδη γινόταν, προκειμένου να προσεγγισθεί πιο γρήγορα κάποιο ελάχιστο. Οι τιμές ενημέρωσης των βαρών αρχικοποιούνται σε τυχαίες μικρές τιμές, οι οποίες όμως δεν επηρεάζουν ουσιαστικά την εξέλιξη του αλγόριθμου.

Για να εξετασθεί η επάρκεια του συστήματος, πραγματοποιήθηκαν ένα σύνολο από πειράματα, τα οποία είχαν στόχο να εκτιμήσουν την ακρίβεια κατηγοριοποίησης του συστήματος για κείμενα που ανήκουν σε συγκεκριμένες θεματικές κατηγορίες. Επιπλέον υλοποιήθηκαν παραλλαγές του βασικού συστήματος, αλλά και συγκρίσεις αυτού με άλλα υπάρχοντα. Στα πειράματα οργάνωσης κειμένων που περιγράφονται στη συνέχεια χρησιμοποιήθηκε η μέθοδος ευπροσάρμοστης ανάστροφης διάδοσης (RPROP).

4.2 Περιγραφή Δεδομένων

Για τα πειράματα αξιολόγησης του συστήματος χρησιμοποιήθηκε αρχικά το σύνολο των 1.005 κειμένων της Ελληνικής Βουλής, που είχαν χρησιμοποιηθεί και στα υφολογικά πειράματα. Στα κείμενα αυτά αποδόθηκε μία από τις εξής πέντε (5) κατηγορίες: Εσωτερικά, Εξωτερικά, Εκπαίδευση, Οικονομία και Υπόλοιπα. Η απόδοση των κειμένων σε κατηγορίες έγινε χειρωνακτικά. Τα κείμενα που ανήκουν στην κατηγορία «Υπόλοιπα» δεν χρησιμοποιήθηκαν για την εκτίμηση της ακρίβειας κατηγοριοποίησης, αλλά χρησιμοποιήθηκαν οι λέξεις και οι προτάσεις τους για την εξαγωγή των συμφραζομένων κατά το στάδιο δημιουργίας του χάρτη λέξεων. Η κατανομή των κειμένων ανά θεματική κατηγορία φαίνεται στον πίνακα 4.1:

Θεματική Κατηγορία	**Αριθμός Κειμένων**
Εσωτερικά	82
Εξωτερικά	207
Παιδεία	70
Οικονομία	201
Υπόλοιπα	445
Σύνολο	**1005**

Πίνακας 4.1. Α' σώμα κειμένων που χρησιμοποιήθηκε

Επειδή η εφαρμογή ενός τέτοιου συστήματος παρουσιάζει πρακτικό ενδιαφέρον για πολύ ευρύτερα σύνολα δεδομένων, συγκροτήθηκε στη συνέχεια ένα άλλο, μεγαλύτερο σώμα κειμένων. Έτσι, κατασκευάστηκε στη σύγχρονη γλώσσα προγραμματισμού C# ένα ρομποτάκι (web crawler) ικανό να συγκεντρώσει ένα πολύ μεγάλο αριθμό κειμένων από το ηλεκτρονικό αρχείο εφημερίδας ευρείας κυκλοφορίας (πιο συγκεκριμένα της Ελευθεροτυπίας, για το διάστημα 2001 – 2006). Από το ηλεκτρονικό αρχείο συγκεντρώθηκαν 64.783 κείμενα, από τα οποία αφαιρέθηκαν οι διπλοεγγραφές (δηλαδή κείμενα με περισσότερες από μία συνδέσεις (links)), καθώς και κείμενα με λιγότερες από 100 λέξεις, με αποτέλεσμα να προκύψουν **36.450** κείμενα, τα οποία, βάσει της δομής του ιστότοπου από τον οποίο προήλθαν, αντιστοιχήθηκαν σε έξι κατηγορίες, **Πολιτική**, **Κόσμος**, **Ελλάδα**, **Οικονομία**, **Τέχνες** και **Αθλητισμός**, και μετά υπέστησαν λημματοποίηση.

Πρέπει να δοθεί ιδιαίτερη έμφαση στο γεγονός ότι, εκτός της λημματοποίησης και της αυτόματης αφαίρεσης των διπλών και των μικρών κειμένων, δεν έγινε καμία περαιτέρω επεξεργασία, πολύ δε περισσότερο χειρωνακτική. Ο τρόπος που συγκεντρώθηκαν τα δεδομένα αυτά τα καθιστά ιδιαίτερα κατάλληλα για την εκτίμηση της επάρκειας του συστήματος σε πραγματικές συνθήκες, που απηχούν απαιτήσεις πρακτικών εφαρμογών. Στον ακόλουθο πίνακα 4.2 παρουσιάζεται η κατανομή των συλλεγμένων κειμένων ανά θεματική κατηγορία.

Θεματική Κατηγορία	**Αριθμός Κειμένων**
Πολιτική	7078
Κόσμος	5598
Ελλάδα	7403
Τέχνες	5315
Οικονομία	5584
Αθλητισμός	5472
Σύνολο	**36450**

Πίνακας 4.2. Β' σώμα κειμένων που χρησιμοποιήθηκε

Το Β' σώμα κειμένων είναι πολύ μεγαλύτερο και βάσει του τρόπου συγκρότησής του ανταποκρίνεται καλύτερα στις απαιτήσεις του προβλήματος. Ωστόσο, για λόγους καθαρά υπολογιστικούς, οι περισσότερες δοκιμές παραλλαγών στις παραμέτρους του προτεινόμενου συστήματος έγιναν για το Α' σύνολο και μόνο μετά τη σταθεροποίηση των παραμέτρων έγιναν δοκιμές και με το σαφώς μεγαλύτερο Β' σύνολο δεδομένων.

4.3 Αυτόνομη αξιολόγηση χαρτών

Η αξιολόγηση της ποιότητας των χαρτών είναι εξαιρετικά δύσκολη διαδικασία αφού δεν μπορεί να υπάρξει διαθέσιμη κάποια πρότυπη κατηγοριοποίηση των λέξεων. Ειδικότερα δε είναι αδύνατο να προκύψει ένα πρότυπο αποτέλεσμα απόλυτα αντικειμενικό και κοινώς αποδεκτό, αφού ακόμη και το ίδιο άτομο ενδεχομένως να κατηγοριοποιήσει τις λέξεις διαφορετικά σε διαφορετικές χρονικές στιγμές ή ακόμη και ανάλογα με τη σειρά που τις βλέπει. Επιπρόσθετα πολλές λέξεις είναι πολύσημες είτε χρησιμοποιούνται μεταφορικά είτε αποκτούν εξειδικευμένο νόημα ανάλογα με τη θεματική του κειμένου και τα συμφραζόμενα. Επίσης, λαμβάνοντας υπόψιν ότι οι απαιτήσεις αφορούν χιλιάδες λέξεις κάθε προσπάθεια χειρωνακτικής ή εμπειρικής αξιολόγησης μπορεί να οδηγηθεί σε αδιέξοδο ή σε παραπλανητικά αποτελέσματα. Για το λόγο αυτό έγινε προσπάθεια να ορισθούν κάποια αντικειμενικά – ποσοτικά κριτήρια χαρακτηρισμού των προκυπτουσών ομάδων.

4.3.1 Χρήση λέξεων-κλειδιών

Για την εκτίμηση της διαχωριστικής ικανότητας του χάρτη στα δεδομένα της Βουλής (Α' σώμα κειμένων) ακολουθήθηκε η εξής διαδικασία: Αρχικά δημιουργήθηκαν 4 σύνολα χαρακτηριστικών λέξεων αντίστοιχα με τις 4 διαφορετικές θεματικές κατηγορίες του συνόλου δεδομένων: (1) εσωτερικά ζητήματα, (2) εξωτερικά ζητήματα, (3) οικονομικά θέματα και (4) παιδεία. Τα σύνολα αυτά συγκροτήθηκαν με τον ακόλουθο τρόπο:

Αρχικά για το πρώτο σύνολο ζητήθηκε από ένα γλωσσολόγο να επιλέξει 100 λέξεις από κάθε θεματική κατηγορία χωρίς να έχει πρόσβαση στα κείμενα ή στους χάρτες λέξεων. Από τις επιλεγμένες λέξεις αυτές αφαιρέθηκαν αυτές που απαντώνται σε περισσότερες από μία κατηγορίες. Το υπόλοιπα τρία σύνολα δημιουργήθηκαν με ημιαυτόματο τρόπο.

Το αρχείο της Βουλής είχε ήδη διαθέσιμη μία ταξινόμηση των κειμένων σε κατηγορίες, των οποίων ο τίτλος χρησιμοποιήθηκε για να δημιουργηθεί ένα σύνολο από λέξεις-κλειδιά. Από το σύνολο αυτό αφαιρέθηκαν οι συχνές λέξεις, κυρίως λειτουργικές, και οι λέξεις που ανήκουν σε περισσότερες από μία κατηγορίες, ενώ στη συνέχεια αφαιρέθηκαν χειρωνακτικά κάποιες άσχετες με το εκάστοτε θέμα, λέξεις με τη βοήθεια ενός γλωσσολόγου.

Τα υπόλοιπα δύο σύνολα δημιουργήθηκαν ημιαυτόματα με τον εντοπισμό ανά κατηγορία εκείνων των λέξεων που καταλάμβαναν το 70%~85% και το 85%-90% του εύρους των εμφανίσεων, σε αντιστοιχία δηλαδή με τον τρόπο που έγινε η εξαγωγή των χαρακτηριστικών των λέξεων. Ουσιαστικά εφαρμόσθηκε η ανάλυση ABC ανά κατηγορία κειμένων. Θα πρέπει να αναφερθεί ότι ορισμένες από τις λέξεις-κλειδιά που προήλθαν με αυτόν τον τρόπο ενδεχομένως να ανήκουν στην κατηγορία B, η οποία όμως δεν απεικονίζεται στο χάρτη, με αποτέλεσμα αυτές ουσιαστικά να αγνοούνται. Σε αυτά τα σύνολα ακολουθήθηκε αντίστοιχη επεξεργασία με τα προηγούμενα. Με τη χρήση αυτών των περιορισμένων συνόλων (δειγματοληψία - sampling) είναι δυνατόν να παρουσιασθεί οπτικά για κάθε σύνολο το κατά πόσο λέξεις της ίδιας κατηγορίας ανήκουν σε γειτονικά κελιά στο χάρτη SOM, κάτι που φυσικά δεν θα ήταν δυνατόν να γίνει για όλες τις λέξεις.

Πέρα όμως από τα οπτικά αποτελέσματα αναζητήθηκαν αντικειμενικά – ποσοτικά κριτήρια στην προσπάθεια να περιγραφεί με ποσοτικό τρόπο η γεωμετρία του χάρτη. Αρχικά μετρήθηκε η μέση τιμή και η διασπορά των αποστάσεων των λέξεων που ανήκουν στην ίδια ομάδα πάνω στο πλέγμα του χάρτη (**κριτήριο 1**). Επιπλέον μετρήθηκε η μέση τιμή και η διασπορά των αποστάσεων αυτών που δεν ανήκουν στην ίδια ομάδα. Σκοπός του κριτηρίου αυτού είναι να μετρήσει το πόσο κοντά προβάλλονται στο χάρτη οι λέξεις-κλειδιά που ανήκουν στην ίδια ομάδα σε σχέση με αυτές που ανήκουν σε διαφορετικές.

Στη συνέχεια μετρήθηκε η μέση τιμή των αποστάσεων για κάθε ομάδα από τον κοντινότερο γείτονα που περιέχει λέξεις που ανήκουν στην κατηγορία σε σχέση με τη μέση τιμή των αποστάσεων του κοντινότερου γείτονα που δεν ανήκει στην κατηγορία (**κριτήριο 2**). Το κριτήριο αυτό προσπαθεί να μετρήσει το βαθμό που οι γειτονικές ομάδες του χάρτη SOM περιέχουν νοηματικά συγγενείς λέξεις, περιοριζόμενο βέβαια στα σύνολα των λέξεων-κλειδιών.

Τέλος, μετρήθηκε το ποσοστό των κελιών – νευρώνων που περιέχουν «καθαρές» ομάδες κάθε κατηγορίας, δηλαδή ομάδες που δεν περιέχουν λέξεις άλλων κατηγοριών (**κριτήριο 3**). Το συγκεκριμένο κριτήριο αντικατοπτρίζει το βαθμό που οι λέξεις-κλειδιά διαχωρίζονται προβαλλόμενες στις ομάδες του χάρτη.

Κάθε ένα από τα κριτήρια αυτά αποσκοπεί στην αποτύπωση συγκεκριμένων γεωμετρικών ιδιοτήτων, μπορεί ωστόσο να είναι παραπλανητικό ανάλογα με τη μορφή του εκάστοτε χάρτη. Για παράδειγμα, το κριτήριο 3 στην περίπτωση αραιών και μικρών ομάδων μπορεί να εμφανίσει αδικαιολόγητες, για την ποιότητα του αποτελέσματος, υψηλές τιμές.

Ένα άλλο κριτήριο (**κριτήριο 4**) αφορά στο κατά πόσο οι ομάδες που αντιστοιχίζονται οι λέξεις στο χάρτη μπορεί να είναι γραμμικά διαχωρίσιμες. Μία σχετική τεχνική είναι η χρήση ενός δικτύου MLP με λίγους νευρώνες (2-3) στο κρυφό επίπεδο, ώστε να μην μπορεί να απεικονίσει πολύπλοκες δομές. Είναι ευρέως γνωστό ότι για να λυθεί το πρόβλημα XOR που χωρίζει το δισδιάστατο χώρο σε τέσσερις το πολύ διαφορετικές περιοχές απαιτούνται κατ' ελάχιστο 2 νευρώνες στο κρυφό επίπεδο ενός δομημένου MLP. Συγκεκριμένα, για να αξιολογηθεί η απόδοση του χάρτη για κάθε σύνολο λέξεων, τοποθετήθηκαν οι λέξεις στο χάρτη με κάθε κατηγορία να φέρει διαφορετική ετικέτα. Ως είσοδος στο δίκτυο χρησιμοποιήθηκαν οι συντεταγμένες(*x*,*y*) στο χάρτη και ως έξοδος μία δυαδική κωδικοποίηση των ετικετών σε τέσσερις εξόδους ([-1 -1 -1 1], [-1 -1 1 -1], [-1 1 -1 -1],[1 -1 -1 -1]). Το μέσο τετραγωνικό σφάλμα που επιτυγχάνει το MLP κατά την εκπαίδευση φανερώνει το βαθμό που τα δεδομένα είναι γραμμικά διαχωρίσιμα και αποτελεί την τιμή του κριτηρίου 4.

Η πιο εφικτή διαδικασία αξιολόγησης της μεθόδου είναι η χρήση της ως μέθοδος επιλογής χαρακτηριστικών στα πλάισια ενός συστήματος κατηγοριοποίησης κειμένων με βάση το θέμα τους και ως εκ τούτου να κριθεί συνολικά ως προς το αποτέλεσμα του συστήματος. Για το λόγο αυτό η αποτελεσματικότητα της μεθόδου μπορεί να κριθεί συνολικά από τα πειραματικά αποτελέσματα του συνολικού συτήματος που παρουσιάζονται στην ενότητα (4.4.1).

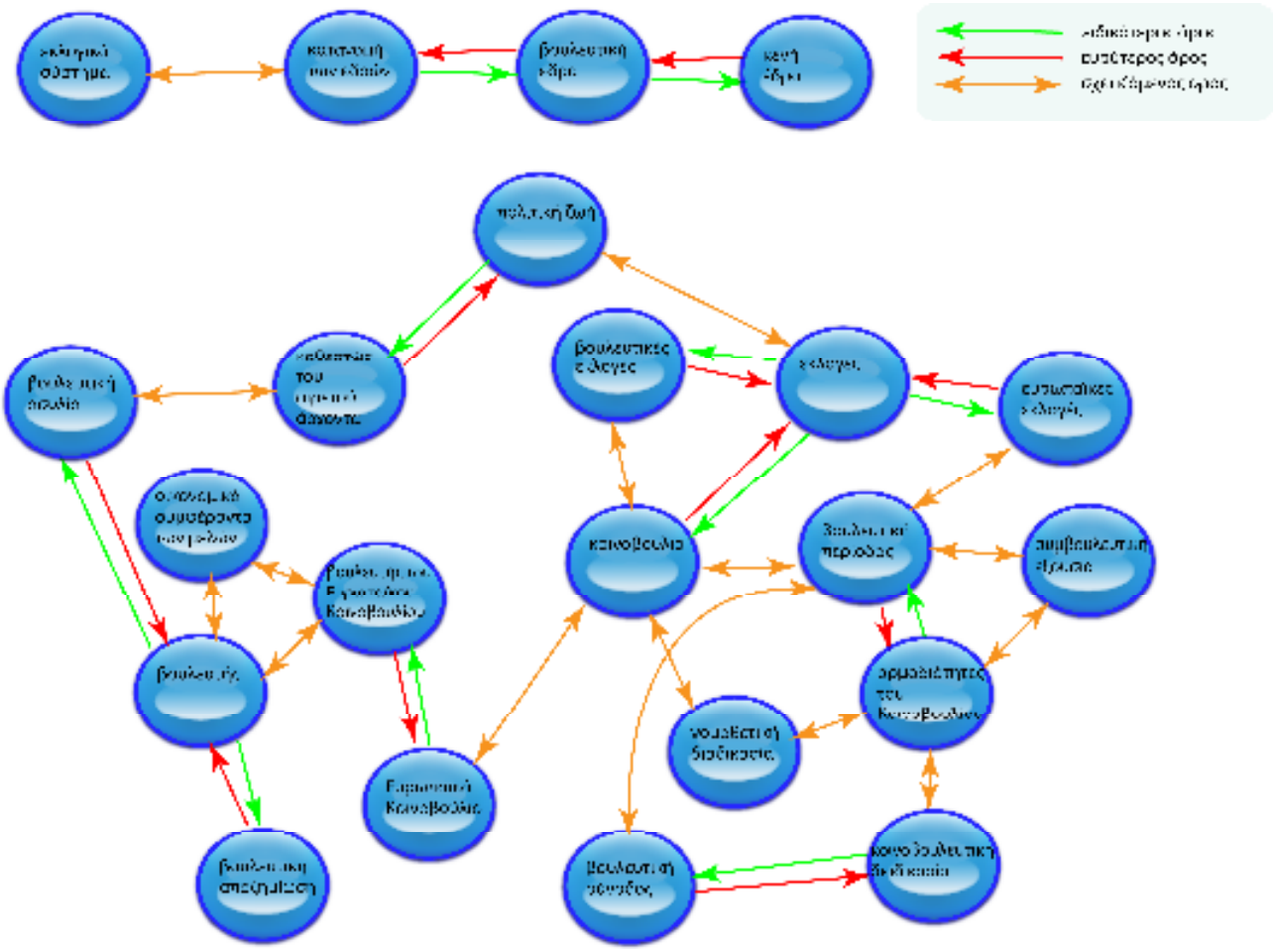
4.3.2 Θησαυρός Eurovoc

Μία επιπλέον προσπάθεια που δοκιμάσθηκε για την αξιολόγηση της ομαδοποίησης του χάρτη λέξεων αφορούσε στη σύγκριση του χάρτη με ένα θησαυρό όρων. Ο θησαυρός συνιστά μία δόμηση λέξεων βάσει κάποιων σημασιακών σχέσεων (sense relations) μεταξύ αυτών, όπως σχέσεις υπωνυμίας (narrow terms και broader terms), δηλαδή την υπαγωγή μίας ειδικής έννοιας σε μία γενικότερη, αλλά και σχέσεις νοηματικής συγγένειας (related terms), που είναι ευρύτερη από τη συνωνυμία. Το είδος των συσχετίσεων που μπορεί να εκφράζει ένας θησαυρός εξαρτάται από την εκάστοτε υλοποίηση, ενώ ένας θησαυρός συνήθως καλύπτει κάποια ορισμένη θεματική περιοχή.

Για την αξιολόγηση των αποτελεσμάτων χρησιμοποιήθηκε ο θησαυρός Eurovoc, ο οποίος σχετίζεται με διοικητικά κυρίως θέματα στους τομείς δραστηριότητας της Ευρωπαϊκής Κοινότητας. Εξαιτίας των θεματικών κατηγοριών που καλύπτει φαίνεται αρκετά συγγενής με το σύνολο δεδομένων της Βουλής. Οι όροι στο Eurovoc είναι μονολεκτικοί και πο-

λυλεκτικοί και συνδέονται με σχέσεις της μορφής: α) εξειδικευμένος όρος (narrow term), β) ευρύτερος όρος (broader term) και γ) σχετιζόμενος όρος (related term).

Υπολογιστικά, ένας θησαυρός μπορεί να αναπαρασταθεί ως ένας γράφος: Κάθε κορυφή/κόμβος του γράφου περιέχει έναν όρο, ενώ οι συνδέσεις μεταξύ των κορυφών εκφράζουν τις σχέσεις. Προφανώς για κάθε είδος σχέσης υπάρχει και ένα διαφορετικό είδος σύνδεσης. Ειδικότερα, οι σχέσεις εξειδικευμένος και ευρύτερος όρος είναι κατευθυνόμενες (directed), ενώ οι σχέσεις σχετιζόμενου όρου όχι. Το Eurovoc είναι σε μορφή XML, με την επεξεργασία του οποίου δημιουργείται ο γράφος του θησαυρού στη μνήμη. Ένα ελάχιστο τμήμα του Eurovoc απεικονίζεται ενδεικτικά σε μορφή γράφου στο σχήμα 4.5.



Σχήμα 4.5. Τμήμα του θησαυρού Eurovoc απεικονιζόμενο σε μορφή γράφου

Ο θησαυρός Eurovoc χρησιμοποιείται ως το μέσο μέτρησης της σχέσης μεταξύ λέξεων του χάρτη. Πιο συγκεκριμένα, για κάθε ζευγάρι λέξεων αναζητείται η πιθανή σύνδεσή τους στο Eurovoc. Η σύνδεση αναφέρεται στο ελάχιστο πλήθος των συνδέσεων που χρειάζεται να διαβεί κάποιος για να περιηγηθεί από τη μία λέξη στην άλλη. Επειδή το πλήθος των

όρων που αντιστοιχεί σε κορφές του χάρτη είναι της τάξεως των χιλιάδων, δεν είναι υπολογιστικά εφικτή η εξαντλητική αναζήτηση στο γράφο.

Για το λόγο αυτό στον εξαντλητικό αλγόριθμο αναζήτησης του Dijkstra (Dijkstra E. W., 1959) προστέθηκε κάποιο κατώφλι, το οποίο ορίζει κάτω από ποιο όρια θεωρείται ότι υπάρχει σύνδεση μεταξύ δύο όρων. Ο αλγόριθμος Dijkstra προσπαθεί από την αρχή να φτάσει στο στόχο επιλέγοντας κάθε φορά γειτονικούς κόμβους που δεν έχει επισκεφθεί και ενημερώνοντας την τιμή της ελάχιστης απόσταση του εκάστοτε επισκεπτόμενου κόμβου από τον αρχικό. Όταν η ελάχιστη απόσταση για το δεδομένο βήμα του αλγόριθμου υπερβεί το κατώφλι, η αναζήτηση σταματάει, αφού όλες οι αποστάσεις είναι θετικές και δεν υπάρχει δυνατότητα να βρεθεί η κορυφή-στόχος στα απαιτούμενα όρια απόστασης.

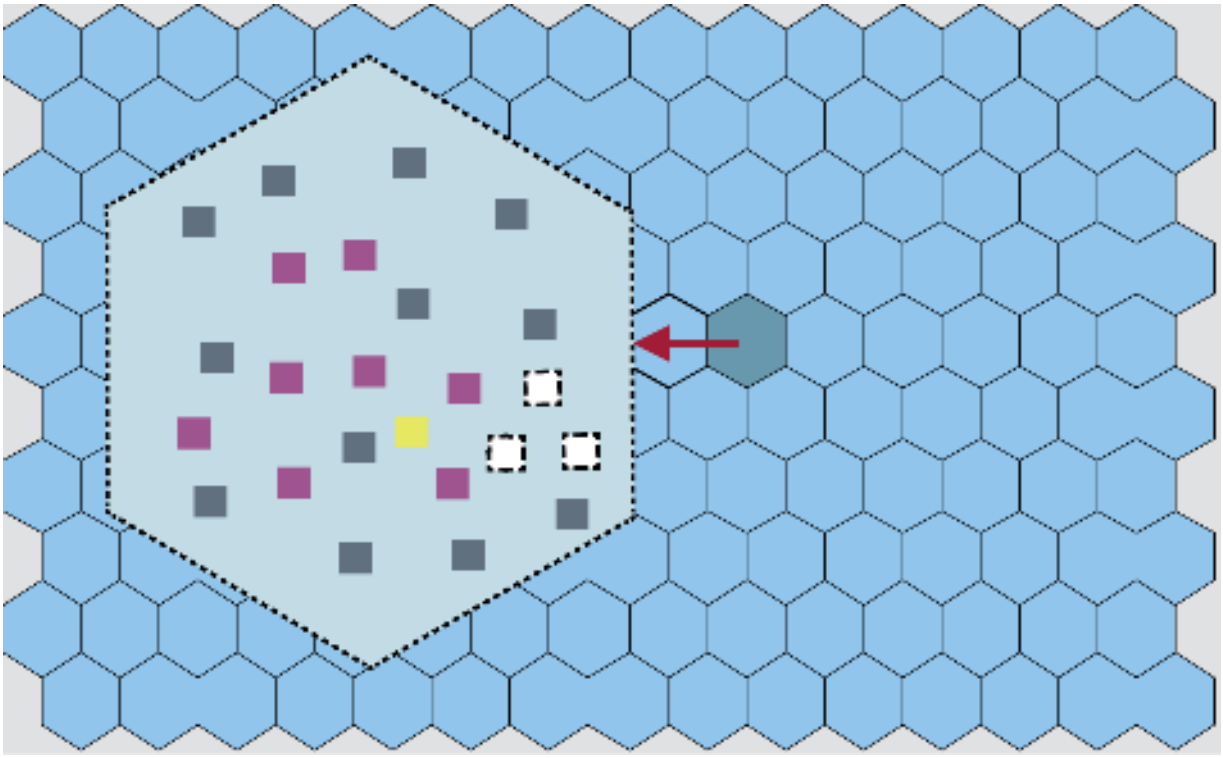
Προκειμένου να ποσοτικοποιηθούν οι συνδέσεις του γράφου στο κόστος των συνδέσεων ευρύτερων (broader term) και εξειδικευμένων όρων (narrow term) δόθηκε η τιμή 4, ενώ στους σχετιζόμενους όρους (related term) η τιμή 5. Η επιλογή αυτή είναι καθαρά εμπειρική και βασίστηκε στο σκεπτικό ότι οι σχετιζόμενοι όροι μπορεί να είναι ευρύτεροι νοηματικά και σταδιακά να μεταβαίνουν σε πιο απομακρυσμένους κόμβους. Τα κατώφλια που επιλέχθηκαν είχαν τιμές από 1-5 φορές το κόστος των εξειδικευμένων όρων (*1*4=4,2*4=8,3*4=12,4*4=16,5*4=20*).

Δεδομένου ότι ο θησαυρός και τα κείμενα της Βουλής έχουν διαφορετική προέλευση, είναι λογικό, παρά τη θεματική τους συγγένεια, να μην εντοπίζονται κάποιες λέξεις των κειμένων της Βουλής στους όρους του θησαυρού. Επιπλέον, οι περισσότεροι όροι του θησαυρού είναι πολυλεκτικοί, οπότε για να εντοπισθεί μία λέξη σε αυτόν αρκεί η λέξη αυτή να αποτελεί τμήμα ενός όρου. Ακόμη, μία λέξη του χάρτη μπορεί να απαντάται σε περισσότερες της μίας κορυφές του Eurovoc, οπότε προτιμάται η κορυφή που δίνει την ελάχιστη απόσταση.

Με βάση τον παραπάνω υπολογισμό της απόστασης μεταξύ λέξεων γίνονται οι εξής μετρήσεις για την ποιότητα του χάρτη: Για κάθε ομάδα του χάρτη SOM μετράται το πλήθος των δυνατών συνδέσεων ($ΠΔΣ_i$) ανά ζεύγος λέξεων στο Eurovoc. Κάθε σύνδεση μεταξύ δύο λέξεων της ομάδας πρέπει να απαντάται τουλάχιστον μία φορά στο Eurovoc (π.χ. η σύνδεση A-B σημαίνει A->B και B->A), ώστε οι λέξεις να θεωρηθούν συνδεδεμένες. Εδώ το $ΠΔΣ_i$ είναι πάντα μικρότερο από την τιμή $(N^2-N)/2$, όπου N το πλήθος των λέξεων της ομάδας του SOM, αφού σπάνια υπάρχουν όλες οι λέξεις της ομάδας στο Eurovoc. Στη συνέχεια, μετράται το πλήθος των πραγματικών συνδέσεων ($ΠΠΣ_i$) για κάθε ομάδα ως το σύνολο των συνδέσεων ανά ζεύγος λέξεων που συνδέονται και στο θησαυρό Eurovoc σε απόσταση μικρότερη από ή ίση με 1,2,3,4,5 φορές το κόστος των εξειδικευμένων όρων.

Η πρώτη μέτρηση (*eurovoc_w)* για το δεδομένο κατώφλι (*c*) προκύπτει ως ο λόγος του συνόλου των υπαρκτών συνδέσεων προς τις δυνατές συνδέσεις για όλες τις ομάδες του χάρτη (*k)* (4-14):

$$eurovoc\_w(c) = \frac{\sum_{i=1}^{k} \Pi\Pi\Sigma_i(c)}{\sum_{i=1}^{k} \Pi\Delta\Sigma_i(c)} \qquad (4\text{-}14)$$



Σχήμα 4.6. Παράδειγμα υπολογισμού του κριτηρίου *eurovoc_w* για το Eurovoc

Ένα παράδειγμα εφαρμογής αυτού του κριτηρίου παρουσιάζεται στο σχήμα 4.6, όπου απεικονίζεται σε μεγέθυνση μία ομάδα του χάρτη SOM. Οι λέξεις αντιστοιχίζονται σε τετράγωνα σχήματα και με το ίδιο χρώμα παρουσιάζονται αυτές που είναι συνδεδεμένες στο Eurovoc, δηλαδή συνδέονται σε απόσταση μικρότερη από το καθορισμένο κατώφλι (*c*). Οι λέξεις που απεικονίζονται με άσπρο χρώμα και διακεκριμένο περίγραμμα είναι αυτές που δεν υπάρχουν στο Eurovoc. Ο υπολογισμός της τιμής του κριτηρίου (4-14) για το σχήμα 4.6 θα ήταν ο εξής:

- Για τις πράσινες λέξεις, που είναι 12, το πλήθος των συνδέσεων είναι *(12*11)/2=66*.

- Για τις κόκκινες λέξεις, που είναι 8, το πλήθος των συνδέσεων είναι *(8*7)/2=28*.
- Η κίτρινη λέξη είναι μοναδική και δεν έχει καμία σύνδεση.
- Οι άσπρες λέξεις δεν εμφανίζονται στο Eurovoc και δε δίνουν καμία σύνδεση.

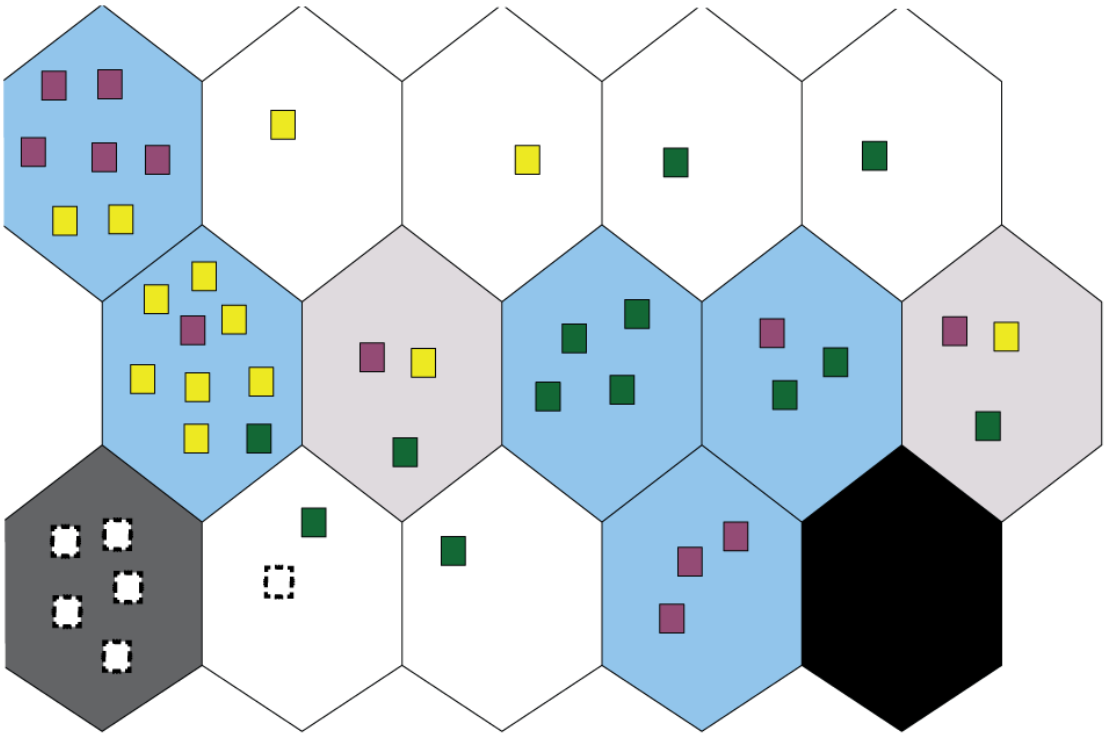
Το πλήθος των πραγματικών συνδέσεων (*ΠΠΣ$_i$*) είναι *28+66=94*.

Το πλήθος των δυνατών συνδέσεων (*ΠΔΣ$_i$*) υπολογίζεται εάν αγνοηθούν οι άσπρες λέξεις από το σύνολο λέξεων (*12+8+1+3=24* ή 21 χωρίς τις άσπρες) και ισούται με (21*20)/2=210. Το συνολικό κριτήριο είναι *eurovoc_w(c)=94/210=0,448*.

Μία δεύτερη μέτρηση που χρησιμοποιήθηκε βασίζεται στο ποσοστό των ομάδων του χάρτη που περιέχουν τουλάχιστον ένα ζεύγος λέξεων συνδεδεμένο και στο θησαυρό Eurovoc (δηλαδή το ποσοστό των ομάδων που δεν είναι εντελώς ασύνδετες). Για κάθε χάρτη μετράται το πλήθος των ομάδων που έχουν μη μηδενικό πλήθος από συνδέσεις ανά ζεύγος λέξεων της ομάδας στο Eurovoc στα όρια του κάθε κατωφλίου. Το πλήθος αυτό διαιρείται με το σύνολο των ομάδων του χάρτη, οι οποίες όμως θα μπορούσαν να έχουν τουλάχιστον μία δυνατή διασύνδεση. Δηλαδή, τουλάχιστον ένα ζεύγος λέξεών τους εμφανίζεται στο Eurovoc, ανεξάρτητα με το εάν εκεί είναι συνδεδεμένο ή όχι. Ο περιορισμός αυτός έχει να κάνει με το γεγονός ότι το Eurovoc δεν καλύπτει όλες τις λέξεις του χάρτη. Πιο συγκεκριμένα το μέτρο δίνεται από τον τύπο (4-15) (*eurovoc_g)* για το κατώφλι (*c)*.

$$eurovoc\_g(c) = \frac{Count_{i=1}^{k}\left[\Pi\Pi\Sigma_i(c) > 0\right]}{k} \qquad (4\text{-}15)$$

Ένα παράδειγμα εφαρμογής του κριτηρίου αυτού παρουσιάζεται στο σχήμα 4.7, όπου εμφανίζεται ένας πολύ μικρός χάρτης με μέγεθος 3x5. Οι ομάδες που έχουν τουλάχιστον ένα συνδεδεμένο ζευγάρι λέξεων στο Eurovoc καλύπτονται με γαλάζιο χρώμα και είναι συνολικά πέντε. Η μαύρη ομάδα είναι τελείως κενή και δεν υπολογίζεται, όπως αντίστοιχα και οι άσπρες ομάδες, οι οποίες δεν περιέχουν κανένα δυνατό ζευγάρι λέξεων, αφού δεν περιλαμβάνουν τουλάχιστον δύο λέξεις που να υπάρχουν στο Eurovoc. Η γκρι σκούρο ομάδα περιλαμβάνει μόνο λέξεις που δεν απαντώνται στο Eurovoc και για το λόγο αυτό δε λαμβάνεται υπόψιν. Τέλος, οι γκρι ομάδες έχουν πάνω από δύο λέξεις που υπάρχουν στο Eurovoc και λαμβάνονται υπόψιν στον υπολογισμό του παρανομαστή του κριτηρίου (4-15). Συγκεκριμένα, η τιμή του κριτηρίου είναι ίση με το πλήθος των μπλε ομάδων προς το πλήθος των μπλε και ανοιχτό γκρι ομάδων: *5/(5+2)=5/7=0,714*.



Σχήμα 4.7 Παράδειγμα υπολογισμού του κριτηρίου *eurovoc_g* για το Eurovoc

Οι μετρήσεις που προέκυψαν συγκρίθηκαν και με χάρτες που δημιουργήθηκαν με τυχαίους τρόπους. Πιο συγκεκριμένα, συγκρίθηκαν με την πιθανότητα δύο οποιωνδήποτε τυχαία επιλεγμένων λέξεων να είναι συνδεδεμένες στο Eurovoc (Monte Carlo Simulation) (MacKay D., 2003) (**μέθοδος 1, κριτήριο *eurovoc_w_rand1***). Επιπλέον συγκρίθηκαν με την πιθανότητα δύο λέξεων προερχόμενων από ένα τυχαία δημιουργημένο χάρτη να είναι συνδεδεμένες στο Eurovoc (**μέθοδος 2, κριτήριο *eurovoc_w_rand2***). Ο τυχαία δημιουργημένος χάρτης έχει ίδιο αριθμό ομάδων με τον εξεταζόμενο χάρτη και κάθε ομάδα έχει ίδιο (σταθερό) πλήθος λέξεων (ισοκατανεμημένες). Στον τυχαία δημιουργημένο χάρτη εξετάζεται συγκριτικά και η πιθανότητα των τυχαίων ομάδων να έχουν τουλάχιστον μία σύνδεση στο Eurovoc (**μέθοδος 3, *eurovoc_g_rand1***).

4.4 Πειράματα

Το σύστημα που χρησιμοποιήθηκε για τα πειράματα κατηγοριοποίησης υλοποιήθηκε χρησιμοποιώντας συνδυασμούς γλώσσας προγραμματισμού C++ και C#. Κάποια τμήματα του συστήματος, που σχετίζονται με πολύπλοκες και επαναλαμβανόμενες μαθηματικές πράξεις όπως οι αλγόριθμοι εκπαίδευσης, έχουν υλοποιηθεί σε γλώσσα C++, η οποία είναι ικανή να εκτελεί αριθμητικές πράξεις και πράξεις μεταξύ πινάκων με πολύ μεγάλη ταχύτητα. Πιο συγκεκριμένα οι αλγόριθμοι λειτουργίας και εκπαίδευσης έχουν κωδικοποιηθεί με C++ σε μία βιβλιοθήκη τύπου dll για Windows και μέσω κλήσεων COM μπορεί να χρησιμοποιείται ο κώδικας και από C#. Ο αλγόριθμος SOM που υλοποιήθηκε αποτελεί μία μεταφορά και τροποποίηση του SOM toolbox (Vesanto J., 2000) που τρέχει σε περιβάλλον Matlab. Τα τμήματα του συστήματος που αφορούν στη μέτρηση χαρακτηριστικών έχουν υλοποιηθεί στην πιο σύγχρονη γλώσσα C# (χρησιμοποιώντας Microsoft .Net Framework 2), η οποία αποτελεί και εξέλιξη της C++. Η γλώσσα αυτή, αν και πιο αργή στις μαθηματικές πράξεις, εντούτοις περιέχει ένα σύνολο από πολύ χρήσιμες βιβλιοθήκες που βοηθούν την επεξεργασία αλλά και τη μέτρηση χαρακτηριστικών από κείμενα. Ιδιαίτερα η βιβλιοθήκη αναζήτησης συμβολοακολουθιών (Regex) με ειδικές εκφράσεις, που ενσωματώνει δυναμική αναζήτηση συμβολοακολουθιών βασιζόμενη σε κανονικές εκφράσεις (regular expressions), είναι σημαντική για την επεξεργασία κειμένων από το σύστημα.

Με σκοπό την αποτίμηση της επάρκειας του προτεινόμενου συστήματος στο δεδομένο πρόβλημα, διεξάχθη ένα σύνολο πειραμάτων με διαφορετικές παραμέτρους για το σύστημα. Στο πρώτο σύνολο πειραμάτων έγινε συγκριτική δοκιμή για τις πιο βασικές παραμέτρους του συστήματος, οι οποίες δίνονται στη συνέχεια:

- Το μέγεθος του χάρτη SOM που χρησιμοποιήθηκε
- Η τιμή *k* του k-means για το στάδιο μείωσης της διάστασης που καθορίζει και το πλήθος χαρακτηριστικών για κάθε κείμενο
- Το πλήθος των νευρώνων του κρυφού επιπέδου για τον κατηγοριοποιητή MLP
- Τα όρια που τίθενται στις κατηγορίες A, B και C κατά το στάδιο της ανάλυσης ABC
- Η εφαρμογή ή μη της κανονικοποίησης στα διανύσματα που αναπαριστούν τα λήμματα στο χάρτη

Η τελευταία αυτή παράμετρος μπορεί να μη φαίνεται σημαντική εκ πρώτης όψεως, ωστόσο είναι πολύ ουσιαστική, καθώς η εφαρμογή της κανονικοποίησης μπορεί να επιφέρει βελτιστοποίηση στο χρόνο εκτέλεσης του συστήματος, ενώ με τη μη εφαρμογή της οι μετρήσεις έχουν καλύτερη πραγματική ερμηνεία, αφού αντιστοιχούν ουσιαστικά στη δεσμευμένη πιθανότητα κατά Bayes (MacKay D. 2003 και Bell T. C., Cleary J. G., & Witten I. H., 1990), δηλαδή, δεδομένου ότι έχει εμφανισθεί το λήμμα *a*, ποια είναι η πιθανότητα να ακολουθήσει στην πρόταση και το χαρακτηριστικό *b*. Οι τιμές που χρησιμοποιήθηκαν για τις παραμέτρους περιγράφονται στη συνέχεια.

Δοκιμάσθηκαν χάρτες δύο διαφορετικών μεγεθών, (α) του 10x15 και (β) του μεγέθους που αποφαίνεται αυτόματα ο αλγόριθμος αρχικοποίησης του SOM (Kohonen 1997) και το οποίο διαφέρει ανάλογα με τις μετρήσεις προς επεξεργασία. Πιο συγκεκριμένα, ο αλγόριθμος αυτός αποτελεί ουσιαστικά μία εμπειρική μέθοδο, η οποία αρχικά υπολογίζει το πλήθος των νευρώνων ανάλογα με το πλήθος των δεδομένων εισόδου. Στη συνέχεια κατανέμει τους νευρώνες στο πλέγμα των δύο διαστάσεων αναλογικά, σύμφωνα με τις δύο μεγαλύτερες ιδιοτιμές του πίνακα αυτοσυσχέτισης των δεδομένων.

Επιλέχθηκαν τα όρια 85% μεταξύ των κατηγοριών A και B με 70% μεταξύ των B και C, σε σύγκριση με τα 85% και 90% αντίστοιχα. Οι τιμές που επιλέχθηκαν για την παράμετρο *k* του συστήματος είναι οι εξής: 70, 60, 50, 40, 30, 20, 10, 6, 4, και 2, οι οποίες όμως δεν είναι πάντα αυτές ακριβώς εξαιτίας του περιορισμού για την ισοκατανομή των κέντρων του k-means στο χάρτη[2]. Όπως είναι γνωστό, ο αριθμός των νευρώνων στο κρυφό επίπεδο του MLP δεν μπορεί να καθοριστεί με ντετερμινιστικό τρόπο και για το λόγο αυτό δοκιμάσθηκαν τιμές από 5 έως και 10.

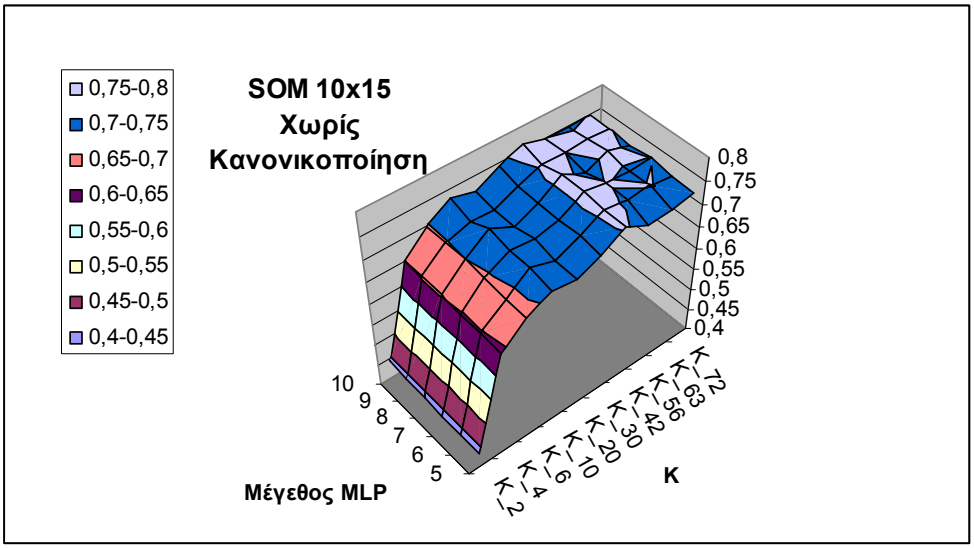
4.4.1 Αποτελέσματα κατηγοριοποίησης

4.4.1.1 Α' Σώμα Κειμένων

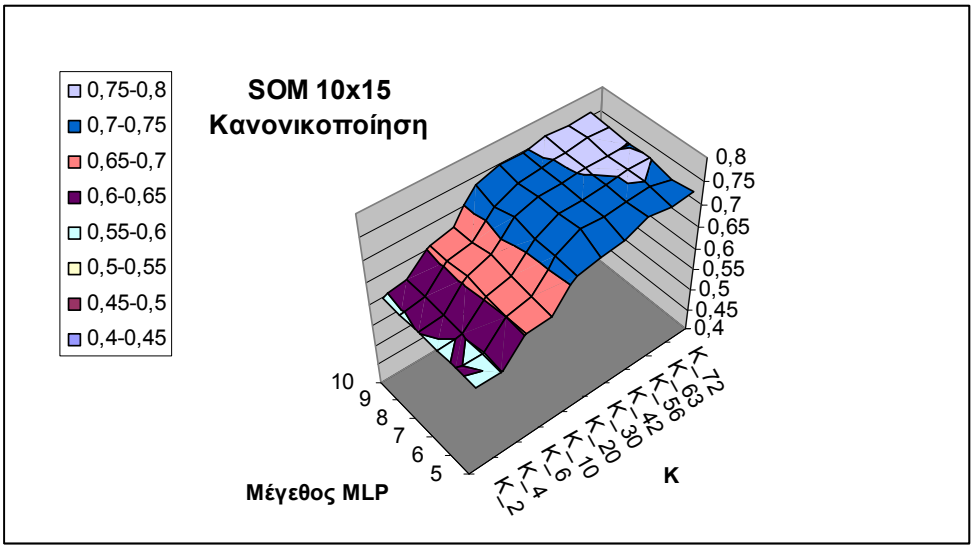
Η τελική ακρίβεια του συστήματος εκτιμήθηκε με την τεχνική της κυκλικής επαλήθευσης (leave-one-out) για 9 όμως σύνολα (επτά (7) για εκπαίδευση, ένα (1) για τερματισμό αυτής και ένα (1) για εκτίμηση της απόδοσης, άρα στο σύνολο 72 συνδυασμοί), όπως είχε εφαρμοσθεί και στην περίπτωση των πειραμάτων αναγνώρισης ύφους. Η ακρίβεια που αναφέρεται αφορά ρητά τη **μέση** απόδοση στα σύνολα δοκιμής που είναι άγνωστα κατά τη διαδικασία της εκπαίδευσης, για όλους τους δυνατούς συνδυασμούς.

[2] Για περισσότερες πληροφορίες ο αναγνώστης παραπέμπεται στην ενότητα 4.2.

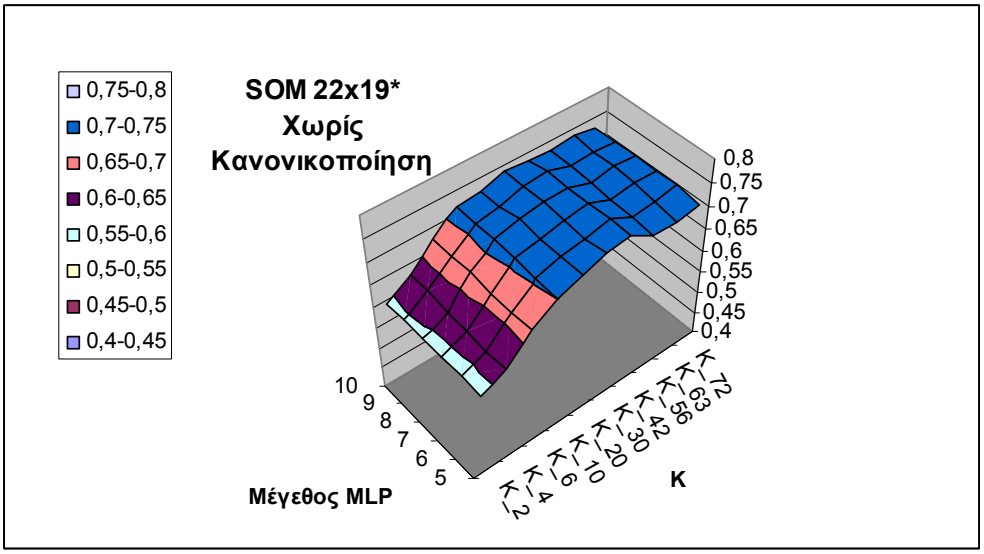
Για κάθε τρόπο μέτρησης των συμφραζομένων των λημμάτων (επιλογή εύρους ΑΒC 70%-85% ή 85%-90%, κανονικοποίηση ή μη) προκύπτει μία γραφική παράσταση για κάθε μέγεθος χάρτη των αποτελεσμάτων με άξονες τις τιμές *k* και το μέγεθος του MLP για τα κείμενα της Βουλής (βλ. σχήματα 4.8 έως 4.15, όπου ο αστερίσκος εκφράζει μεγέθη SOM που επιλέγονται αυτόματα από τη διαδικασία αρχικοποίησης).



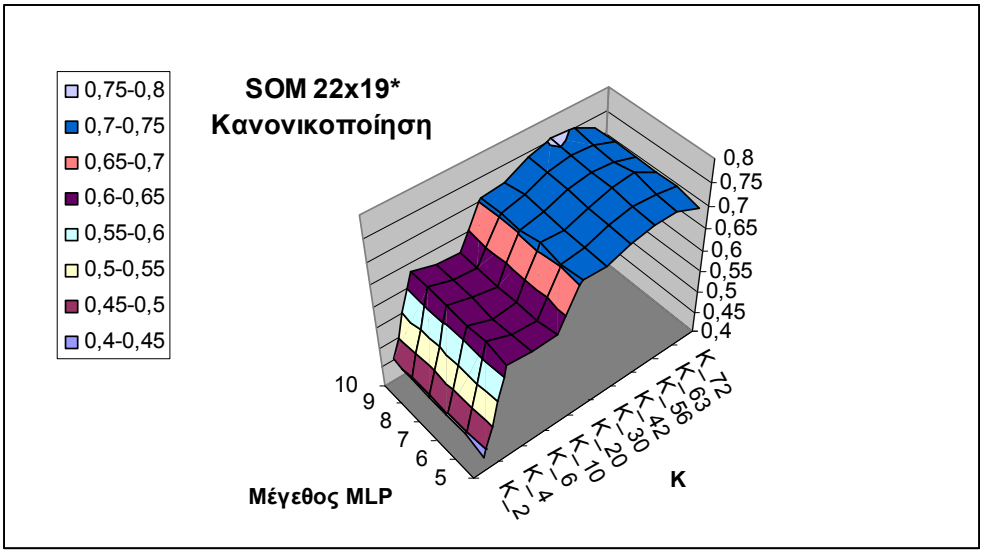
Σχήμα 4.8. Απόδοση του χάρτη SOM 10x15, χωρίς κανονικοποίηση, για το εύρος 70%-85% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



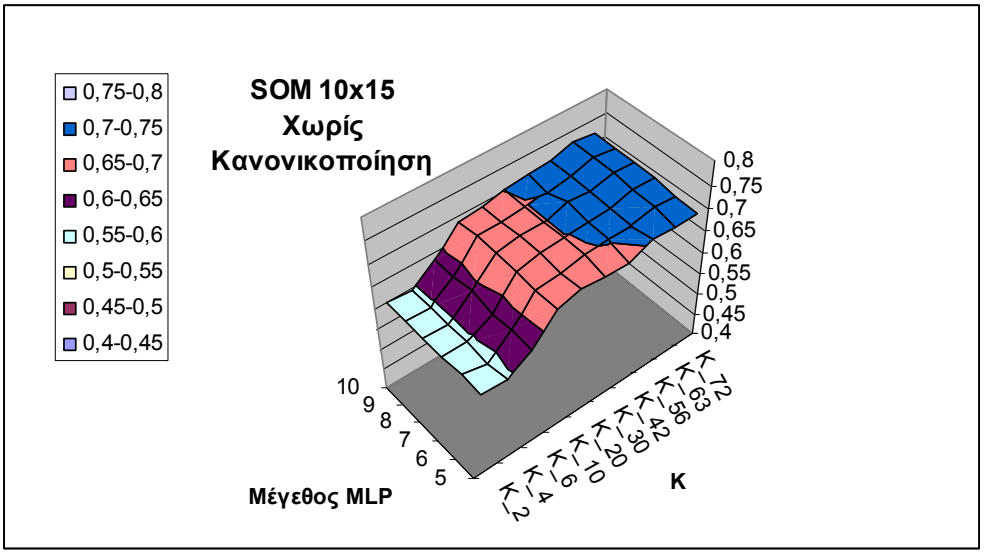
Σχήμα 4.9. Απόδοση του χάρτη SOM 10x15, με κανονικοποίηση, για το εύρος 70%-85% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



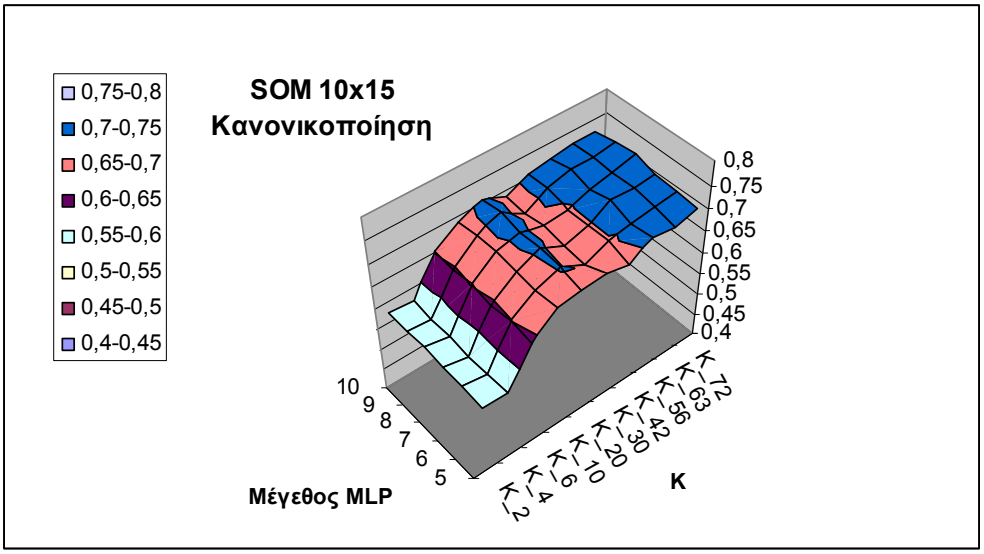
Σχήμα 4.10. Απόδοση του χάρτη SOM 22x19, αυτόματα προσδιοριζόμενου, χωρίς κανονικοποίηση, για το εύρος 70%-85% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



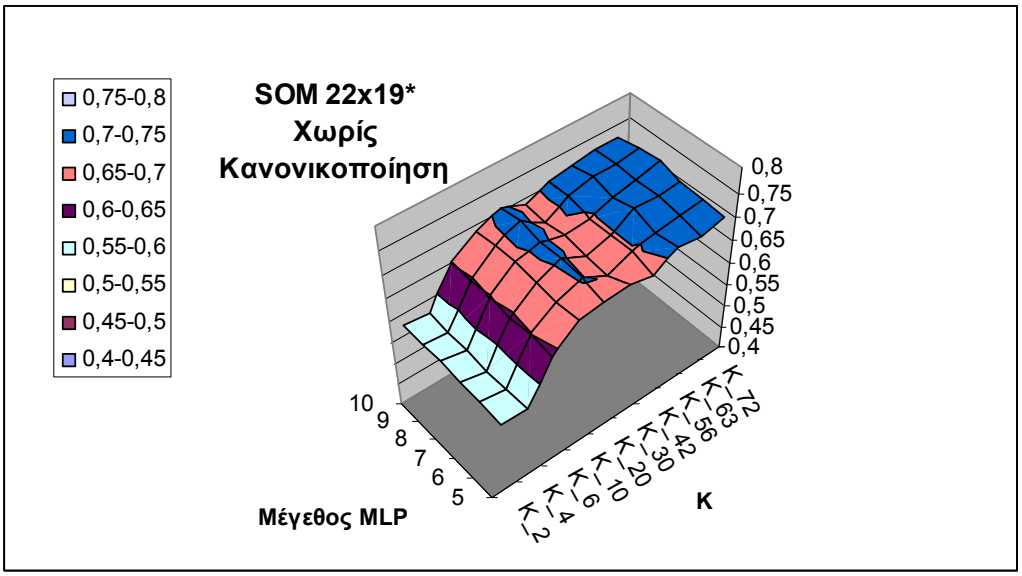
Σχήμα 4.11. Απόδοση του χάρτη SOM 22x19, αυτόματα προσδιοριζόμενου, με κανονικοποίηση, για το εύρος 70%-85% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



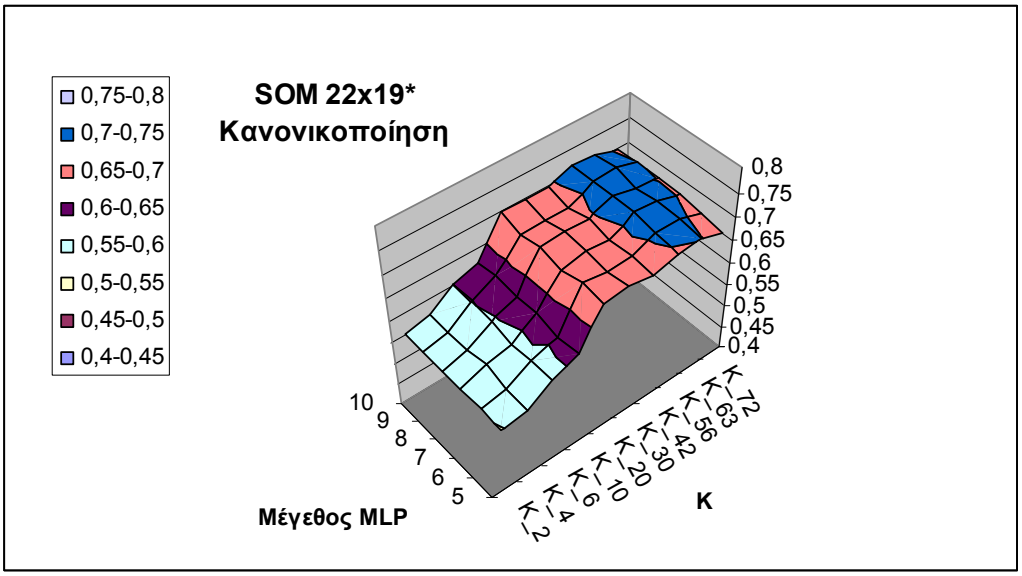
Σχήμα 4.12. Απόδοση του χάρτη SOM 10x15, χωρίς κανονικοποίηση, για το εύρος 85%-90% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



Σχήμα 4.13. Απόδοση του χάρτη SOM 10x15, με κανονικοποίηση, για το εύρος 85%-90% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



Σχήμα 4.14. Απόδοση του χάρτη SOM 22x19, αυτόματα προσδιοριζόμενου, χωρίς κανονικοποίηση, για το εύρος 85%-90% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)



Σχήμα 4.15. Απόδοση του χάρτη SOM 22x19, αυτόματα προσδιοριζόμενου, με κανονικοποίηση, για το εύρος 85%-90% ως προς το μέγεθος του MLP και την τιμή του *k* για τα δεδομένα της Βουλής (Α' σώμα κειμένων)

Η ακρίβεια κατηγοριοποίησης που παρατηρήθηκε υπερβαίνει το 70% για τους καλύτερους συνδυασμούς παραμέτρων. Αναλυτικότερα, το βέλτιστο σύστημα είναι αυτό που χρησιμοποιεί το εύρος 70%-85%, χωρίς κανονικοποίηση, με μέγεθος χάρτη 10x15, για *k=42* και 10 κρυφούς νευρώνες με απόδοση 76,5%. (βλ. σχήμα 4.8). Η απόδοση και των υπόλοιπων συνδυασμών κινείται περίπου στο ίδιο επίπεδο (π.χ. 75,9%, 74,1% 75,2%, 74% κ.λπ.), όταν χρησιμοποιείται επαρκής αριθμός ομάδων (τουλάχιστον 10) του k-means.

Από τα εν λόγω πειραματικά αποτελέσματα συνάγεται ότι τα συστήματα με εύρος 70%-85% είναι κατά κανόνα καλύτερα από αυτά με 85%-90%. Επιπλέον, ο ρόλος της κανονικοποίησης στα αποτελέσματα φαίνεται να είναι μηδαμινός, αφού συστηματικά δε συμβάλλει ούτε σε καλύτερη ούτε σε χειρότερη απόδοση, συνεπώς, δε διαφαίνεται κάποια χρησιμότητα στο να υποστούν τα δεδομένα επιπρόσθετη επεξεργασία. Με αντίστοιχο σκεπτικό φαίνεται ότι ο πιο μικρός χάρτης 10x15 αποδίδει πολλές φορές το ίδιο ή ακόμη και καλύτερα από τους πολύ μεγαλύτερους 22x19, με αποτέλεσμα να μην αιτιολογείται η χρησιμοποίησή τους.

Από τα σχήματα 4.8 – 4.15 μπορεί επίσης να διαφανεί ότι η διαφοροποίηση του μεγέθους του MLP δεν επιφέρει μεγάλη αλλαγή στα αποτελέσματα, καθώς η ακρίβεια κατηγοριοποίησης τροποποιείται κατά λιγότερο από 2%. Αντιθέτως, το *k* αναδεικνύεται σε σημαντικότερη παράμετρο, αφού με τη μεγάλη μείωσή του, δηλαδή κάτω από 6, η απόδοση του συστήματος χειροτερεύει. Είναι εντυπωσιακό ότι ακόμη και για μικρές τιμές του *k*, γύρω στο 10, το σύστημα σταθερά κατηγοριοποιεί τα κείμενα με ακρίβεια άνω του 70%. Με την αύξηση, ωστόσο, της τιμής του *k* πάνω από το 56, για το βέλτιστο σύστημα, η απόδοση χειροτερεύει, γεγονός που ενδεχομένως οφείλεται στην αύξηση της πολυπλοκότητας του διαχωριστικού προβλήματος που καλείται να επιλύσει το MLP.

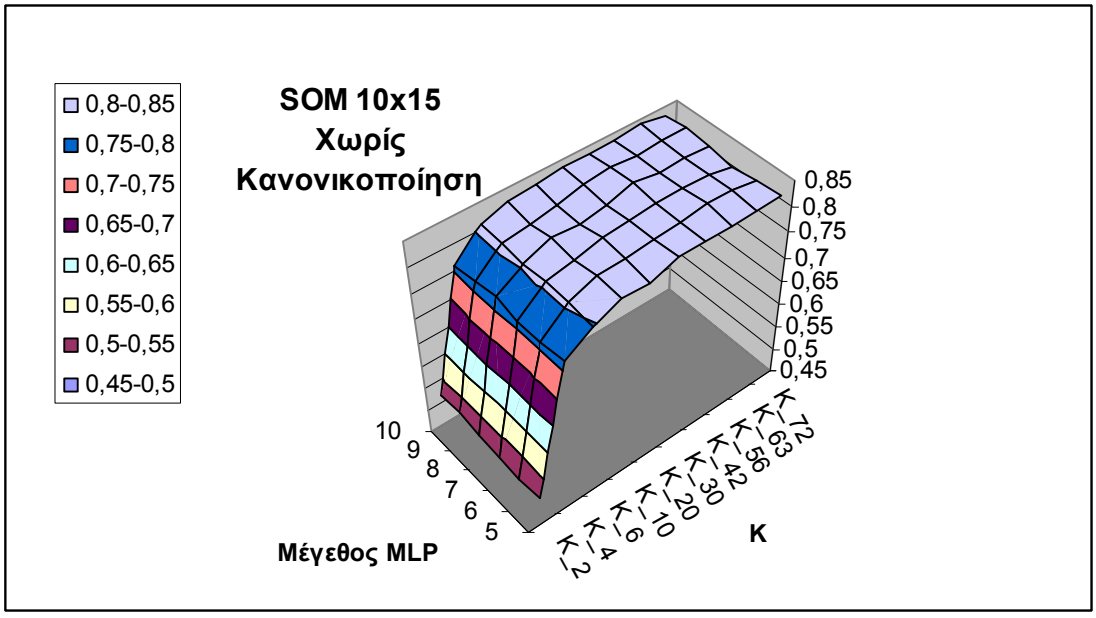
Συνοψίζοντας, **οι προτιμητέες παράμετροι** για το σύστημα, από πλευράς απόδοσης και οικονομίας σε υπολογιστική ισχύ, είναι οι ακόλουθες:

- **μη κανονικοποίηση**
- **εύρος 70%-85%** για την ανάλυση ABC
- **μέγεθος χάρτη 10x15**
- **απουσία συσχέτισης** μεγέθους του MLP και αποτελεσμάτων[3]
- **κύμανση** της **τιμής** της **παραμέτρου *k*** στο εύρος **10-56**, καθορίζοντας έτσι το όριο μεταξύ απόδοσης του συστήματος και πολυπλοκότητας (level-of-detail).

[3] Εντούτοις, τα αποτελέσματα για μεγαλύτερο πλήθος νευρώνων είναι καλύτερα, έστω και κατ' ελάχιστο.

Το σύστημα με τις προαναφερθείσες παραμέτρους με *k=42* και 10 κρυφούς νευρώνες θα αποτελέσει το σύστημα αναφοράς για τις συγκρίσεις στα υπόλοιπα πειράματα.

Τα προαναφερθέντα πειράματα αφορούν στο πρόβλημα της οργάνωσης των κειμένων της Βουλής σε 4 κατηγορίες. Από τις κατηγορίες αυτές, η κατηγορία Εσωτερικά είναι αρκετά γενική και μπορεί να περιλαμβάνει και κείμενα με θέμα την Παιδείας ή και την Οικονομία. Για το λόγο αυτό, τα πειράματα επαναλήφθηκαν με λιγότερα δεδομένα για 3 μόνο κατηγορίες. Στο σχήμα 4.16 παρουσιάζεται ενδεικτικά η απόδοση του συστήματος με τα καλύτερα αποτελέσματα στο πείραμα με τις 3 κατηγορίες.

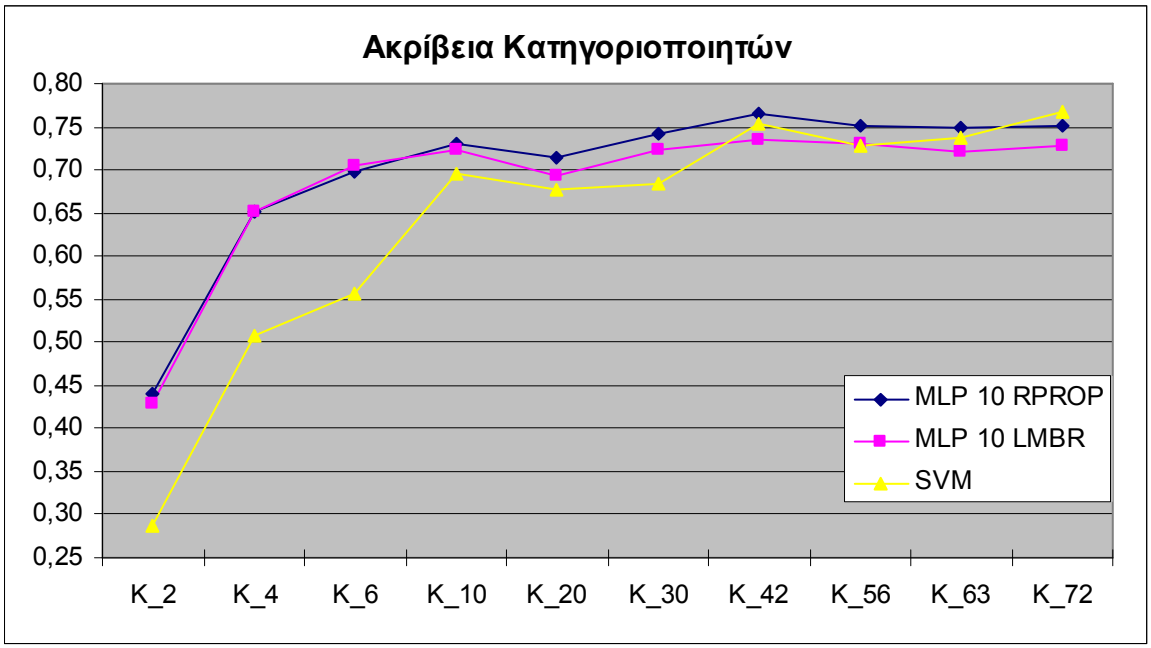


Σχήμα 4.16. Απόδοση του χάρτη SOM 10x15, χωρίς κανονικοποίηση, για το εύρος 70%-85% ως προς το μέγεθος του MLP και την τιμή του *k* για το πρόβλημα κατηγοριοποίησης για τρεις κατηγορίες για τα δεδομένα της Βουλής (Α' σώμα κειμένων)

Τα αποτελέσματα δείχνουν ότι η απόδοση του συστήματος αυξήθηκε κατά περίπου 10% σε σχέση με το πείραμα των 4 κατηγοριών. Η καλύτερη απόδοση ήταν 84,3% για *k=56* και 10 νευρώνες στο κρυφό επίπεδο. Επιπλέον, και σε αυτά πειράματα διαφάνηκε ότι το μέγεθος του MLP δεν επιδρά σημαντικά στα αποτελέσματα, ενώ το σύστημα δίνει αρκετά αποδεκτά αποτελέσματα ακόμη και για μικρές τιμές του *k*.

Στη συνέχεια έγινε προσπάθεια να βελτιωθεί η απόδοση του συστήματος χρησιμοποιώντας κάποιο διαφορετικού είδους κατηγοριοποιητή. Συγκεκριμένα, δοκιμάσθηκαν αφενός για το MLP ο πιο σύνθετος αλγόριθμος εκπαίδευσης LMBR, ο οποίος περιορίζει και την πολυπλοκότητα του μοντέλου, και αφετέρου το μοντέλο SVM, τα οποία είχαν εφαρμοσθεί με επιτυχία στα πειράματα αναγνώρισης ύφους.

Για το MLP δοκιμάσθηκαν τα ίδια μεγέθη δικτύων, ενώ για το SVM δοκιμάσθηκαν διάφορες τιμές της ακτίνας του Γκαουσσιανού πυρήνα και της τιμής *C* του SVM, εκ των οποίων παρουσιάζονται οι καλύτερες. Το σύνολο των πειραμάτων επαναλήφθηκε (εύρος ABC: 70%-85%, 85%-90%, κανονικοποίηση, μέγεθος MLP, τιμές *k*), ώστε να διαπιστωθεί κατά πόσο οι νέοι αλγόριθμοι μπορούν να επιτύχουν καλύτερα αποτελέσματα. Στη συνέχεια, για λόγους οικονομίας χώρου παρουσιάζεται η συγκριτική απόδοση των κατηγοριοποιητών (σχήμα 4.3) μόνο για εκείνη τη δομή χάρτη που απέδωσε καλύτερα στα προηγούμενα πειράματα των 4 κατηγοριών που χρησιμοποιούσαν τη μέθοδο εκπαίδευσης RPROP για το MLP.



Σχήμα 4.17. Συγκριτική απόδοση των 3 κατηγοριοποιητών του χάρτη SOM 10x15, χωρίς κανονικοποίηση, για το εύρος 70%-85%, με μέγεθος MLP 10 νευρώνων και με παράμετρο την τιμή του *k* για το πρόβλημα κατηγοριοποίησης για 4 κατηγορίες

Το συμπέρασμα που προκύπτει είναι ότι ο αλγόριθμος RPROP είναι σταθερά καλύτερος από τον LMBR, ο οποίος είναι και πολυπλοκότερος ως προς τον τρόπο λειτουργίας του, αλλά και απαιτητικότερος σε μνήμη και αριθμητικές πράξεις (αντιστροφή πινάκων κ.λπ.). Το καλύτερο αποτέλεσμα με βάση τα προηγούμενα πειράματα προκύπτει για *k=42* και 10 κρυφούς νευρώνες, το οποίο είναι 76,5%. Το αποτέλεσμα αυτό υπερβαίνεται από το SVM για *k=72*, αφού επιτυγχάνει ακρίβεια 76,8%.

Εντούτοις αυτό είναι και το μοναδικό σημείο που το SVM υπερέχει του MLP RPROP, αφού για όλες τις άλλες τιμές του *k* υστερεί. Επιπλέον, είναι αμφίβολο το εάν η βελτίωση 0,03% δικαιολογεί τη χρήση 30 επιπρόσθετων χαρακτηριστικών. Λαμβάνοντας υπόψιν τα αποτελέσματα, σε συνδυασμό με την απλότητα της RPROP, κρίνεται ότι είναι η προτιμητέα μέθοδος για το προτεινόμενο σύστημα.

Για λόγους οικονομίας χώρου στον παρακάτω πίνακα (4.3), παρουσιάζονται τα αποτελέσματα για όλους τους παραπάνω χάρτες που εξετάσθηκαν χωρίς την κανονικοποίηση των διανυσμάτων, ωστόσο τα συμπεράσματα που προέκυψαν και από τους υπόλοιπους χάρτες ήταν αντίστοιχα.

Όρια ABC	**Μέγεθος SOM**	**RPROP** *h=8 k=70*	**LMBR** *h=8 k=70*	**SVM** *k=70*	**RPROP** *(καλύτερη απόδοση)*	**LMBR** *(καλύτερη απόδοση)*	**SVM** *(καλύτερη απόδοση)*
70% 85%	**[25 17]**	72,7%	70,5%	72,7%	**74,1%** *h=9 & k=40*	73,5% *(h=8 & k=40)*	75,6% *k=42*
70% 85%	**[10 15]**	74,4%	72,9%	76,5%	**76,5%** *h=10 & k=40*	74,7% *(h=7 & k=70)*	**76,5%** *k=70*
85% 90%	**[26 16]**	72,3%	70,5%	74,0%	74,0% *h=9 & k=50*	72,4% *(h=10 & k=60)*	**74,7%** *k=56*
85% 90%	**[10 15]**	71,8%	70,3%	72,9%	**72,6%** *h=10 & k=60*	71,4% *h=9 & k=70*	72,9% *k=70*

Πίνακας 4.3. Συγκριτική απόδοση των κατηγοριοποιητών LMBR, RPROP και SVM χωρίς κανονικοποίηση για τα κείμενα της Βουλής (Α' σώμα κειμένων) για δεδομένες παραμέτρους μαζί με εκείνες που οδήγησαν στην καλύτερη απόδοση

Ένα σημαντικό χαρακτηριστικό που παρουσιάζει η LMBR είναι ότι η συνάρτηση αξιολόγησης που χρησιμοποιεί (τύπος 3-21) επιβάλλει μικρές τιμές στα βάρη του δικτύου. Με τον τρόπο αυτό και δεδομένου ότι οι τιμές εισόδου είναι κανονικοποιημένες στο ίδιο εύρος, είναι εφικτή η αξιολόγηση των καλύτερων χαρακτηριστικών ανάλογα με τις απόλυτες τιμές των βαρών (τύπος 3-35) για όλους τους συνδυασμούς συνόλων εκπαίδευσης - επαλήθευσης - δοκιμής, όπως ακριβώς έγινε στα πειράματα αναγνώρισης συγγραφέα. Με βάση την περιγραφή του συστήματος οι ομάδες λέξεων που δημιούργησε το μοντέλο SOM αποτελούν και τα χαρακτηριστικά μέσω των οποίων έγινε η αναπαράσταση των κειμένων. Είναι εφικτό λοιπόν να βρεθούν εκείνες οι ομάδες που είχαν τη μεγαλύτερη συμβολή στο τελικό αποτέλεσμα.

Στον πίνακα 4.4 παρουσιάζονται οι τέσσερις καλύτερες ομάδες, όπως αυτές αξιολογήθηκαν από την LMBR για το σύστημα με μέγεθος χάρτη 10x15, MLP με 10 νευρώνες στο κρυφό επίπεδο και τιμή *k* ίση με το 70, που είναι και η μέγιστη. Η μέγιστη τιμή για το *k* επιλέχθηκε, επειδή ο μεγαλύτερος αριθμός ομάδων συνεπάγεται λιγότερες λέξεις ανά ομάδα, γεγονός που οδηγεί σε πιο κατανοητά και ευδιάκριτα πειραματικά αποτελέσματα. Αυτός είναι και ο λόγος που τα αποτελέσματα παρουσιάζονται μόνο για το μικρότερο σώμα κειμένων της Βουλής (Α' σώμα κειμένων).

Ομάδα 1 *(124 λέξεις συνολικά)*	**Ομάδα 2** *(209 λέξεις συνολικά)*
"εξοντώνω", "επαρκώς", "σπατάλη", "συγκριτικά", "ενέχω", "απορροφώ", "πακέτο", "επενδύω", "κατασπατάληση", "εντυπωσιακός", "χρηματοδοτώ", "εξοικονόμηση", "εγκληματικός", "κατάφωρος", "προσελκύω", "τρις", "απορρόφηση", "αθέμιτος", "χρηματοδότηση", "επιπόλαιος", "υπερκαλύπτω", "ώθηση", "πλουτοπαραγωγικός", "συνυπολογίζω", "συγχρηματοδότηση", "αντιπλημμυρικός", "εμπόρευμα", "κατατάσσω", "ξεμπλοκάρω", "επιδοματικός", "μακέτα", "γιγαντιαίος", "πλουτισμός", "πασιφανής", "κατασπαταλώ", "επωφελούμαι", "επενδυτικός", "λιγοστεύω", "δαπανηρός", "αυτοχρηματοδότηση", "αθλητισμός", "άντληση", "πάγωμα", "φουσκώνω", "διασπάθιση", "ανακύκλωση", "αντιμονοπωλιακός", "λεηλασία", "απορροφητικότητα", "πλαίσιο", "κατανομή", "ορθολογικός", "πακτωλός"	"εκπόνηση", "αξιόποινος", "αδίκημα", "απόσυρση", "δικηγορικός", "Βουλή", "υπόμνημα", "διαμαρτυρόμενος", "διανοούμαι", "αρχείο", "τεκμηριώνω", "συντάσσω", "διακομματικός", "συστήνω", "ψήφιση", "συνέδριο", "πρακτικά", "μεθόδευση", "πρωθυπουργία", "καταψηφίζω", "ομιλητής", "αόριστος", "ψευδορκία", "ευγένεια", "φακέλωμα", "γραμματεία", "ανησυχία", "επικυρώνω", "συμπληρωματικός", "παράβαση", "συνέντευξη", "απολογισμός", "απόσπασμα", "προσβάλλω", "αίτηση", "δυσφήμηση", "ποινικός", "συμπαιγνία", "συνδικαλισμός", "υπηρεσιακός", "φάκελος", "επιτροπή", "αποσύρω", "πόρισμα", "συμβασιούχος", "έγγραφο", "λεπτομερώς", "τροποποιώ", "διευθυντικός", "διαμαρτυρία", "εισαγγελικός", "ποινικοποίηση", "αντιπροσωπεία", "εισαγγελέας", "αντισυνταγματικός", "νόμος", "διαπλεκόμενος", "Άρειος", "ευρωκοινοβούλιο"
Ομάδα 3 *(39 λέξεις συνολικά)*	**Ομάδα 4** *(25 λέξεις συνολικά)*
"συμβουλευτικός", "επιστροφή", "αναμφισβήτητα", "απόσταση", "παραδεκτός", "ενδεχόμενος", "λήψη", "επαγρυπνώ", "αποφασιστικός", "επηρεάζω", "υπαρκτός", "καθιστώ", "θεμελιώνω", "τολμηρός", "εκτόνωση", "ειλικρινής", "ώριμος", "Σένγκεν", "δεικνύω", "ΟΚΕ", "ιδίως", "προσχωρώ", "συσχετισμός", "πειθώ", "ριζικά", "καμπή", "διαπραγματευτικός", "πνεύμα", "μετατρέπω", "αντιμέτωπος", "οπωσδήποτε", "ερμηνεύω", "ιεραρχικός", "εκβιασμός", "υποδεέστερος", "ταλάντευση", "συγκρότηση", "επικαλούμενος", "συμβόλαιο"	"εισακτέος", "διατροφή", "διορισμός", "οδικός", "ασεπ", "πιλοτικός", "βρεφονηπιακός", "μεταπτυχιακός", "παιδικός", "σταθμός", "μάνα", "βιβλιοθήκη", "ηλικία", "ώσπου", "σύζυγος", "ΚΕΘΕΑ", "εξήμισι", "προξενείο", "χρονών", "υποπολλαπλάσιο", "δεκάξι", "ελαφρά", "συγγραφέας", "Ανθόπουλος", "ποδάρι"

Πίνακας 4.4. Οι τέσσερις καλύτερες ομάδες λέξεων, όπως αυτές αξιολογήθηκαν από την LMBR για τα κείμενα της Βουλής (Α' σώμα κειμένων)

Όπως φαίνεται στον πίνακα 4.4,η καλύτερη ομάδα (Ομάδα 1) περιέχει λέξεις σχετικές με οικονομικά κυρίως θέματα, ενώ η δεύτερη ομάδα (Ομάδα 2) περιέχει λέξεις που αφορούν νομικά-διοικητικά κυρίως θέματα. Η τρίτη ομάδα είναι γενικότερου περιεχομένου με όρους σχετικούς με την πολιτική, ενώ η τέταρτη ομάδα περιλαμβάνει λέξεις σχετικές με θέματα παιδείας. Πρέπει εδώ να αναφερθεί ότι οι δύο πρώτες ομάδες περιέχουν πάνω από 50 λέξεις, ωστόσο στον πίνακα εμφανίζονται επιλεκτικά 50 από αυτές, ώστε να είναι ευανάγνωστα τα αποτελέσματα.

Η μέτρηση των συμφραζόμενων για την περιγραφή των λημμάτων κατά τη φάση της δημιουργίας του χάρτη βασίστηκε στην παραδοχή ότι οι λέξεις που βρίσκονται στην ίδια περίοδο θα πρέπει να συνδέονται νοηματικά. Η σειρά εμφάνισης των λέξεων, δεδομένης της ελεύθερης σειράς των όρων στη Νέα Ελληνική, και κατά συνέπεια οι αποστάσεις μεταξύ των λέξεων δεν ελήφθησαν υπόψιν. Για να ελεγχθεί ωστόσο η προσέγγιση αυτή, δοκιμάσθηκε ένας διαφορετικός τρόπος μέτρησης των χαρακτηριστικών. Πιο συγκεκριμένα, η συνεμφάνιση μίας λέξης με ένα χαρακτηριστικό (δηλαδή λέξη της ομάδας Β όπως προέκυψε από την ανάλυση ABC) σε επίπεδο περιόδου δε θα έχει πάντα την ίδια τιμή (δηλαδή 1, όπως στα προηγούμενα πειράματα), αλλά θα λαμβάνει μία τιμή μεταβαλλόμενη αντιστρόφως ανάλογα της απόστασης μεταξύ των λέξεων, δηλαδή όσο μεγαλύτερη είναι η απόσταση μεταξύ δύο λέξεων, τόσο λιγότερο θα μετράει η συνεμφάνισή τους. Για αυτές τις μετρήσεις χρησιμοποιήθηκαν οι συναρτήσεις (4-14) και (4-15):

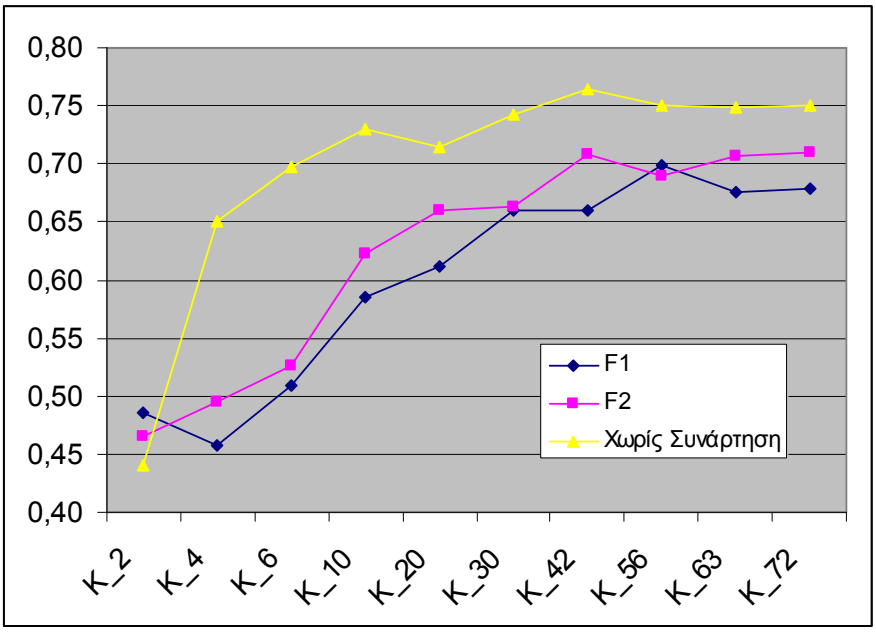
F1: $exp(\frac{1-d}{L})$ (4-14)

όπου *L* είναι μία παράμετρος, που στα πειράματα είχε την τιμή 7 και εκφράζει το πόσο απότομα μειώνεται η τιμή της συνάρτησης, και *d* είναι η απόσταση σε αριθμό λέξεων μεταξύ της λέξης-στόχου και της λέξης-χαρακτηριστικού.

F2: $\frac{1}{exp(-A+B\cdot d)+1}$ (4-15)

όπου *B* είναι μία παράμετρος, που στα πειράματα είχε την τιμή 2, *A* είναι επίσης μία παράμετρος που έχει την τιμή 15 και *d* είναι η απόσταση σε αριθμό λέξεων μεταξύ της λέξης-στόχου και της λέξης-χαρακτηριστικού.

Όταν χρησιμοποιήθηκαν οι συναρτήσεις αυτές αγνοήθηκαν από τα συμφραζόμενα οι λειτουργικές λέξεις. Αυτό έγινε, προκειμένου να μην επηρεάσουν τις αποστάσεις μεταξύ των λέξεων περιεχομένου. Τα αποτελέσματα που προέκυψαν σχετικά με την απόδοση του συστήματος φαίνονται στο σχήμα 4.18:

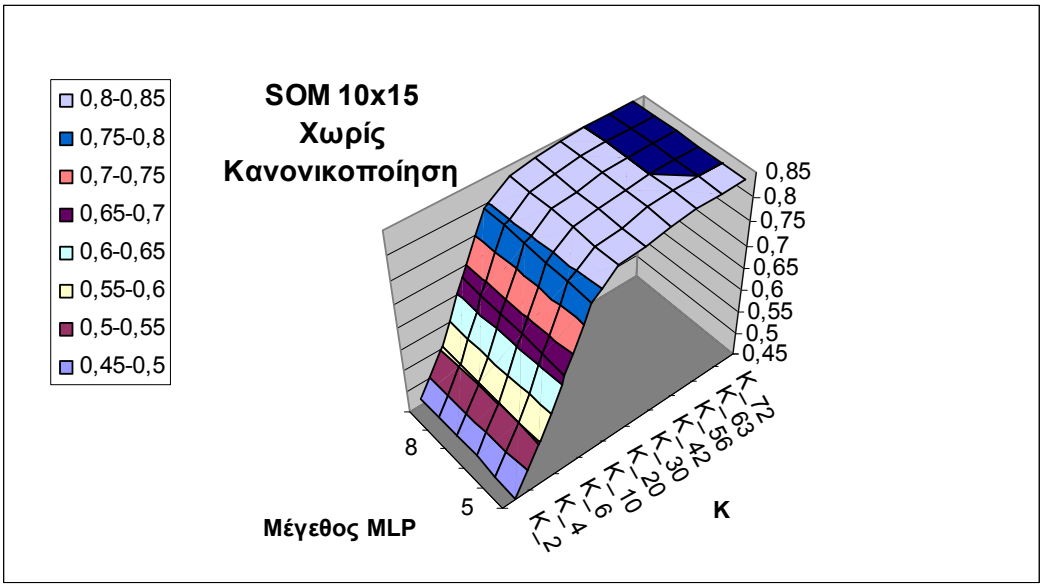


Σχήμα 4.18. Συγκριτική απόδοση των 3 διαφορετικών τρόπων μέτρησης συμφραζομένων με μέγεθος του κρυφού επιπέδου MLP ίσο με 10 ως προς την τιμή του *k* για το πρόβλημα κατηγοριοποίησης κειμένων της Βουλής σε τέσσερις κατηγορίες

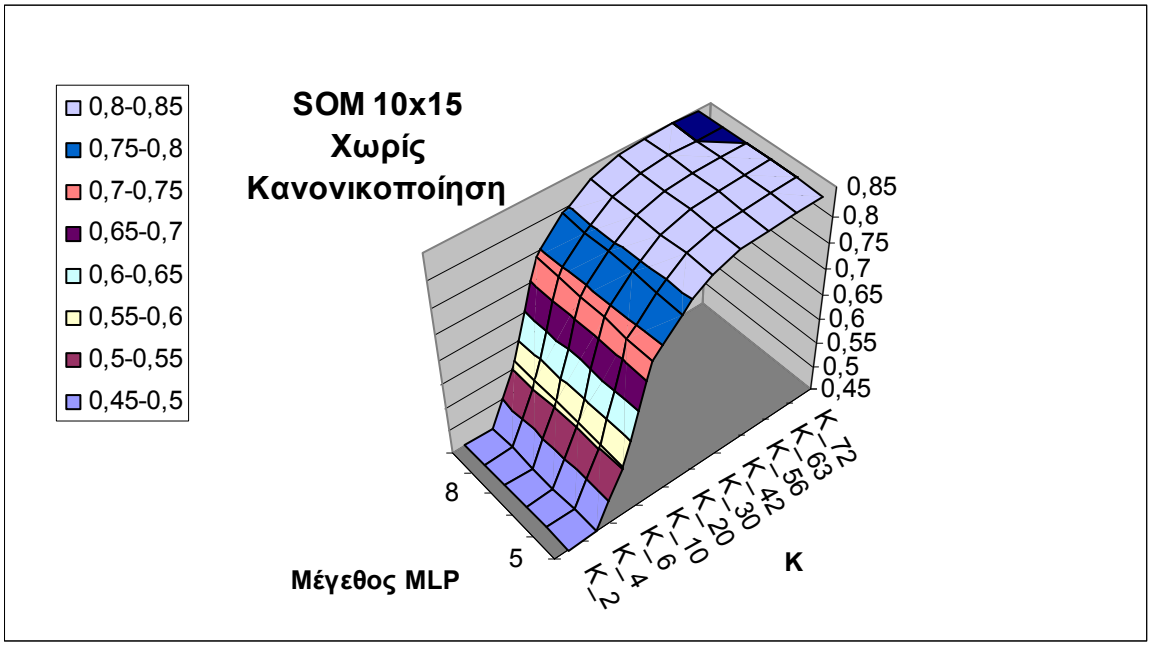
Από τα παραπάνω αποτελέσματα γίνεται σαφές ότι οι μετρήσεις συμφραζομένων χωρίς χρήση συναρτήσεων σχετικής απόστασης αποδίδει πολύ καλύτερα, αφού οδηγεί σχεδόν πάντα σε απόδοση βελτιωμένη κατά 5%. Μία πιθανή ερμηνεία είναι ότι οι μετρήσεις συμφραζομένων οδηγούν σε αρκετά μικρές τιμές, οι οποίες γίνονται ακόμη μικρότερες με τη χρήση των συναρτήσεων αυτών. Το γεγονός αυτό σε συνδυασμό με τη μεγάλη διάσταση του διανύσματος εισόδου μπορεί να επεξηγήσει τη χειρότερη απόδοση του συστήματος με τη χρήση συναρτήσεων σχετικής απόστασης.

4.4.1.2 Β' Σώμα Κειμένων

Το προτεινόμενο σύστημα θα είχε ακόμα μεγαλύτερο πρακτικό ενδιαφέρον κατά την εφαρμογή του σε ένα πολύ μεγαλύτερο σύνολο δεδομένων. Έτσι, εφαρμόσθηκε με τον ίδιο ακριβώς τρόπο στο Β' σύνολο δεδομένων (κείμενα εφημερίδων), που ήταν πολύ μεγαλύτερο (36.450). Ο χάρτης που επελέγη είχε διαστάσεις 10x15, ο οποίος κρίθηκε προτιμητέος στα πειράματα του σώματος κειμένων της Βουλής. Επιπλέον επιλέχθηκε ο αλγόριθμος εκπαίδευσης RPROP, ο οποίος είναι ταχύτερος από τον LMBR και παρουσιάζει μεγαλύτερη ακρίβεια κατηγοριοποίησης. Τα σχετικά αποτελέσματα δίνονται στα σχήματα (4-19) και (4-20):



Σχήμα 4.19. Απόδοση του χάρτη SOM 10x15, χωρίς κανονικοποίηση, για το εύρος 70%-85% ως προς το μέγεθος του MLP και την τιμή του *k* για το πρόβλημα κατηγοριοποίησης στο πολύ μεγαλύτερο σύνολο δεδομένων των εφημερίδων (Β' σώμα κειμένων)



Σχήμα 4.20. Απόδοση του χάρτη SOM 10x15, χωρίς κανονικοποίηση, για το εύρος 85%-90% ως προς το μέγεθος του MLP και την τιμή του *k* για το πρόβλημα κατηγοριοποίησης στο πολύ μεγαλύτερο σύνολο δεδομένων των εφημερίδων (Β' σώμα κειμένων)

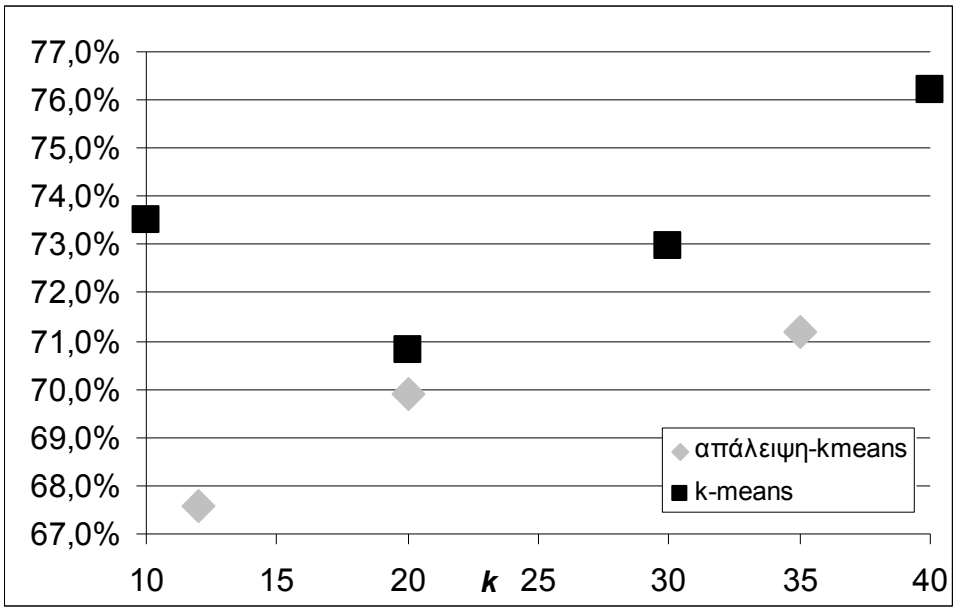
Είναι ιδιαίτερα σημαντικό και ελπιδοφόρο το ότι το σύστημα αντεπεξήλθε σε ένα πολύ δυσκολότερο και πολυπλοκότερο πρόβλημα, δεδομένου ότι το δεύτερο σύνολο αφορούσε 6 θεματικές κατηγορίες. Η καλύτερη απόδοση που επιτυγχάνει το σύστημα για το εύρος 70%-85%, για *k=72* και MLP με 10 νευρώνες στο κρυφό επίπεδο, είναι 85,4%, δηλαδή ένα πολύ υψηλό ποσοστό. Εξίσου σημαντικό είναι το γεγονός ότι ακόμη και με *k=56*, το σύστημα επιτυγχάνει απόδοση 85,2%, ενώ για *k=42*, που ήταν και το προηγούμενο σύστημα αναφοράς στην ενότητα 4.4.1.1, επιτυγχάνει 84,6%.

Επιπλέον, στα σχήματα 4.19 και 4.20 παρατηρείται ότι το μέγεθος του MLP δεν επηρεάζει σημαντικά την ακρίβεια κατηγοριοποίησης. Το καλύτερο αποτέλεσμα για το εύρος 85%-90% είναι ελάχιστα χειρότερο, με τιμή 85,3% για *k=72* και MLP με 10 νευρώνες στο κρυφό επίπεδο. Όταν όμως επιλέγονται λιγότερες ομάδες, π.χ. *k=56* και *k=42*, τα αποτελέσματα διατηρούνται πάλι σε υψηλό επίπεδο, 84,8% και 84,4% αντίστοιχα, αλλά είναι συγκριτικά χειρότερα από τα αποτελέσματα για το εύρος 70%-85%. Η επιλογή του εύρους 70%-85% φαίνεται ότι είναι και εδώ προτιμότερη.

Τα αποτελέσματα αυτά είναι ιδιαίτερα ενθαρρυντικά και δείχνουν ότι ο μεγαλύτερος όγκος των δεδομένων μπορεί να οδηγήσει σε πιο περιγραφικούς χάρτες ικανούς να απεικονίσουν καλύτερα τις θεματικές κατηγορίες των λημμάτων και κατ' επέκταση των κειμένων.

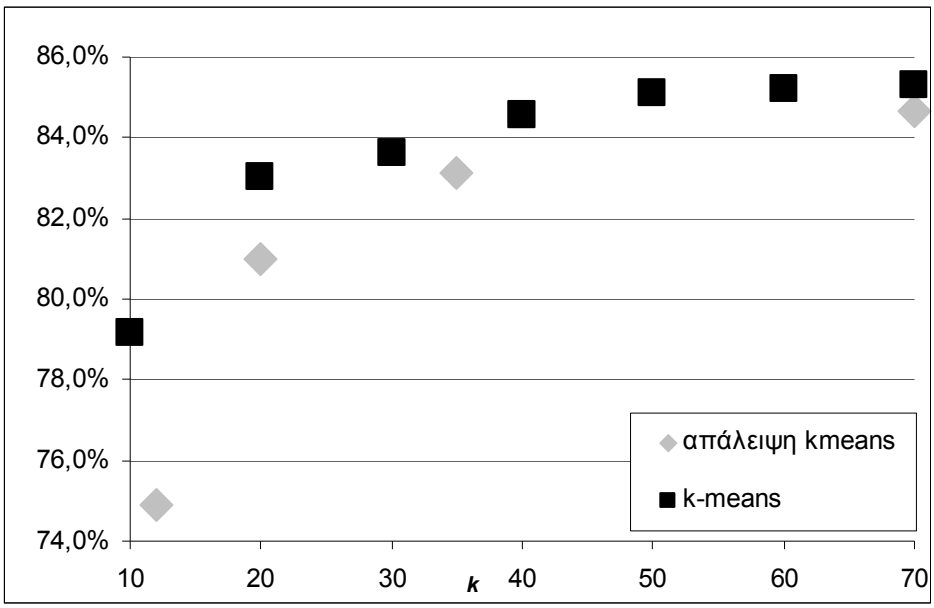
4.4.2 Αποτελέσματα με παράλειψη του υποσυστήματος *k-means*

Η απόδοση του συστήματος εκτιμήθηκε και όταν στο σύστημα δε συμπεριελήφθη το στάδιο της μείωσης της διάστασης, με στόχο την απλοποίηση του συστήματος. Όταν παρακάμπτεται το *k*-means, για το χάρτη λέξεων SOM επιλέγεται ένα πολύ μικρότερο πλήθος ομάδων, αφού το πλήθος των ομάδων του χάρτη πλέον καθορίζει τη διάσταση της εισόδου του κατηγοριοποιητή MLP. Διαφορετικά μεγέθη πλέγματος δοκιμάσθηκαν για το SOM και πιο συγκεκριμένα πλέγματα με 5x7=35 νευρώνες, 4x5=20 νευρώνες και 3x4=12 νευρώνες, όπου ο αριθμός των νευρώνων είναι κοντά στις τιμές *k* του *k-means* που χρησιμοποιήθηκε στα προηγούμενα πειράματα. Στη συνέχεια παρουσιάζονται τα αποτελέσματα για MLP με 8 κρυφούς νευρώνες, τα οποία εκπαιδεύτηκαν με τον αλγόριθμο RPROP για τα κείμενα της Βουλής (σχήμα 4.21) και τα κείμενα των εφημερίδων (σχήμα 4.22).



Σχήμα 4.21. Συγκριτική απόδοση του συστήματος χωρίς το *k-means* στο σώμα κειμένων της Βουλής (Α' σώμα κειμένων)

Η παράλειψη της εκτέλεσης του *k*-means επιφέρει μεν μείωση στην ακρίβεια του συστήματος για το ίδιο περίπου πλήθος ομάδων – η ακρίβεια δεν υπερβαίνει το 72,0%, ποσοστό πολύ μικρότερο από εκείνο που επιτυγχάνει το πλήρες σύστημα (76,5%) - ωστόσο, οδηγεί σε ταχύτερη εκπαίδευση του συστήματος, αλλά και σε εξαρχής μικρότερο μέγεθος του χάρτη SOM, με αποτέλεσμα ταχύτερες αναζητήσεις για το νικητή νευρώνα κατά τη διαδικασία της εκπαίδευσης, αλλά και λιγότερες ενημερώσεις γειτονικών νευρώνων. Η εν λόγω βελτιστοποίηση στην απόδοση μπορεί να είναι ιδιαίτερα χρήσιμη, όταν το σύστημα εφαρμόζεται σε πολύ μεγάλο πλήθος δεδομένων.



Σχήμα 4.22. Συγκριτική απόδοση του συστήματος χωρίς το k-means στο σώμα κειμένων των εφημερίδων (Β' σώμα κειμένων)

Η παράλειψη της εκτέλεσης του *k*-means στο Β' σώμα κειμένων, όπου η πρακτική αξία του εγχειρήματος είναι μεγαλύτερη (σχήμα 4.22), οδηγεί σε ελάχιστα μικρότερη ακρίβεια κατηγοριοποίησης για τα μεγέθη 5x7 (= 35 νευρώνες) και 7x10 (= 70 νευρώνες): 83,1% και 84,6% αντίστοιχα, σε σχέση με την τιμή 85,2% που επιτυγχάνει το πλήρες σύστημα. Λαμβάνοντας υπόψιν το μεγάλο μέγεθος του Β' σώματος κειμένων (συνολικά 36.450 κείμενα), η παράλειψη του *k*-means μπορεί να είναι προτιμητέα ακόμη και αν συνεπάγεται μείωση της ακρίβειας κατηγοριοποίησης, καθώς αυτή είναι ελάχιστη.

4.4.3 Αποτελέσματα ανεξάρτητης αξιολόγησης χαρτών

4.4.3.1 Πειράματα αξιολόγησης με λέξεις-κλειδιά

Στα πλαίσια του ευρύτερου στόχου της αξιολόγησης του συστήματος, εκτελέστηκε ένα σύνολο από πειράματα με σκοπό την αυτόνομη αξιολόγηση του χάρτη λέξεων που δημιουργήθηκε χωρίς επίβλεψη κατά το πρώτο στάδιο του συστήματος με τη χρήση του νευρωνικού δικτύου SOM. Πιο συγκεκριμένα για διαφορετικά μεγέθη και παραμέτρους χάρτη, όπως το εύρος του ABC (70%-85% και 85%-90%), η κανονικοποίηση ή μη και το μέγεθος του χάρτη, εξετάσθηκαν οι τιμές των κριτηρίων για τις διαδικασίες που παρουσιάσθηκαν στο κεφάλαιο 4.3 τόσο με χρήση λέξεων-κλειδιών όσο και με τη χρήση του θησαυρού Eurovoc. Κύριος στόχος της προσπάθειας αυτής είναι να διαπιστωθεί το κατά πόσο η ανεξάρτητη ενδιάμεση αξιολόγηση του χάρτη λέξεων μπορεί να συντελέσει στην επιλογή των κατάλληλων εκείνων παραμέτρων, που θα οδηγήσουν στην καλύτερη αναπαράσταση των κειμένων στο χώρο προτύπων με αποτέλεσμα τον καλύτερο διαχωρισμό αυτών βάσει του περιεχομένου τους.

Στους πίνακες 4.5 έως 4.8 παρουσιάζονται, για τις διαφορετικές παραμέτρους του χάρτη που εξετάσθηκαν (εύρος ABC, κανονικοποίηση, μέγεθος χάρτη), οι τιμές των επιλεγμένων κριτηρίων, που παρουσιάσθηκαν αναλυτικά στην ενότητα 4.3.1 και αφορούσαν στην κατανομή των λέξεων-κλειδιών στο χάρτη. Θα πρέπει να τονιστεί ότι τα κριτήρια αυτά αντικατοπτρίζουν γεωμετρικές ιδιότητες του χάρτη και συνεπώς οι συγκρίσεις των τιμών για χάρτες διαφορετικού μεγέθους είναι επισφαλείς. Τα πειράματα που παρουσιάζονται εδώ αφορούν στα κείμενα της Βουλής (Α’ σώμα κειμένων). Οι κατηγορίες λέξεων είναι οι εξής: Κ1 = εσωτερικά, Κ2 = εξωτερικά, Κ3 = οικονομία και Κ4= παιδεία.

Όρια ABC	Μέγεθος SOM	Μετρήσεις κριτηρίου 1	Λέξεις γλωσσολόγου	Λέξεις τίτλων	Λέξεις 70%-85%	Λέξεις 85%-90%
70% 85%	[10 15] κανονικ.	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,87 0,60 ***0,71** 0,80	0,77 0,65 0,87 ***0,37**	0,79 0,39 0,45 ***0,28**	0,72 0,56 0,81 0,82
70% 85%	[10 15]	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,88 ***0,56** 0,81 ***0,66**	***0,76** ***0,50** 0,83 0,49	***0,52** ***0,38** ****0,44** 0,42	***0,69** ***0,39** 0,68 0,72
85% 90%	[10 15] κανονικ.	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	***0,81** 0,67 0,73 0,92	0,95 0,70 0,61 0,80	1,00 0,57 0,74 0,57	0,94 0,53 ***0,31** 0,72
85% 90%	[10 15]	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,92 0,66 0,78 0,76	1,04 0,71 ***0,57** 0,50	0,81 0,42 0,67 0,45	0,97 0,53 0,41 0,66
70% 85%	[25 17] κανονικ.	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,93 0,58 0,78 0,79	0,86 0,70 0,88 0,42	0,85 0,49 0,55 0,40	0,95 0,50 0,76 0,78
70% 85%	[25 17]	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,84 0,59 0,78 0,77	0,81 0,57 0,87 0,53	0,54 0,42 ****0,44** 0,40	0,79 0,46 0,51 0,85
85% 90%	[26 16] κανονικ.	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,85 0,77 0,79 0,89	0,84 0,84 0,88 0,63	0,98 0,56 0,56 0,60	0,99 0,69 0,73 0,77
85% 90%	[26 16]	D{K1} / D{~K1} D{K2} / D{~K2} D{K3} / D{~K3} D{K4} / D{~K4}	0,94 0,77 0,87 0,73	1,03 0,81 0,70 0,52	0,83 0,50 0,72 0,38	0,99 0,61 0,65 ***0,65**

Πίνακας 4.5. Ο λόγος των αποστάσεων των λέξεων-κλειδιών ανά κατηγορία ως προς τις λέξεις των άλλων κατηγοριών για διάφορες παραμέτρους του χάρτη SOM (κριτήριο 1). Με κόκκινο χρώμα σημειώνονται οι ιδιαίτερα χαμηλές επιδόσεις των χαρτών (από 0,85 και πάνω) και με αστεράκι οι καλύτερες ανά σύνολο λέξεων και κατηγορία.

Με D{Kx} συμβολίζεται η μέση τιμή της απόστασης των λέξεων που ανήκουν στην κατηγορία Kx για το κριτήριο 1 και με D{~Kx} αυτών που δεν ανήκουν στην κατηγορία Kx (όπου το x κυμαίνεται από 1 έως 4 για κάθε κατηγορία λέξεων). Ο λόγος D{Kx} προς D{~Kx} είναι επιθυμητό να είναι ελάχιστος και να μην υπερβαίνει τη μονάδα, γιατί τότε οι λέξεις διαφορετικής κατηγορίας βρίσκονται πιο κοντά από ό,τι οι λέξεις της ίδιας κατηγορίας. Με κόκκινο χρώμα σημειώνονται οι ιδιαίτερα χαμηλές επιδόσεις των χαρτών (από 0,85 και πάνω). Το σύστημα που επιλέχθηκε από τα πειράματα κατηγοριοποίησης (μέγεθος χάρτη 10x15, παράλειψη κανονικοποίησης, εύρος ανάλυσης ABC 70%-85%, μέγεθος MLP 10 νευρώνες και k=42 για το k-means) στην ενότητα 4.4.1.1 δηλώνεται στον πίνακα 4.5 με υπογράμμιση. Με αστερίσκο σημειώνονται για κάθε ομάδα οι καλύτερες τιμές του κριτηρίου από το σύνολο των χαρτών που εξετάσθηκαν (δηλαδή συνδυασμός μεγέθους χάρτη και κανονικοποίησης ή μη). Σε περίπτωση ισοψηφίας οι καλύτερες τιμές σημειώνονται με δύο αστερίσκους. Όπως φαίνεται από τον πίνακα 4.5, το επιλεγμένο σύστημα επιτυγχάνει και την καλύτερη απόδοση στο κριτήριο 1, καθώς μόνο σε 7 από τις 16 περιπτώσεις είναι προτιμητέο κάποιο από τα υπόλοιπα 7 συστήματα.

Όρια ABC	Μέγεθος SOM	Μετρήσεις κριτηρίου 2	Λέξεις γλωσσολόγου	Λέξεις τίτλων	Λέξεις 70%-85%	Λέξεις 85%-90%
70% 85%	[10 15] κανονικ.	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	1,52 0,69 1,90 0,71	1,90 0,40 0,63 1,17	1,12 0,27 0,47 0,14	1,31 0,28 0,95 0,73
70% 85%	[10 15]	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	2,23 0,73 1,03 1,19	2,62 0,57 1,00 2,25	1,49 0,26 0,72 0,17	1,60 0,25 4,15 0,51
85% 90%	[10 15] κανονικ.	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	1,39 0,59 1,47 1,19	1,62 0,53 0,88 0,91	1,42 0,67 1,60 0,77	1,22 0,42 1,24 0,73
85% 90%	[10 15]	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	2,39 0,47 1,85 1,55	0,80 0,34 0,35 1,33	0,55 0,30 0,61 0,29	1,66 0,39 0,94 1,08
70% 85%	[25 17] κανονικ.	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	1,33 0,88 1,14 0,78	2,10 0,66 0,68 1,43	1,16 0,24 0,85 0,36	1,55 0,39 3,47 0,87
70% 85%	[25 17]	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	1,61 0,82 1,40 0,49	1,38 0,56 0,81 3,06	0,58 0,33 0,48 0,08	1,60 0,48 1,40 0,57
85% 90%	[26 16] κανονικ.	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	1,45 0,81 1,24 1,22	2,44 0,59 0,52 1,34	1,22 0,45 0,61 0,45	1,48 0,47 2,15 0,73
85% 90%	[26 16]	N{K1} / N{~K1} N{K2} / N{~K2} N{K3} / N{~K3} N{K4} / N{~K4}	2,12 0,59 1,31 1,33	0,74 0,37 0,44 0,99	0,65 0,15 0,43 0,32	1,39 0,63 2,55 1,06

Πίνακας 4.6. Ο λόγος των αποστάσεων από τον κοντινότερο γείτονα των λέξεων-κλειδιών ανά κατηγορία ως προς τις λέξεις των άλλων κατηγοριών του κοντινότερου γείτονα για διάφορες παραμέτρους του χάρτη SOM (κριτήριο 2).

Με N{Kx} συμβολίζεται η τιμή της απόστασης του κοντινότερου γείτονα που περιέχει λέξεις που ανήκουν στην κατηγορία Kx και με N{~Kx} αυτού που περιέχει λέξεις που δεν ανήκουν στην κατηγορία Kx για το κριτήριο 2. Ο λόγος N{Kx} προς N{~Kx} είναι επιθυμητό να είναι όσο το δυνατό πιο μικρός. Όπως παρατηρείται όμως, με το κριτήριο 2 εμφανίζει σε πολλές περιπτώσεις ανεπιθύμητα υψηλές τιμές, συχνότατα δε μεγαλύτερες της μονάδας. Το φαινόμενο αυτό οφείλεται στο γεγονός ότι οι λέξεις-κλειδιά μπορεί να είναι διάσπαρτα κατανεμημένες στο χάρτη SOM, δηλαδή να τοποθετούνται σε περισσότερες ομάδες, οι οποίες να είναι και μεταξύ τους απομακρυσμένες. Το σύστημα που επιλέχθηκε από τα πειράματα κατηγοριοποίησης δηλώνεται στον πίνακα 4.6 με υπογράμμιση. Συνολικά, οι τιμές του κριτηρίου ήταν αρκετά υψηλότερες από το προσδοκώμενο για το πλήθος των περιπτώσεων. Καλύτερη εικόνα για τα αποτελέσματα του κριτηρίου 2 θα φανούν και στην συνέχεια όπου και θα παρουσιαστεί η απεικόνιση της κατανομής των λέξεων-κλειδιών στο χάρτη.

Όρια ABC	Μέγεθος SOM	Μετρήσεις κριτηρίου 3	Λέξεις γλωσσολόγου	Λέξεις τίτλων	Λέξεις 70%-85%	Λέξεις 85%-90%
70% 85%	[10 15] κανονικ.	G NG Q=G/(NG+G)	53 67 0,44	37 22 0,63	43 5 0,90	60 15 0,80
70% 85%	[10 15]	G NG Q=G/(NG+G)	58 73 0,44	42 24 0,64	44 14 0,76	53 28 0,65
85% 90%	[10 15] κανονικ.	G NG Q=G/(NG+G)	61 74 0,45	45 31 0,59	40 11 0,78	49 21 0,70
85% 90%	[10 15]	G NG Q=G/(NG+G)	52 80 0,39	48 20 0,71	36 6 0,86	44 19 0,70
70% 85%	[25 17] κανονικ.	G NG Q=G/(NG+G)	108 44 **0,71***	55 18 0,75	64 8 0,89	74 18 0,80
70% 85%	[25 17]	G NG Q=G/(NG+G)	91 61 0,60	58 10 **0,85***	62 4 0,94	81 12 0,87
85% 90%	[26 16] κανονικ.	G NG Q=G/(NG+G)	97 59 0,62	58 24 0,71	47 4 0,92	70 10 **0,88***
85% 90%	[26 16]	G NG Q=G/(NG+G)	89 77 0,54	63 16 0,80	51 0 **1,00***	67 12 0,85

Πίνακας 4.7. Το ποσοστό των ομάδων του χάρτη με λέξεις-κλειδιά μίας μόνο κατηγορίας για διάφορες παραμέτρους του χάρτη SOM (κριτήριο 3)

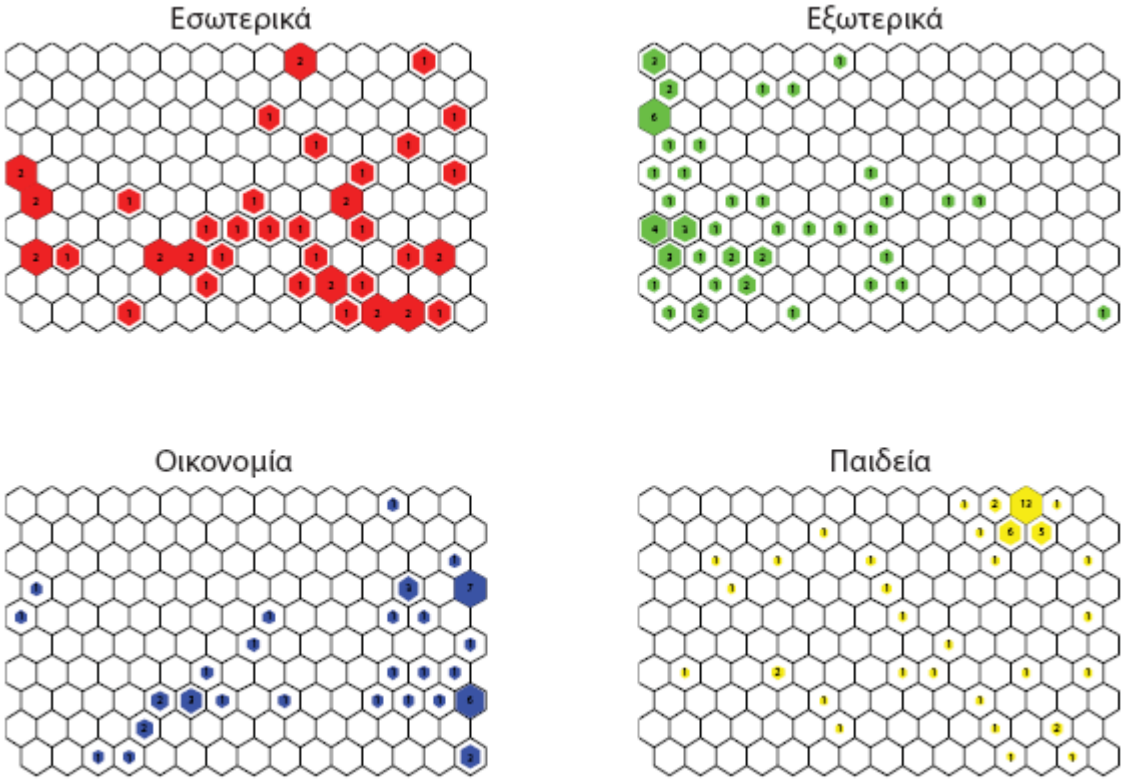
Στον πίνακα 4.7, το G συμβολίζει το πλήθος των ομάδων με λέξεις-κλειδιά μίας μόνο κατηγορίας, ενώ το NG το πλήθος των ομάδων με λέξεις-κλειδιά από περισσότερες κατηγορίες. Το Q, ο λόγος του πλήθους των ομάδων με λέξεις-κλειδιά μίας κατηγορίας προς το συνολικό πλήθος των ομάδων, θα πρέπει να είναι όσο το δυνατό μεγαλύτερο, χωρίς όμως να υπερβαίνει τη μονάδα. Το κριτήριο 3 φαίνεται να δίνει συστηματικά μεγαλύτερες τιμές στα συστήματα SOM με περισσότερους νευρώνες (βλ. τα καλύτερα σκορ που σημειώνονται ανά κατηγορία με έντονα γράμματα), αφού αυτά είναι πιο πιθανό να περιέχουν «καθαρές» ομάδες, αλλά σε μικρότερους χάρτες (μεγέθους [10 15]) δε φαίνεται (με βάση τα σημειωμένα με αστερίσκο αποτελέσματα) να προκρίνει σαφώς μία λύση, ενώ είναι χαρακτηριστικό το ότι δεν προκρίνει σε καμία περίπτωση το επιλεγμένο από τα πειράματα κατηγοριοποίησης σύστημα.

Όρια ABC	Μέγεθος SOM	Μετρήσεις κριτηρίου 4	Λέξεις γλωσσολόγου	Λέξεις τίτλων	Λέξεις 70%-85%	Λέξεις 85%-90%
70% 85%	[10 15] κανονικ.	MLP *h=2* MLP *h=3*	0,85 0,86	0,86 0,87	0,90 0,92	0,87 0,89
70% 85%	[10 15]	MLP *h=2* MLP *h=3*	0,86 **0,87*****	**0,88**** 0,88	**0,91**** **0,94***	**0,90*** **0,90***
85% 90%	[10 15] κανονικ.	MLP *h=2* MLP *h=3*	0,84 0,85	0,87 **0,89****	0,86 0,89	0,87 0,87
85% 90%	[10 15]	MLP *h=2* MLP *h=3*	0,85 0,86	0,87 0,87	0,87 0,88	0,88 0,88
70% 85%	[25 17] κανονικ.	MLP *h=2* MLP *h=3*	0,85 **0,87*****	0,86 0,87	0,90 0,93	0,88 0,88
70% 85%	[25 17]	MLP *h=2* MLP *h=3*	**0,87***** **0,87*****	**0,88**** **0,89****	**0,91**** 0,93	0,89 0,89
85% 90%	[26 16] κανονικ.	MLP *h=2* MLP *h=3*	0,85 0,85	0,85 0,86	0,88 0,90	0,87 0,88
85% 90%	[26 16]	MLP *h=2* MLP *h=3*	0,84 0,85	0,86 0,86	0,90 0,90	0,86 0,87

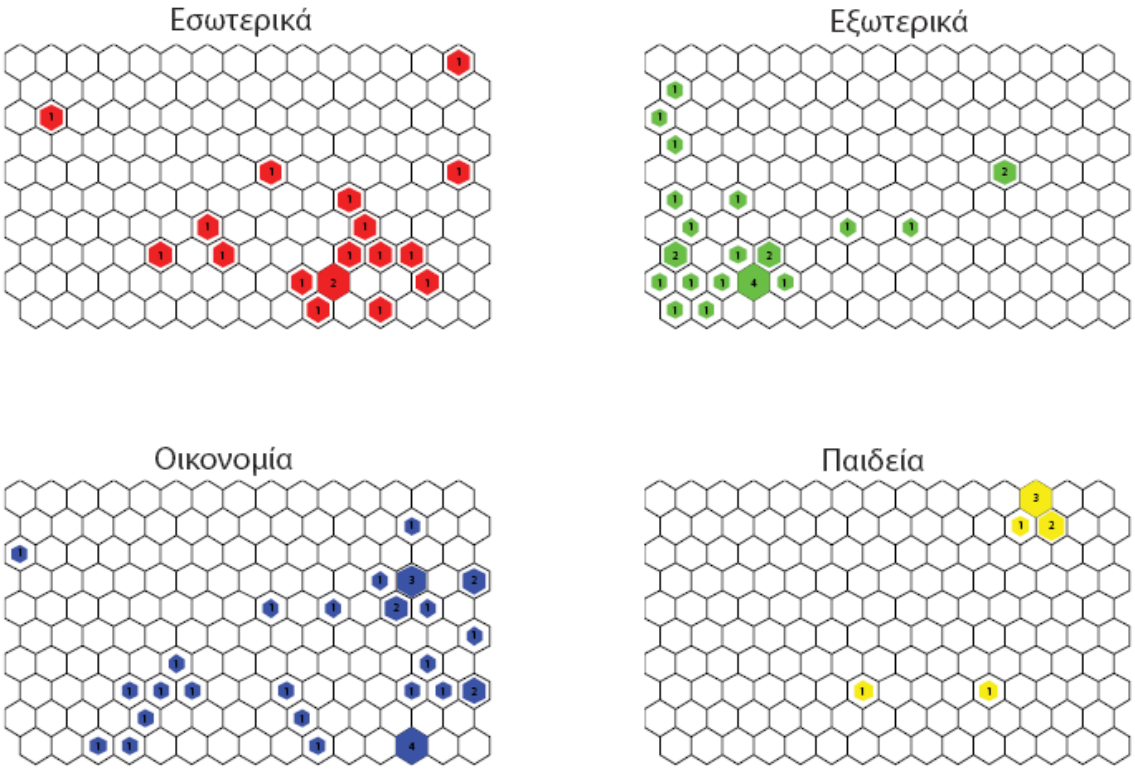
Πίνακας 4.8. Η ακρίβεια κατηγοριοποίησης του MLP που δείχνει το βαθμό στον οποίο οι λέξεις-κλειδιά προβάλλονται σε ομάδες γραμμικά διαχωρίσιμες στο χάρτη SOM για διάφορες παραμέτρους του χάρτη (κριτήριο 4).

Στον πίνακα 4.8, απεικονίζεται η ακρίβεια κατηγοριοποίησης του MLP με 2 και 3 νευρώνες στο κρυφό επίπεδο για τις λέξεις-κλειδιά, η οποία και αντικατοπτρίζει το βαθμό που οι λέξεις-κλειδιά προβάλλονται σε ομάδες γραμμικά διαχωρίσιμες ανά θεματική κατηγορία. Πρέπει να αναφερθεί ότι η ακρίβεια κατηγοριοποίησης του MLP για 2 και 3 νευρώνες συνιστά διαφορετικό κριτήριο. Με αστερίσκο σημειώνεται η καλύτερη επίδοση ανά χάρτη, ενώ οι περιπτώσεις ισοβαθμιών επισημαίνονται με δύο ή τρεις αστερίσκους ανάλογα με το πλήθος των ισόβαθμων στοιχείων. Όπως δείχνουν οι μετρήσεις, το προτιμητέο σύστημα (της ενότητας 4.4.1.1) και εδώ συμπίπτει με τα αποτελέσματα κατηγοριοποίησης αλλά και με αυτά του κριτηρίου 1, καθώς εμφανίζει την καλύτερη απόδοση σε 6 από τα 8 πειράματα (παρά το γεγονός ότι για τις τιμές με περισσότερους του ενός αστερίσκους και άλλα συστήματα απέδωσαν εξίσου καλά). Το σημαντικό, όμως, είναι ότι όλες οι τιμές παρουσιάζουν μικρές διαφοροποιήσεις, κάτι που καθιστά λιγότερο ασφαλή τη λήψη κατηγορηματικής απόφασης. Επίσης, οι τιμές είναι αρκετά υψηλές, γεγονός που υποδηλώνει το ότι οι λέξεις-κλειδιά των θεματικών κατηγοριών (εσωτερικά, εξωτερικά, οικονομία, παιδεία) προβάλλονται στους χάρτες με τρόπο διαχωρίσιμο.

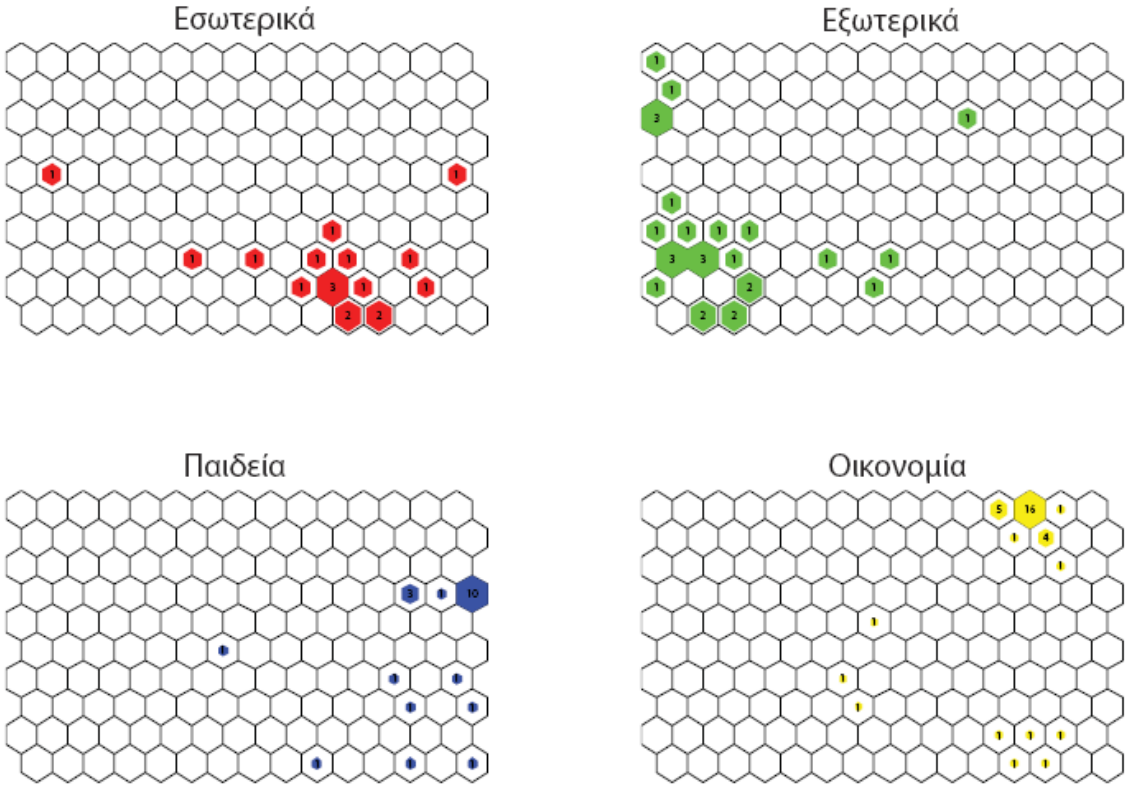
Για την καλύτερη κατανόηση των μετρήσεων των κριτηρίων παρουσιάζεται για το επιλεγμένο σύστημα η προβολή των λέξεων-κλειδιών στο χάρτη (σχήματα 4.23 έως 4.26). Στα σχήματα αυτά προβάλλονται οι ομάδες στις οποίες αντιστοιχίζονται οι λέξεις-κλειδιά για τα διάφορα σύνολα (λέξεις γλωσσολόγου, λέξεις τίτλων, λέξεις εύρους 70%-85%, λέξεις εύρους 85%-90%). Στο εσωτερικό των κόμβων αναγράφεται το πλήθος λέξεων που αποδίδεται σε αυτούς.



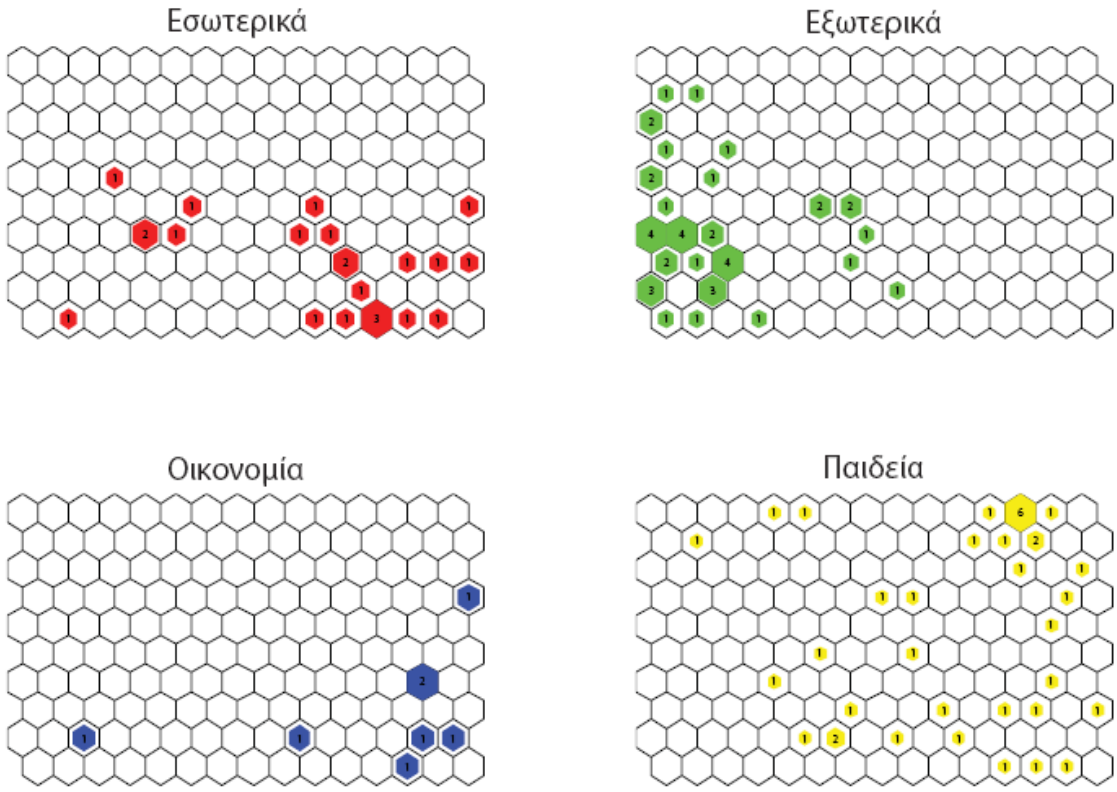
Σχήμα 4.23. Προβολή των λέξεων-κλειδιών που προήλθαν από γλωσσολόγο (1ο σύνολο) στο χάρτη SOM



Σχήμα 4.24. Προβολή των λέξεων-κλειδιών που προήλθαν από τους τίτλους (2ο σύνολο) στο χάρτη SOM



Σχήμα 4.25. Προβολή των λέξεων-κλειδιών στο χάρτη SOM που προήλθαν από το εύρος 70%-85% (3ο σύνολο) ανά κατηγορία κειμένων



Σχήμα 4.26. Προβολή των λέξεων-κλειδιών στο χάρτη SOM που προήλθαν από το εύρος 85%-90% (4ο σύνολο) ανά κατηγορία κειμένων

Από τα παραπάνω σχήματα (4.23 έως 4.26) συνάγεται ότι η κατηγορία των εσωτερικών για το σύνολο των δεδομένων είναι ιδιαίτερα διεσπαρμένη, ενώ οι πιο συγκεντρωμένες κατηγορίες είναι αυτές της οικονομίας και της παιδείας.

4.4.3.2 Πειράματα αξιολόγησης με τη χρήση θησαυρού

Σε συνέχεια των πειραμάτων για την ανεξάρτητη ενδιάμεση αξιολόγηση των χαρτών, χρησιμοποιήθηκε ο θησαυρός Eurovoc, με τη βοήθεια του οποίου μετρήθηκαν τα κριτήρια που παρουσιάσθηκαν στην ενότητα 4.3.2. Για διάφορες τιμές του κατωφλίου στον παραγόμενο από το θησαυρό γράφο, παρουσιάζονται στον πίνακα 4.9 οι τιμές των σχετικών κριτηρίων (*eurovoc_w*: τύπος (4-14), *eurovoc_g*: τύπος (4-15), *eurovoc_w_rand1*, *eurovoc_w_rand2* και *eurovoc_g_rand*). Οι τιμές αυτές είναι επιθυμητό να είναι μεγαλύτερες, αφού στην περίπτωση των *eurovoc_w* και *eurovoc_g* εκφράζουν το βαθμό ομοιότητας των συνδέσεων του χάρτη με το θησαυρό, ενώ για τα κριτήρια *eurovoc_w_rand1*, *eurovoc_w_rand2* και *eurovoc_g_rand* μετράνε το βαθμό τα αποτελέσματα του χάρτη SOM είναι καλύτερα από τα τυχαία παραγόμενα. Προφανώς, για την τελευταία κατηγορία μετρήσεων είναι καλό οι τιμές να είναι αρκετά μεγαλύτερες της μονάδας.

Όρια ABC	Μέγεθος SOM	Τιμές κριτήρι-ων (*eurovoc)*	Κατώφλι *c=1*	Κατώφλι *c=2*	Κατώφλι *c=3*	Κατώφλι *c=4*	Κατώφλι *c=5*
70% 85%	[10 15] κανονικ.	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	0,03 **0,55ˇˇ** 2,97 454,61 **0,55ˇˇ**	0,07 0,64 2,39 365,56 0,64	0,10 0,69 1,95 298,40 0,69	0,18 0,79 1,68 257,13 0,79	0,30 0,87 1,43 218,25 0,87
70% 85%	[10 15]	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	0,03 **0,55ˇˇ** 3,03 463,55 **0,55ˇˇ**	0,07 **0,69ˇˇ** 2,50 382,81 **0,69ˇˇ**	0,10 **0,78ˇˇ** 1,91 293,22 **0,78ˇˇ**	0,18 **0,83ˇˇ** 1,67 255,97 **0,83ˇˇ**	0,31 **0,89ˇˇ** 1,44 221,06 **0,89ˇˇ**
85% 90%	[10 15] κανονικ.	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	0,03 0,41 3,25 497,83 0,41	**0,08ˇˇ** 0,56 2,76 422,31 0,56	**0,12ˇˇ** 0,65 2,26 346,69 0,65	**0,21ˇˇ** 0,74 **1,96ˇˇ** **299,79ˇ** 0,74	**0,33ˇˇ** 0,83 **1,56ˇ** **239,26ˇ** 0,83
85% 90%	[10 15]	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	**0,04ˇ** 0,54 **3,44ˇ** **526,15ˇ** 0,54	**0,08ˇˇ** **0,69ˇˇ** **2,81ˇ** **430,66ˇ** **0,69ˇˇ**	**0,12ˇˇ** 0,73 **2,27ˇ** **348,13ˇ** 0,73	**0,21ˇˇ** 0,80 **1,96ˇˇ** 299,50 0,80	**0,33ˇˇ** 0,87 1,55 237,10 0,87
70% 85%	[25 17] κανονικ.	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	0,04 0,20 3,80 1683,49 0,26	0,09 **0,31^** 3,12 1384,19 **0,32^**	0,13 **0,37^** 2,38 1054,83 **0,37^**	0,21 **0,45^** 1,96 867,81 **0,45^**	0,33 **0,53^** 1,54 684,37 **0,53^**
70% 85%	[25 17]	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	0,04 0,19 4,08 1840,85 0,26	0,08 0,29 2,94 1326,81 0,29	0,12 0,36 2,30 1036,31 0,36	0,20 0,46 1,93 872,16 0,46	0,33 0,56 1,53 692,72 0,56
85% 90%	[26 16] κανονικ.	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	**0,05^^** 0,18 4,58 2033,14 0,23	**0,10^^** 0,29 3,52 **1563,05^** 0,29	**0,16^** 0,34 **2,96^** **1314,58^** 0,34	**0,25^^** 0,40 **2,37^** **1051,25^** 0,40	**0,38^^** 0,47 **1,79^** **795,20^** 0,47
85% 90%	[26 16]	*_w* *_g* *_w_rand1* *_w_rand2* *_g_rand*	**0,05^^** **0,22^** **4,77^** **2104,42^** **0,28^**	**0,10^^** 0,29 **3,53^** 1556,72 0,30	0,15 0,34 2,81 1240,81 0,34	**0,25^^** 0,41 2,32 1025,91 0,41	**0,38^^** 0,51 1,78 784,94 0,51

Πίνακας 4.9. Τιμές των κριτηρίων του Eurovoc για διάφορες παραμέτρους του χάρτη SOM

Από τον πίνακα 4.9 προκύπτει ότι τα κριτήρια που χρησιμοποιούνται για τη σύγκριση με το θησαυρό δείχνουν σαφή προτίμηση στα συστήματα με περισσότερες ομάδες SOM. Για το λόγο αυτό σημειώνονται με διαφορετικά σύμβολα (ˇ,^) οι καλύτερες τιμές των κριτηρίων για τους χάρτες μεγέθους 10x15 από τους υπόλοιπους. Σε περιπτώσεις που υπάρχουν περισσότεροι από ένας νικητές, το επιλεγμένο σύμβολο εμφανίζεται περισσότερες φορές.

Επιπλέον, επειδή τα διάφορα κριτήρια παρουσιάζουν αρκετά διαφορετικά εύρη τιμών, για να δοθεί ένα μέτρο αξιολόγησης της σημασίας της κλίμακάς τους, παρουσιάζεται η διασπορά τους ανά μέγεθος χάρτη στον πίνακα 4.10 για διάφορες τιμές κατωφλίου.

Μέγεθος SOM	**Κριτήρια**	**Κατώφλι *c=1***	**Κατώφλι *c=2***	**Κατώφλι *c=3***	**Κατώφλι *c=4***	**Κατώφλι *c=5***
[10 15]	*_w*	0,00003	0,00003	0,00013	0,00030	0,00023
	_g	0,00469	0,00377	0,00309	0,00140	0,00063
	_w_rand1	0,04629	0,04097	0,03769	0,02709	0,00483
	_w_rand2	1080,15	972,98	892,34	619,30	116,49
	_g_rand	0,00469	0,00377	0,00309	0,00140	0,00063
[25 17] [26 16]	*_w*	0,00003	0,00009	0,00033	0,00069	0,00083
	_g	0,00029	0,00010	0,00022	0,00087	0,00142
	_w_rand1	0,19916	0,08709	0,10383	0,05390	0,02087
	_w_rand2	36310,40	14479,83	18924,84	9584,93	3464,94
	_g_rand	0,00043	0,00020	0,00022	0,00087	0,00142

Πίνακας 4.10. Τιμές των διασπορών των κριτηρίων για διάφορα μεγέθη του χάρτη SOM

Το μόνο κριτήριο που φαίνεται να προτιμά το καλύτερο στα πειράματα κατηγοριοποίησης σύστημα είναι το κριτήριο *eurovoc_g* που σχετίζεται με τις ομάδες. Στα υπόλοιπα κριτήρια οι διαφοροποιήσεις μεταξύ χαρτών αντίστοιχων μεγεθών δεν μπορούν να δείξουν σαφή προτίμηση (κυρίως για τα μεγέθη 10x15) σε κάποιες παραμέτρους. Τα κριτήρια *eurovoc_rand_w1* και *eurovoc_rand_w2*, που σχετίζονται με το πόσες φορές είναι καλύτερες κάποιες μετρήσεις σε σχέση με τους τυχαίους χάρτες, έχουν τιμές μεγαλύτερες της μονάδας, όχι όμως και το κριτήριο *eurovoc_rand_g*. Αυτό οφείλεται στο γεγονός ότι οι τυχαία παραγόμενοι χάρτες έχουν ισοκατανεμημένες λέξεις στις ομάδες, κάτι που δεν ισχύει για τους χάρτες του συστήματος SOM. Ως εκ τούτου είναι πιο πιθανό να υπάρχει ταυτόχρονα τουλάχιστον μία σύνδεση ανάμεσα στις λέξεις των ομάδων και στο Eurovoc.

Οι ομάδες του SOM έχουν το μειονέκτημα ότι περιλαμβάνουν λιγότερες λέξεις και πολλές από αυτές δεν απαντώνται στο Eurovoc, με αποτέλεσμα οι χάρτες να εμφανίζονται λιγότερο συνδεδεμένοι. Οι υψηλές τιμές που παρατηρούνται στα κριτήρια που συσχετίζουν λέξεις και όχι ομάδες, δηλαδή τα *eurovoc_rand_w1* και *eurovoc_rand_w2*, αποτελούν ένα στοιχείο τεκμηρίωσης της αποτελεσματικότητας της μεθόδου που ακολουθήθηκε αφού δείχνουν ότι οι ομάδες λέξεων που δημιουργούνται με την προτεινόμενη διαδικασία περιλαμβάνουν σαφώς περισσότερες σχετιζόμενες λέξεις από ό,τι οι τυχαία παραγόμενοι χάρτες, βάσει του θησαυρού (Eurovoc).

Δυστυχώς όμως η μέθοδος αξιολόγησης αυτή δεν μπορεί να επιλέξει το προτιμητέο σύστημα. Άλλη μία παρατήρηση είναι ότι, καθώς αυξάνει η τιμή του κατωφλίου απόστασης στο γράφο (*c)*, βάσει του οποίου ορίζονται οι σχετιζόμενες λέξεις στο Eurovoc, οι μετρήσεις στους χάρτες SOM γίνονται πιο συμβατές με το Eurovoc, αφού εμφανίζουν μεγαλύτερες τιμές, γεγονός που οφείλεται στο ότι η έννοια της γειτονίας στο θησαυρό επεκτείνεται και περιλαμβάνει περισσότερες λέξεις.

Η ανεξάρτητη ενδιάμεση αξιολόγηση των χαρτών δεν μπορεί να οδηγήσει με ακρίβεια στην επιλογή των παραμέτρων του συστήματος. Αυτό οφείλεται στο γεγονός ότι είναι πολύ δύσκολο να βρεθούν σύνολα δεδομένων, όπως λέξεις-κλειδιά και χάρτες, πλήρως εναρμονισμένα με τα κείμενα προς κατηγοριοποίηση. Για παράδειγμα, ο θησαυρός Eurovoc έχει φτιαχτεί με διαφορετικό σκοπό και περιέχει ιεραρχίες και σχέσεις διαφορετικές από ό,τι απαιτεί η θεματική κατάταξη των κειμένων της Βουλής. Αυτό γίνεται αντιληπτό και από το ότι πολλές λέξεις του χάρτη SOM δεν απαντώνται στο Eurovoc. Ακόμη και οι λέξεις-κλειδιά προκύπτουν από δειγματοληψία, και ενδέχεται να απεικονίζουν μόνο ένα μέρος των χαρακτηριστικών των χαρτών. Επιπλέον η προσπάθεια αξιολόγησης γεωμετρικών ιδιοτήτων με αριθμητικά κριτήρια δεν μπορεί να εκφράσει απόλυτα τις ιδιότητες αυτές, ενώ επηρεάζεται σημαντικά από τα μεγέθη των χαρτών που εφαρμόζεται.

Παρ' όλα αυτά, οι αυτόνομες αξιολογήσεις αποτελούν ένα βήμα τεκμηρίωσης της τεχνικής που εφαρμόζεται για την απεικόνιση των κειμένων στο χώρο προτύπων, αφού εμφανίζουν την αυξημένη συσχέτιση που υπάρχει ανάμεσα στους παραγόμενους με το SOM χάρτες λέξεων και στο θησαυρό Eurovoc που προέρχεται από ανεξάρτητη με τα δεδομένα πηγή. Οι αυτόνομες αξιολογήσεις μπορούν να αποκαλύψουν εν μέρει τις δυνατότητες του συστήματος.

ΚΕΦΑΛΑΙΟ 5ο

5.1 Σύστημα οργάνωσης κειμένων βασισμένο σε ACO

Ο αλγόριθμος ACO (Ant Colony Optimization) είναι ένας αλγόριθμος βελτιστοποίησης βασισμένος στον τρόπο λειτουργίας των αποικιών εντόμων και ειδικότερα των μυρμηγκιών (Kennedy J. & Eberhart R. C. 2001) αναφορικά προς τη μετακίνησή τους και τη μεταφορά της τροφής τους για αποθήκευση. Πρόκειται για μία πιθανοτική μέθοδο που χρησιμοποιείται ευρέως για την επίλυση δύσκολων υπολογιστικών προβλημάτων, των οποίων η αναλυτική λύση χαρακτηρίζεται από εκθετική πολυπλοκότητα (Hromkovic J., 2002, Martens D., De Backer M., Haesen R., Baesens, B., Mues C. & Vanthienen, J, 2006 και Lessing L., Dumitrescu I. & Stützle T, 2004).

Πιο συγκεκριμένα, στη φύση, τα μυρμήγκια αρχικά περιπλανώνται τυχαία στο χώρο μέχρι να βρούνε τροφή, οπότε και επιστρέφουν στην αποικία. Κατά τη διαδρομή της επιστροφής αφήνουν ίχνη φερομόνης, προκειμένου κάποιο άλλο μυρμήγκι, εντοπίζοντας αυτό το ίχνος, να σταματήσει να περιπλανάται τυχαία και να το ακολουθήσει. Το δεύτερο μυρμήγκι με τη σειρά του θα ελευθερώσει νέα ποσότητα φερομόνης, ενισχύοντας το ίχνος και ενθαρρύνοντας και άλλα μέλη να ακολουθήσουν το ίδιο δρομολόγιο. Όσο μεγαλύτερη είναι η διαδρομή, τόσο περισσότερο αργεί το μυρμήγκι να τη διανύσει και το ίχνος σταδιακά χάνει την έντασή του. Αντίθετα, μία πιο σύντομη διαδρομή προς την ίδια πηγή τροφής δέχεται φερομόνη συχνότερα και το ίχνος ανανεώνεται συχνά παραμένοντας ισχυρό. Συνεπώς, με τον εντοπισμό συντομότερων μονοπατιών επιτυγχάνεται η ταχύτερη μεταφορά τροφής στην αποικία.

Η εξάτμιση της φερομόνης, που έχει ως αποτέλεσμα να μειώνεται η ελκυστικότητα των μη πολυσύχναστων διαδρομών, αποτελεί ένα μηχανισμό προστασίας της αναζήτησης από τα τοπικά ελάχιστα, καθώς καθιστά μεν δυνατή τη συνεχή αναζήτηση στο χώρο, χωρίς ωστόσο να βαραίνουν μονοσήμαντα οι λύσεις που έχουν βρεθεί έως εκείνη τη στιγμή. Επιπλέον, ο τρόπος που κινούνται τα μέλη της αποικίας, παρά το γεγονός ότι επηρεάζεται από την ποσότητα φερομόνης που υπάρχει κατανεμημένη στο χώρο, είναι στοχαστικός, με αποτέλεσμα να δίνεται η δυνατότητα εξερεύνησης νέων διαδρομών.

Με βάση τον τρόπο λειτουργίας των αποικιών κατασκευάστηκαν προγράμματα για ηλεκτρονικούς υπολογιστές, τα οποία προσομοιώνουν τη διαδικασία αναζήτησης τροφής με σκοπό την εξεύρεση λύσεων σε προβλήματα βελτιστοποίησης. Απαραίτητη προϋπόθεση στα προβλήματα αυτά είναι η ύπαρξη αντικειμενικής συνάρτησης, η οποία και έχει ως σκοπό την αξιολόγηση των εκάστοτε λύσεων με βάση τα επιθυμητά αποτελέσματα.

Οι αλγόριθμοι της οικογένειας ACO έχουν χρησιμοποιηθεί αρκετά για την εξεύρεση λύσεων στο πρόβλημα του πλανόδιου πωλητή (TSP – Traveling Salesman Problem) (Hromkovic J., 2002 και Dorigo M. & Gambardella M., 1997) ο οποίος πρέπει να διέλθει από ένα σύνολο σημείων με το ελάχιστο δυνατό κόστος. Το κόστος ή, ακριβέστερα, η συνολική απόσταση της διαδρομής αποτελεί την αντικειμενική συνάρτηση αξιολόγησης των λύσεων. Η φύση του προβλήματος αυτού έχει σαφείς ομοιότητες με τη διαδικασία συλλογής τροφής που περιγράφηκε. Οι αλγόριθμοι ACO παρουσιάζουν ομοιότητες με τους εξελικτικούς αλγόριθμους (Evolutionary Computation Algorithms), αλλά και τους αλγόριθμους αναζήτησης που εμπνέονται από τη μελέτη σμηνών (Particle Swarm Optimization,) αλλά λόγω της φύσης τους ενδείκνυνται περισσότερο για προβλήματα αναζήτησης διαδρομών ιδιαίτερα σε δυναμικά περιβάλλοντα.

5.1.1 Περιγραφή συστήματος

Η μοντελοποίηση που ακολουθείται για την επίλυση προβλημάτων TSP αποτέλεσε το κίνητρο για την εφαρμογή της τεχνικής αυτής ως εναλλακτικής για τη δημιουργία ενός χάρτη λέξεων που να οργανώνει τις λέξεις ανάλογα με τη νοηματική τους συνάφεια. Ο αλγόριθμος ACO χρησιμοποιήθηκε αντί για μοντέλο SOM, όπως αυτό παρουσιάστηκε στην ενότητα 4.1.1.2, ενώ στη συνέχεια το σύστημα παραμένει ως έχει με τη χρήση του κατηγοριοποιητή MLP.

Βασική παραδοχή του προτεινόμενου συστήματος είναι ότι οι λέξεις συνδέονται με κάποια «νοηματική» σχέση, η οποία και είναι επιθυμητό να αντικατοπτρίζεται από κάποιο μαθηματικά εκφρασμένο μέτρο βασισμένο στις κοινές εμφανίσεις λέξεων στα πλαίσια μίας περιόδου.

Έτσι ορίζεται αρχικά μία τιμή τ_{ij} για κάθε ζεύγος λέξεων *(i,j)*, η οποία ισούται με το λόγο του πλήθους των κοινών εμφανίσεων των λέξεων σε επίπεδο περιόδου προς τη μικρότερη από τις δύο ανεξάρτητη συχνότητα εμφάνισης των λέξεων αυτών.

$$\tau_{ij} = \frac{\#\underset{s}{Count}(i, j : i \in s \wedge j \in s)}{min[\#\underset{s}{Count}(i : i \in s), \#\underset{s}{Count}(j : j \in s)]} \quad (5\text{-}1)$$

όπου *s* η εκάστοτε πρόταση και *#Count()* η συνάρτηση που μετράει το πλήθος των προτάσεων που ικανοποιούν τον περιορισμό της παρένθεσης. Σημειώνεται ότι κάθε τ_{ij} έχει τιμή μικρότερη ή ίση της μονάδας. Όταν η τιμή του τ_{ij} ισούται με τη μονάδα, οι δύο λέξεις i και j εμφανίζονται μαζί σε κάθε πρόταση, ενώ όταν η τιμή του τ_{ij} είναι μηδέν, οι λέξεις αυτές δεν εμφανίζονται πότε μαζί στην ίδια πρόταση.

Η συνολική όμως ομοιότητα μεταξύ δύο λέξεων περιλαμβάνει έναν επιπλέον όρο, ο οποίος αρχικά έχει τυχαία τιμή, αλλά με την εξέλιξη του αλγόριθμου ACO μεταβάλλεται και αντικατοπτρίζει τις προτιμήσεις της αποικίας. Ο όρος αυτός συμβολίζεται με *d* και αντιστοιχεί στην ποσότητα φερομόνης που υπάρχει στη διασύνδεση μεταξύ δύο λέξεων ή, αλλιώς, σε ένα κομμάτι της συνολικής διαδρομής. Η συνολική ομοιότητα των λέξεων ανά ζευγάρι δίνεται από τον τύπο (5-2):

$$sim_{ij} = (\tau_{ij})^a \cdot (d_{ij})^b \quad (5\text{-}2)$$

όπου sim_{ij} η συνολική ομοιότητα μεταξύ των λέξεων *(I,j)* και *a,b* οι παράμετροι που εστιάζουν στον ένα ή στον άλλο όρο.

Σε κάθε βήμα(εποχή) του αλγόριθμου δημιουργείται ένα σύνολο από άτομα με τον εξής τρόπο: Για κάθε λέξη *i* επιλέγεται με τυχαίο τρόπο, βάσει όμως των τιμών του κριτηρίου sim_{ij}, μία λέξη *j*, οι οποίες στη συνέχεια συνδέονται (εικονικά). Η επιλογή αυτή μοιάζει με το roulette selection των γενετικών αλγορίθμων (Goldberg D. E., 1989). Η τιμή sim_{ij} επηρεάζει την πιθανότητα να επιλεγεί το ζεύγος *(i,j)* και ζεύγη με μεγάλη ομοιότητα επιλέγονται συχνότερα. Αφού για όλες τις λέξεις επιλεγεί τουλάχιστον μία (εικονική) σύνδεση, δημιουργείται ένα σύνολο από ομάδες.

Κάθε ομάδα περιλαμβάνει λέξεις οι οποίες συνδέονται (εικονικά) μεταξύ τους άμεσα ή έμμεσα μέσω άλλων λέξεων, ενώ καμία ομάδα δεν περιλαμβάνει λέξη που να έχει έστω και μία (εικονική) σύνδεση με οποιαδήποτε λέξη άλλης ομάδας. Το αποτέλεσμα αυτό επιτυγχάνεται με τη σύμπτυξη των ομάδων, των οποίων οι λέξεις έχουν μία τουλάχιστον (εικονική) σύνδεση (agglomerative clustering). Αν το αποτέλεσμα που θα προκύψει περιλαμβάνει δύο ή λιγότερες ομάδες, τότε το συγκεκριμένο άτομο καταστρέφεται και δεν αξιολογείται καθόλου στη συνέχεια, αφού ομαδοποιήσεις λέξεων με τόσο λίγες ομάδες είναι σαφώς ανεπιθύμητες.

Στη συνέχεια, οι περισσότερο όμοιες ομάδες συνενώνονται ώστε να προκύψουν λιγότερες ομάδες από κάποιο μέγιστο κατώφλι, ώστε σε επόμενο βήμα να μπορέσει ο κατηγοριοποιητής MLP να χειριστεί τη διάσταση του διανύσματος εισόδου, κάτι ανάλογο δηλαδή με το στάδιο της μείωσης διάστασης του συστήματος που παρουσιάσθηκε στο κεφάλαιο 4 με χρήση *k-means*. Η ομοιότητα των ομάδων ορίζεται χρησιμοποιώντας τη μέση τιμή του d_{ij} για όλους τους συνδυασμούς λέξεων, όπου το *i* αντιστοιχίζεται στις λέξεις της μίας ομάδας και το *j* στις λέξεις της άλλης (average linkage (Duda R.O., Hart P.E. & Stork D.G. 2001)).

Έπειτα οι ομάδες διατάσσονται σε μορφή λίστας, με μία υποτυπώδη δομή γειτονίας. Ειδικότερα, κάθε ομάδα έχει δύο συνδέσεις και μπορεί να συνδεθεί με τις περισσότερο όμοιες προς αυτήν ομάδες. Οι συνδέσεις αυτές είναι αμφίδρομες και καταλήγουν σε λίστα μίας διάστασης. Οι συνδέσεις της αρχής με το τέλος απαγορεύεται να ενεργοποιηθούν και η λίστα έχει απαραίτητα διακριτή αρχή και διακριτό τέλος χωρίς να υπάρχουν κύκλοι. Ένα παράδειγμα δημιουργίας της λίστας από μεμονωμένες ομάδες παρουσιάζεται στο σχήμα 5.1.

Ο λόγος δημιουργίας μίας τέτοιας δομής είναι ότι διευκολύνει την αξιολόγηση του αποτελέσματος. Πιο συγκεκριμένα, για κάθε λέξη κάθε πρότασης του σώματος κειμένων εντοπίζεται η αντίστοιχη ομάδα. Έπειτα υπολογίζεται η απόσταση σε επίπεδο πρότασης για όλα τα ζεύγη λέξεων-ομάδων από τη λίστα ομάδων. Η καλύτερη λύση είναι αυτή που δίνει τη μικρότερη συνολική απόσταση. Στη συνέχεια, με βάση μόνο την καλύτερη λύση, ενημερώνονται οι τιμές της φερομόνης d_{ij} για τα επιλεγμένα ζευγάρια λέξεων *(i,j)* κατά τη φάση δημιουργίας του καλύτερου ατόμου. Η επιλογή μόνο του καλύτερου ατόμου για την ενημέρωση των αποτελεσμάτων προέκυψε για λόγους βελτιστοποίησης των υπολογισμών, καθώς έχει ήδη χρησιμοποιηθεί με επιτυχία (Stützle T. & Hoos H., 2000) και παρουσίασε καλύτερα αποτελέσματα από την ενημέρωση που περιλαμβάνει όλες τις λύσεις ανάλογα με το βαθμό της αξιολόγησής τους:

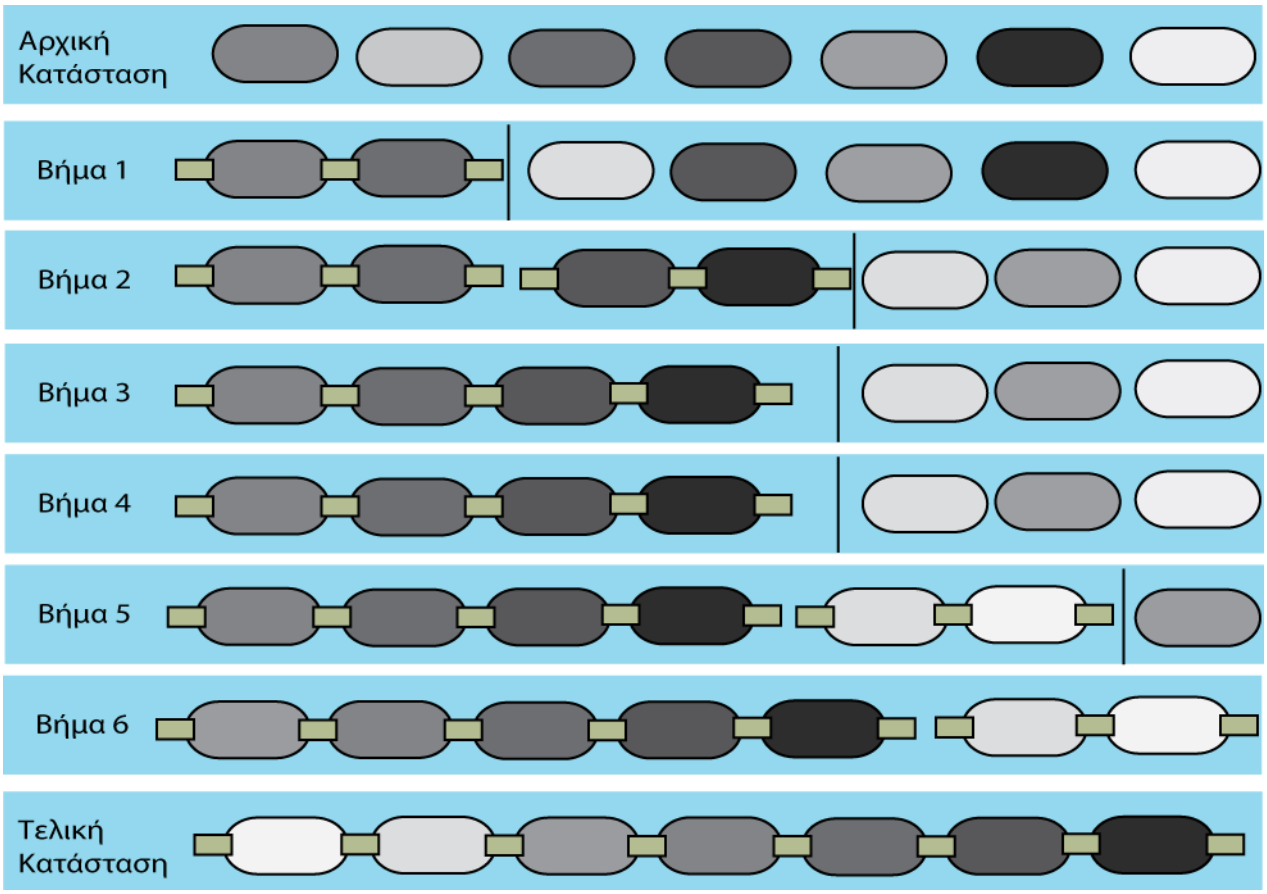
$$d'_{ij} = \begin{cases} f \cdot d_{ij} + Q & \text{Αν } (i,j) \text{ ανήκει στο καλύτερο άτομο} \\ f \cdot d_{ij} & \text{αλλιώς} \end{cases} \quad (5\text{-}3)$$

όπου d'_{ij} είναι η νέα τιμή για τη φερομόνη της διαδρομής (i,j), *F* είναι η παράμετρος που ορίζει το βαθμό «εξάτμισης» της φερομόνης και *Q* είναι μία σταθερή ποσότητα, η οποία υπολογίζεται από τον τύπο (5-4) με σκοπό τη σταθερότητα της συνολικής φερομόνης του συστήματος.

$$Q = \frac{(1-f)\cdot \sum_{j}\sum_{j} d_{ij}}{\#Count(i,j \in Max)} \qquad (5\text{-}4)$$

όπου η συνάρτηση στον παρονομαστή μετράει το πλήθος των διαδρομών - ζευγών (i,j), τα οποία έχουν επιλεγεί από το βέλτιστο άτομο, και το διπλό άθροισμα στον αριθμητή μετράει τη συνολική ποσότητα φερομόνης του συστήματος.

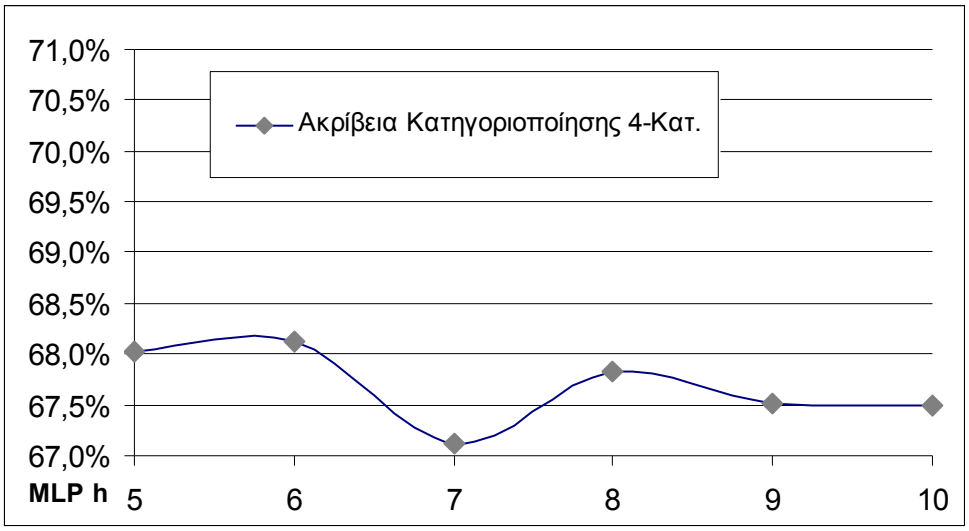
Η διαχρονικά καλύτερη λύση φυλάσσεται (Ελιτισμός), αφού ο αλγόριθμος ACO δεν παρέχει κάποια εγγύηση ότι η καλύτερη λύση της τελευταίας επανάληψης θα είναι καλύτερη ή έστω ισοδύναμη με κάποια ενδιάμεση βέλτιστη λύση. Η ομαδοποίηση που δίνει την καλύτερη τιμή αντικειμενικής συνάρτησης αντικαθιστά το χάρτη SOM και χρησιμοποιείται για την προβολή των κειμένων στο χώρο προτύπων κατ' αντιστοιχία με το σύστημα του κεφαλαίου 4.



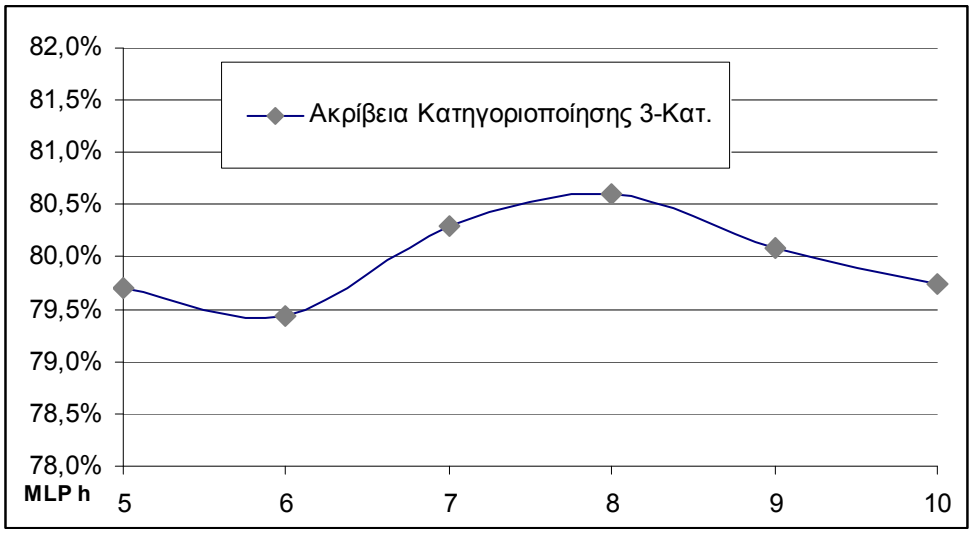
Σχήμα 5.1. Παράδειγμα εκτέλεσης διαδικασίας δημιουργίας γειτονίας ομάδων από μεμονωμένες ομάδες. Τα χρώματα των ομάδων προσομοιώνουν την ομοιότητά τους.

5.1.2 Πειράματα ACO

Για την εκτίμηση της απόδοσης του συστήματος έγιναν κάποια πειράματα με το σώμα κειμένων της Βουλής (Α' σώμα κειμένων). Οι παράμετροι που επιλέχτηκαν για το σύστημα ήταν οι εξής: οι τιμές *a,b* για την απόσταση (5-2) ήταν ίσες με 1, ο συντελεστής «εξάτμισης» της φερομόνης ίσος με *0,97* (δηλαδή το 3% της φερομόνης εξατμίζεται σε κάθε επανάληψη), το μέγιστο μήκος πρότασης τέθηκε στο *50*, καθαρά για υπολογιστικούς λόγους, και το πλήθος των απαιτούμενων ομάδων ήταν *50*. Οι πληθυσμοί που εξετάσθηκαν είχαν μέγεθος 30 ατόμων. Η ακρίβεια που επιτεύχθηκε για διάφορα μεγέθη MLP δίνεται στο σχήμα 5.2 για το πείραμα των 4 κατηγοριών και στο σχήμα 5.3 για το πείραμα των 3.



Σχήμα 5.2. Ακρίβεια κατηγοριοποίησης για το σύστημα ACO για διάφορα μεγέθη MLP για το σώμα κειμένων της Βουλής (Α' σώμα κειμένων) για τις 4 κατηγορίες



Σχήμα 5.3. Ακρίβεια κατηγοριοποίησης για το σύστημα ACO για διάφορα μεγέθη MLP για το σώμα κειμένων της Βουλής (Α' σώμα κειμένων) για τις 3 κατηγορίες

Όπως δείχνουν τα πειράματα, η ακρίβεια κατηγοριοποίησης που επιτυγχάνει το σύστημα ACO είναι χαμηλότερη από την ακρίβεια κατηγοριοποίησης του αντίστοιχου συστήματος SOM. Πιο συγκεκριμένα, για το πρόβλημα των 4 κατηγοριών το σύστημα ACO δεν υπερβαίνει το 68,1% σε σχέση με το 76,5% του συστήματος SOM. Ομοίως, για το πείραμα των 3 κατηγοριών η μέγιστη ακρίβεια που δίνει το σύστημα ACO είναι 80,6%, δηλαδή αρκετά μικρότερη από την αντίστοιχη του SOM (84,3%).

Πρέπει εδώ να αναφερθεί ότι το σύστημα ACO που εξετάσθηκε και είχε 50 ομάδες είναι αρκετά πιο αργό από τα συστήματα SOM, καθώς απαιτεί περίπου δεκαπλάσιο χρόνο από την πλήρη εκπαίδευση του SOM για την εκτέλεση 40 επαναλήψεων. Αυτό οφείλεται στο γεγονός ότι η μοντελοποίηση των τ_{ij} και sim_{ij} απαιτεί χώρο μνήμης αντίστοιχο του τετραγώνου των εξεταζόμενων λέξεων (οι οποίες είναι της τάξεως των 7000 για τα σχετικά περιορισμένα κειμενικά δεδομένα της Βουλής). Επιπλέον, το απαιτούμενο πλήθος των μεταβλητών καθιστά τη δημιουργία νέων ατόμων του πληθυσμού αλλά και την ενημέρωση των τιμών της φερομόνης εξαιρετικά χρονοβόρα διαδικασία.

Το πρόβλημα γίνεται ιδιαίτερα έντονο, εάν ληφθεί υπόψιν ότι το συγκεκριμένο σύστημα δεν μπόρεσε να εφαρμοσθεί στο μεγαλύτερο σώμα κειμένων της εφημερίδας (Β' σώμα κειμένων), αφού οι απαιτήσεις σε μνήμη και υπολογισμούς κατέστησαν την εφαρμογή του ανέφικτη. Παρά την ενδιαφέρουσα προσπάθεια για επίλυση του προβλήματος με χρήση ACO, εντούτοις η μεθοδολογία αυτή σε πραγματικές συνθήκες της δεδομένης εφαρμογής κρίνεται ανεπαρκής. Πιθανώς η μετατροπή του αλγόριθμου και η εκτέλεση του σε παράλληλες αρχιτεκτονικές υπολογιστών να μπορούσε να βελτιώσει τα αποτελέσματα, όμως αυτό το ζήτημα ξεφεύγει από το αντικείμενο της παρούσας μελέτης.

5.2 Σύστημα οργάνωσης κειμένων βασισμένο σε GMM - RBF

5.2.1 Περιγραφή GMM

Το μοντέλο GMM (Gaussian Mixture Models) εφαρμόζεται σε προβλήματα ομαδοποίησης δεδομένων (clustering) (Everitt B. S., Landau S., Leese M. 2001) και βασίζεται στην παραδοχή ότι όλα τα δεδομένα προέρχονται από ένα σύνολο Γκαουσσιανών κατανομών με διακριτά πολυδιάστατα κέντρα και διασπορές. Οι παράμετροι του μοντέλου αποτελούνται από:

1) τα **κέντρα**, δηλαδή διανύσματα που τοποθετούν χωροταξικά τις ομάδες,
2) τον **πίνακα συμμεταβλητότητας** (**covariance**), που ουσιαστικά επηρεάζει την ακτίνα των κέντρων ανά διάσταση και
3) το **διάνυσμα των βαρών**, που εκφράζει το βαθμό που περιγράφει τα δεδομένα η κάθε κατανομή.

Ο βαθμός που ένα δεδομένο αποδίδεται σε μία ομάδα (βαθμός συμμετοχής) εξαρτάται από την απόστασή του από αυτήν και δίνεται από τον ακόλουθο τύπο, ο οποίος εκφράζει την Γκαουσσιανή συνάρτηση:

$$\mu(\vec{x},\vec{m}_t) = \frac{1}{(2\pi)^{N/2}\cdot|\Sigma_i|^{1/2}} \cdot exp[-\frac{1}{2}(\vec{x}-\vec{m}_t)^T \Sigma_i^{-1}(\vec{x}-\vec{m}_t)] \quad (5\text{-}5)$$

όπου $\vec{x}$ το εκάστοτε δεδομένο, $\vec{m}_t$ το κέντρο της ομάδας *i*, Σ_i ο πίνακας συμμεταβλητότητας και *N* το μέγεθος του διανύσματος εισόδου που είναι ίσο με τη διάσταση των $\vec{x}$ και $\vec{m}_t$.

Η συνάρτηση (5-5) δεν έχει απαραίτητα πεδίο τιμών το [0,1] και ως εκ τούτου δεν μπορεί να θεωρηθεί σε καμία περίπτωση ως πιθανότητα. Για το λόγο αυτό χρησιμοποιείται εδώ ο όρος «βαθμός συμμετοχής». Η Γκαουσσιανή συνάρτηση είναι γνωστή ως μία συνάρτηση πυκνότητας πιθανότητας (probability density function), αλλά στα GMM χρησιμοποιείται ως συνάρτηση που εκτιμά τη σχέση ενός δεδομένου με κάποια ομάδα.

Ο αλγόριθμος εκπαίδευσης των GMM (Bilmes J. A., 1997) βασίζεται στην εφαρμογή του γνωστού αλγόριθμου EM (Expectation – Maximization) (Dempster A., Laird N. & Rubin D., 1977) και σκοπεύει στο να προσαρμόσει το μοντέλο (δηλαδή τα κέντρα ομάδων και τον πίνακα συμμεταβλητότητας), ώστε αυτό με τη σειρά του να δίνει για το σύνολο δεδομένων μεγαλύτερους βαθμούς συμμετοχής.

Ο αλγόριθμος ΕΜ αναζητεί τοπικά λύσεις (greedy algorithm) και κάθε βήμα του βελτιώνει απαραιτήτως τη λύση που έχει εντοπιστεί, έως ότου ο ρυθμός βελτίωσης μειωθεί και συγκλίνει σε μία τελική λύση, η οποία αποτελεί συνήθως τοπικό ακρότατο. Στο πρώτο βήμα του αλγόριθμου (E-step) υπολογίζεται η σχετική συμβολή της κάθε ομάδας σε κάθε δεδομένο $\vec{x}$ ως εξής:

$$p(\vec{m}_i, \vec{x}) = \frac{\mu(\vec{m}_i, \vec{x}) \cdot P(\vec{m}_i)}{\sum_{i=1}^{k} \mu(\vec{m}_i, \vec{x}) \cdot P(\vec{m}_i)} \tag{5-6}$$

όπου το $P(\vec{m}_t)$ αποτελεί παράμετρο του μοντέλου, η οποία εκφράζει το βαθμό που η κατανομή $\vec{m}_t$ της ομάδας *i* περιγράφει τα δεδομένα, και το άθροισμα στον παρονομαστή ισχύειγια το σύνολο των ομάδων *k*. Με βάση τον τύπο (5-6) ενημερώνονται οι παράμετροι του μοντέλου ως εξής (M-Step):

Για τα βάρη:

$$P(\vec{m}_i) = \frac{1}{L} \sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j) \tag{5-7}$$

όπου *L* είναι το πλήθος των δεδομένων.

Για τα κέντρα:

$$\vec{m}_i = \frac{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j) \cdot \vec{x}_j}{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j)} \tag{5-8}$$

Για τον πίνακα συμμεταβλητότητας υπάρχουν 3 παραλλαγές, οι οποίες επιβάλλουν γεωμετρικούς περιορισμούς στις ομάδες, όπως φαίνεται στο σχήμα 5.4, και είναι οι εξής:

1) **Σφαιρικές** γεωμετρίες, όταν ο πίνακας *Σ* είναι της μορφής *σI* (όπου *I* ο μοναδιαίος πίνακας):

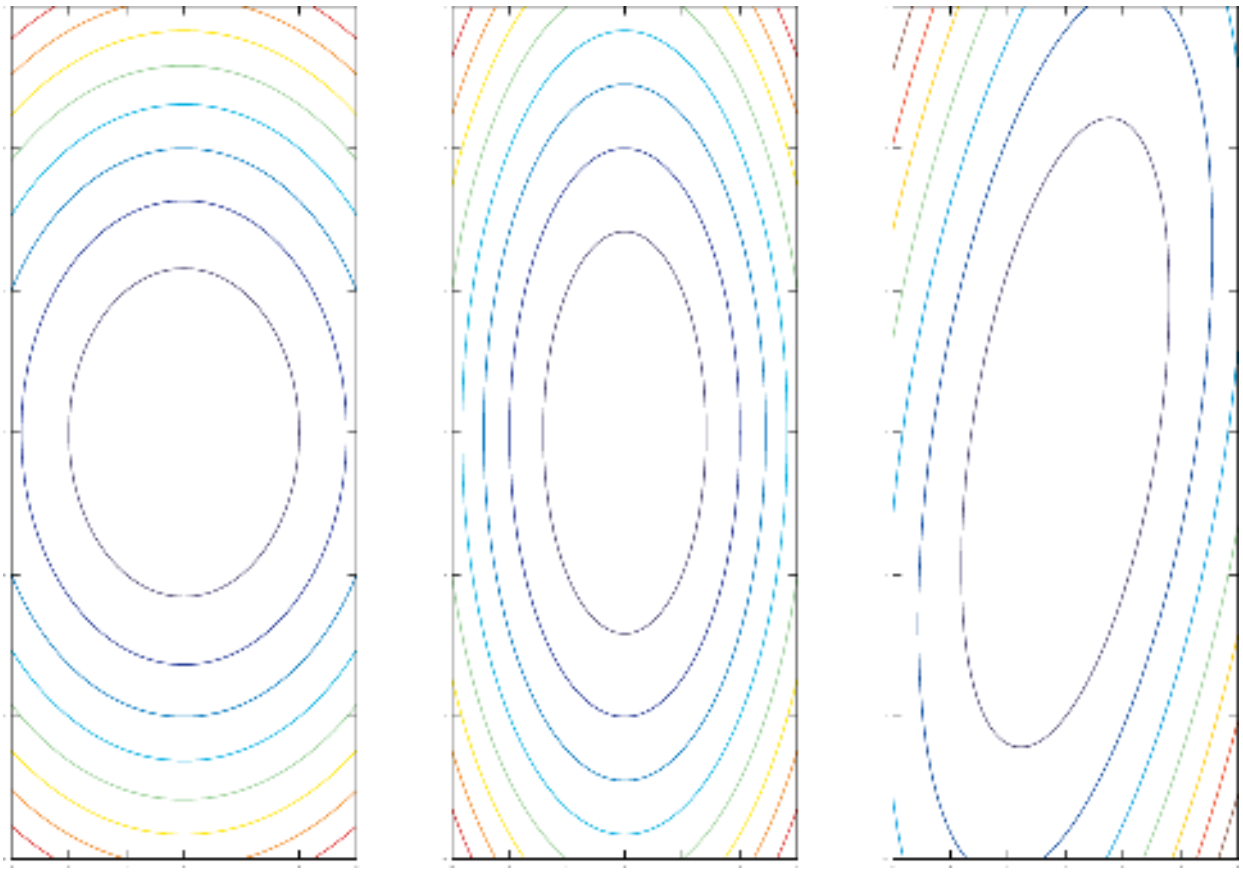
$$(\sigma_i)^2 = \frac{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j) \cdot \left\| \vec{x}_j - \vec{m}_i \right\|^2}{\mathrm{N} \sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j)} \tag{5-9α}$$

2) **Ελλειψοειδείς** ομάδες με τους **άξονες παράλληλους** στο σύστημα συντεταγμένων, όταν ο πίνακας *Σ* είναι διαγώνιος και έχει μη μηδενικά μόνο τα *k* στοιχεία της διαγωνίου *(σ_k)*:

$$(\sigma_{ik})^2 = \frac{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j) \cdot (x_{jk} - m_{ik})^2}{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j)} \qquad (5\text{-}9\beta)$$

3) Για **ελλειψοειδείς** ομάδες με **οποιουσδήποτε άξονες** (γενικότερη περίπτωση):

$$\Sigma_i = \frac{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j) \cdot (\vec{x}_j - \vec{m}_i)(\vec{x}_j - \vec{m}_i)^T}{\sum_{j=1}^{L} p(\vec{m}_i, \vec{x}_j)} \qquad (5\text{-}9\gamma)$$



Σχήμα 5.4. Γεωμετρική αναπαράσταση των ομάδων που δημιουργούν τα GMM σε δύο διαστάσεις ανάλογα με τον πίνακα συμμεταβλητότητας. Από αριστερά προς τα δεξιά έχουμε: Σφαιρικές, Ελλειψοειδείς και Ελλειψοειδείς με εστραμμένους άξονες.

Τα βήματα Ε και Μ επαναλαμβάνονται έως ότου το κριτήριο πιθανοφάνειας για το σύνολο των δεδομένων για διαδοχικές επαναλήψεις να σταματήσει να μεταβάλλεται σημαντικά, οπότε και συγκλίνει στην τελική λύση. Το συνολικό κριτήριο πιθανοφάνειας για το σύνολο των δεδομένων *X* δίνεται από τον τύπο (5-10):

$$s(X) = \prod_{l=1}^{L} \left[\sum_{i=1}^{k} \mu(\vec{m}_i, \vec{x}_l) \cdot P(\vec{m}_i) \right] \tag{5-10}$$

Ένα πολύ σημαντικό στοιχείο στην υλοποίηση του μοντέλου GMM είναι η αρχικοποίησή του. Δυστυχώς όμως δεν υπάρχει κάποιος ντετερμινιστικός αλγόριθμος που να τοποθετεί τις παραμέτρους σε τιμές που θα εξασφαλίσουν την καλύτερη εξέλιξή του.

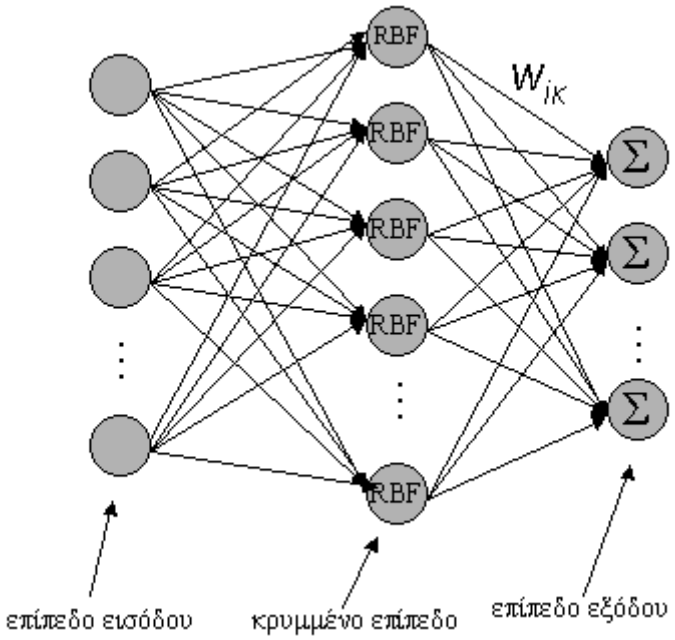
Επιπλέον, ο αλγόριθμος εκπαίδευσης ΕΜ πραγματοποιεί αναζήτηση καλύτερων λύσεων σε τοπικές περιοχές, γεγονός που καθιστά την αρχικοποίηση πολύ σημαντική για το τελικό αποτέλεσμα. Συνηθέστερη τακτική είναι η χρησιμοποίηση ενός κοινού αλγόριθμου ομαδοποίησης όπως το *k-means*, ώστε να βρεθεί ένα καλό αρχικό σημείο για τον αλγόριθμο (Everitt B. S., Landau S. & Leese M. 2001). Διάφορες παραλλαγές για τον αλγόριθμο εκπαίδευσης GMM έχουν προταθεί (Figueiredo M.A.T. & Jain A.K., 2002 και Pernkopf F. & Bouchaffra D., 2005), οι οποίες όμως χρησιμοποιούν και τον ΕΜ.

5.2.2 Περιγραφή RBF

Τα δίκτυα ακτινικών συναρτήσεων βάσης (Radial Basis Function Networks - RBF) αποτελούνται από τρία επίπεδα, όπως φαίνεται στο σχήμα 5.5 (Haykin S. 1999):

- Το **επίπεδο εισόδου**, όπου και εφαρμόζονται οι είσοδοι στο σύστημα.
- Το **ενδιάμεσο κρυφό επίπεδο** και
- Το **επίπεδο εξόδου**

Τα RBF διαφοροποιούνται από τα MLP στο ότι στο κρυφό επίπεδο χρησιμοποιούνται ακτινικές συναρτήσεις ενεργοποίησης.



Σχήμα 5.5. Αρχιτεκτονική των νευρωνικών δικτύων RBF

Όλοι οι νευρώνες του επιπέδου εισόδου συνδέονται απαραίτητα με κάθε νευρώνα του κρυφού επιπέδου και αποτελούν τις εισόδους στις ακτινικές συναρτήσεις βάσης. Μεταξύ του επιπέδου εισόδου και του κρυφού επιπέδου οι συνδέσεις αυτές έχουν σταθερά βάρη ίσα με τη μονάδα $w^1_{ij}=1$.

Οι νευρώνες του κρυφού επιπέδου συνδέονται με το επίπεδο εξόδου με μεταβλητά βάρη w^2_{ij}. Στους κόμβους του επιπέδου εξόδου μερικές φορές χρησιμοποιούνται συναρτήσεις ενεργοποίησης όμοιες με αυτές των δικτύων MLP, συχνότερα ωστόσο χρησιμοποιούνται απλές γραμμικές συναρτήσεις.

Οι νευρώνες του κρυφού επιπέδου χρησιμοποιούν ακτινικές συναρτήσεις ενεργοποίησης, δηλαδή συναρτήσεις των οποίων οι τιμές εξαρτώνται από την απόσταση του διανύσματος εισόδου από κάποια κέντρα. Πιο συγκεκριμένα, οι ακτινικές συναρτήσεις είναι της μορφής $\Phi\left(\left\|\vec{x}-\vec{m}\right\|\right)$. Τέτοια συνάρτηση είναι και η συνάρτηση Gauss, που συνιστά την πιο συνηθισμένη ακτινική συνάρτηση (βλ. τύπο (5-5)).

Η έξοδος του δικτύου RBF μπορεί να υπολογιστεί ως ένας γραμμικός συνδυασμός των εξόδων των νευρώνων του κρυφού επιπέδου που χρησιμοποιούν ακτινικές συναρτήσεις βάσης. Πιο αναλυτικά, αν συμβολίσουμε με s_i την *i* έξοδο του δικτύου, τότε η τιμή της εξόδου είναι:

$$s_i = \sum_{j=1}^{h} w_{ji}\Phi_j\left(\left\|\vec{x}-\vec{m}_j\right\|\right) \qquad (5\text{-}11)$$

όπου w_{ji} το βάρος που συνδέει τον *j* νευρώνα του κρυφού επιπέδου με τον *i* νευρώνα εξόδου, Φ_j η ακτινική συνάρτηση του *j* νευρώνα του κρυφού επιπέδου, m_j το κέντρο αυτής και *h* ο αριθμός των νευρώνων του κρυφού επιπέδου. Επειδή μεταβλητά βάρη υπάρχουν μόνο σε ένα επίπεδο (μεταξύ κρυφού επιπέδου και εξόδου), δε χρειάζεται επιπλέον δείκτης για τα βάρη που να δηλώνει το επίπεδο αυτών. Γενικά χρησιμοποιείται ο ίδιος τύπος ακτινικής συνάρτησης για όλους τους νευρώνες του δικτύου RBF, ενώ αλλάζει μόνο το κέντρο και η ακτίνα της ακτινικής συνάρτησης ενεργοποίησης.

Μία πολύ σημαντική ιδιότητα των δικτύων RBF, όπως και στην περίπτωση των δικτύων MLP, είναι η δυνατότητα που έχουν να «μαθαίνουν», δηλαδή να προσαρμόζουν τις διάφορες παραμέτρους τους, ώστε να παράγουν εξόδους οι οποίες να πλησιάζουν τις «επιθυμητές». Στην περίπτωση των δικτύων RBF, η εκπαίδευση γίνεται με χρήση τεχνικών επιβλεπόμενης μάθησης (supervised learning). Σκοπός της διαδικασίας μάθησης είναι και σε αυτή την περίπτωση, η όσο το δυνατόν καλύτερη προσέγγιση των εξόδων του δικτύου στις επιθυμητές εξόδους των προτύπων, που παρουσιάζονται στο δίκτυο κατά τη φάση της εκπαίδευσης.

Η ιδιομορφία που παρουσιάζουν τα δίκτυα ακτινικών συναρτήσεων βάσης σε σχέση με τα πολυστρωματικά δίκτυα είναι ότι αποτελούνται από διαφορετικά είδη νευρώνων. Εξαιτίας αυτής τους της ιδιομορφίας, δεν μπορούν να εκπαιδευτούν αποτελεσματικά με αλγόριθμο ανάστροφης διάδοσης. Αυτό οφείλεται στο γεγονός ότι οι παράμετροι είναι ανομοιογενείς και επηρεάζουν με διαφορετικό τρόπο η καθεμία τη λειτουργία του δικτύου. Πιο συγκεκριμένα, τα δίκτυα ακτινικών συναρτήσεων βάσης έχουν τις ακόλουθες παραμέτρους που πρέπει να υπολογιστούν:

- Τα κέντρα των ακτινικών συναρτήσεων βάσης, τα οποία είναι διανύσματα με διάσταση ίση με τον αριθμό εισόδων του δικτύου.
- Τις ακτίνες των ακτινικών συναρτήσεων βάσης, που είναι καθαροί αριθμοί.
- Τα βάρη του γραμμικού στρώματος εξόδου

Επομένως, για ένα δίκτυο RBF που έχει *R* εισόδους, *H* νευρώνες στο κρυφό επίπεδο και *M* εξόδους πρέπει να υπολογιστούν *RxH* παράμετροι για τα κέντρα, *H* για τις ακτίνες και *(H+1)xM* βάρη και κατώφλια. Κατά τον προσδιορισμό αυτών των παραμέτρων, προκειμένου να μειωθεί το σφάλμα στο σύνολο των δεδομένων εκπαίδευσης, διακρίνονται δύο φάσεις:

1. Υπολογισμός των παραμέτρων των ακτινικών συναρτήσεων βάσης (τοποθεσίες των κέντρων, ακτίνες των κέντρων, πίνακας συμμεταβλητότητας) χρησιμοποιώντας ένα αλγόριθμο ομαδοποίησης, όπως τα μοντέλα GMM που αναφέρθηκαν στην προηγούμενη ενότητα.
2. Δεδομένου ότι είναι γνωστές οι επιθυμητές έξοδοι, προσαρμογή των βαρών του επιπέδου εξόδου, ελαχιστοποιώντας το μέσο τετραγωνικό σφάλμα. Πιο αναλυτικά: Έστω y (διάσταση *1xM*) το διάνυσμα των εξόδων του δικτύου για όλα τα δεδομένα και *S* το διάνυσμα των εξόδων του κρυφού επιπέδου (με διάσταση *1xH*). Στον πίνακα *W* των βαρών προστίθενται και τα κατώφλια, εάν θεωρηθεί ότι και αυτά είναι βάρη δεχόμενα είσοδο πάντα ίση με ένα, οπότε ο πίνακας των βαρών έχει διάσταση *(H+1)xM*. Και το διάνυσμα *S* θα έχει διάσταση *1x(H+1)*, αν στο τέλος του προσθέσουμε τη μονάδα (ως είσοδο για τα κατώφλια). Με βάση τα παραπάνω ισχύει:

$$y = Sw \tag{5-12}$$

Είναι επιθυμητή η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος *J*, το οποίο σε μορφή πινάκων είναι:

$$J = \frac{1}{2}\|d - wS\|^2 \tag{5-13}$$

όπου *d* είναι οι επιθυμητές έξοδοι για το δίκτυο.

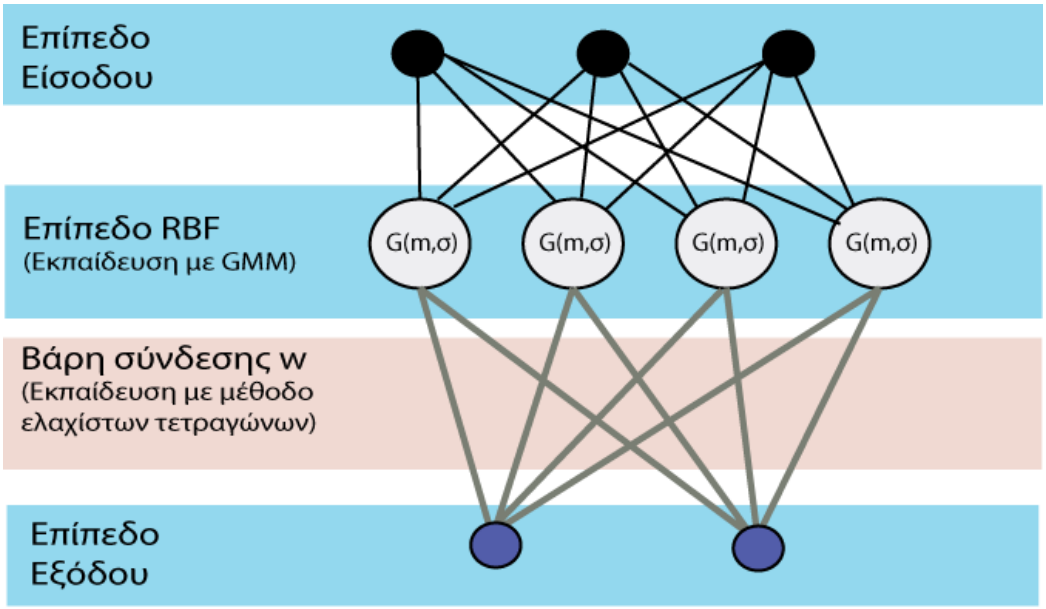
Αναζητώντας το ελάχιστο σφάλμα, τίθεται

$$\frac{dJ}{dw} = 0 \Rightarrow (d - Sw)^{\mathrm{T}}(d - Sw) = 0 \Rightarrow S^{\mathrm{T}}Sw - S^{\mathrm{T}}b = 0 \tag{5-14}$$

η οποία τελικά καταλήγει στη λύση του συστήματος:

$$w = (S^{\mathrm{T}}S)^{-1}S^{\mathrm{T}}d \tag{5-15}$$

Ένα χαρακτηριστικό του αλγόριθμου εκπαίδευσης είναι ότι τα κέντρα των ακτινικών συναρτήσεων βάσης δεν αλλάζουν, αλλά παραμένουν όπως υπολογίστηκαν αρχικά, ενώ τα βάρη υπολογίζονται με μονοσήμαντο τρόπο, ο οποίος δεν είναι επαναληπτικός. Μία συνοπτική εικόνα του συνδυασμού GMM-RBF σε συνδυασμό με τους αλγόριθμους εκπαίδευσης που χρησιμοποιούνται σε κάθε επίπεδο δίνεται στο σχήμα 5.6.



Σχήμα 5.6. Συνοπτική εικόνα του μοντέλου GMM-RBF σε συνδυασμό με τους αλγόριθμους εκπαίδευσης που χρησιμοποιούνται σε κάθε επίπεδο

5.2.3 Περιγραφή συστήματος

Χρησιμοποιώντας τα νευρωνικά δίκτυα RBF σε συνδυασμό με τα GMM (σχήμα 5.6) προκύπτει ένας αλγόριθμος κατηγοριοποίησης ικανός να αντικαταστήσει το MLP που χρησιμοποιήθηκε στο σύστημα της ενότητας 4. Τα GMM υποθετικά θα μπορούσαν να αντικαταστήσουν και τον αλγόριθμο SOM, όμως κάτι τέτοιο δεν είναι εφικτό καθώς τα GMM απαιτούν αριθμό παραμέτρων ανάλογο της διάστασης εισόδου του προβλήματος, η οποία για την περίπτωση των κειμένων της Βουλής (Α' σώμα κειμένων) είναι από 560 έως 640 περίπου, όσο είναι δηλαδή το πλήθος των λέξεων της κατηγορίας Β της ανάλυσης ABC.

Κάθε ομάδα του GMM απαιτεί *M* παραμέτρους για τη διάσταση του κέντρου της, *MxM* για τον πίνακα συμμεταβλητότητας και μία παράμετρο για το βάρος της. Αντίθετα, το μοντέλο SOM απαιτεί για κάθε ομάδα μόλις *M* παραμέτρους, οι οποίες και αποτελούν το κέντρο της. Οι υπολογιστικές αυτές ελαφρύνσεις σε συνδυασμό με την αποτελεσματικότητα του SOM σε προβλήματα μεγάλων διαστάσεων απέτρεψαν την προσπάθεια για αντικατάστασή του από τα GMM. Συνοψίζοντας, ο αλγόριθμος GMM σε συνδυασμό με το RBF χρησιμοποιήθηκε εναλλακτικά του κατηγοριοποιητή MLP στο δεύτερο (και τελευταίο) στάδιο του συστήματος.

5.2.4 Πειράματα

Για την αξιολόγηση του συνδυασμού RBF-GMM μετρήθηκε η ακρίβεια κατηγοριοποίησης για διάφορες παραμέτρους των GMM χρησιμοποιώντας το επιλεγμένο σύμφωνα με την ενότητα 4.4.1.1 μοντέλο χάρτη SOM για το Α' σώμα κειμένων. Δηλαδή χρησιμοποιήθηκε χάρτης μεγέθους 10x15 κατασκευασμένος με μετρήσεις χωρίς κανονικοποίηση και για διάφορες τιμές *k* ομάδων για το *k-means*.

Τα πειράματα υλοποιήθηκαν και για τα δύο σύνολα δεδομένων. Για το Α' σώμα κειμένων εξετάσθηκε τόσο το πλήρες πρόβλημα (4 κατηγορίες) όσο και αυτό χωρίς την κατηγορία «Εσωτερικά» (3 κατηγορίες). Για το Β' σώμα κειμένων επιλέχτηκε ο χάρτης SOM με μέγεθος 4x5 χωρίς το *k-means*, ώστε το πρόβλημα να είναι απλούστερο για το εξεταζόμενο σύστημα. Για τα GMM και κατά συνέπεια για τα RBF επιλέχθηκαν από 3 έως 20 κέντρα Γκαουσιανών (ή νευρώνες RBF) για το Α' σώμα κειμένων και από 10 έως 60 για το Β'. Η υλοποίηση των πειραμάτων έγινε σε περιβάλλον MATLAB χρησιμοποιώντας τα ελεύθερα διαθέσιμα εργαλεία BayesNetToolbox (Murphy K., 2001) και netlab για τα GMM, ενώ για το RBF έγινε καινούρια υλοποίηση συμβατή και με τα δύο αυτά εργαλεία. Επιπλέον εξετάστηκαν πίνακες συμμεταβλητότητας πλήρους μορφής καθώς επίσης και διαγώνιοι.

Είναι σημαντικό να αναφερθεί ότι υπήρξαν παράμετροι του GMM που είχαν ως αποτέλεσμα την αποτυχία του συστήματος στην εξεύρεση εφικτών λύσεων. Μάλιστα, εξαιτίας των αρχικών συνθηκών, πολλές φορές κατέστη αναγκαία η επανάληψη του ίδιου πειράματος αρκετές φορές ώσπου να βρεθεί αποδεκτή λύση. Αυτό οφείλεται στο γεγονός ότι ο πίνακας συμμεταβλητότητας πρέπει να είναι αντιστρέψιμος, αλλά και θετικά ορισμένος. Δυστυχώς αυτός ο περιορισμός πολλές φορές παραβιάζεται με αποτέλεσμα να είναι αναγκαία κάποια στρατηγική επαναφοράς του. Οι στρατηγικές που ακολουθήθηκαν ήταν α) η επαναρχικοποίηση του πίνακα συμμεταβλητότητας με τις αρχικές του τιμές, όπως υλοποιείται από το πακέτο netlab, και β) η πρόσθεση κάποιας μικρής θετικής τιμής στα στοιχεία της διαγώνιου, όπως υλοποιείται από το πακέτο BayesNetToolbox. Αυτές οι στρατηγικές επιδιόρθωσης του πίνακα αποτελούν παραβίαση της εξέλιξης του ΕΜ, με αποτέλεσμα να μην είναι απόλυτα εγγυημένο ότι το επόμενο βήμα του αλγόριθμου θα οδηγεί σε περιοχές μεγαλύτερης πιθανοφάνειας. Η πολυπλοκότητα των δεδομένων όπως και η ύπαρξη χαρακτηριστικών με ελάχιστη πληροφορία είναι παράγοντες που καθιστούν την εξέλιξη του αλγόριθμου προβληματική.

Η καλύτερη επίδοση για το σύστημα GMM-RBF στο Α' σώμα κειμένων για το πρόβλημα με τις 4 κατηγορίες ήταν 63,0% και σημειώθηκε για *k=10* χρησιμοποιώντας την πλήρη μορφή του πίνακα συμμεταβλητότητας. Ομοίως, για το πρόβλημα με τις 3 κατηγορίες η καλύτερη επίδοση ήταν 77,8% για τις ίδιες παραμέτρους. Και οι δύο όμως επιδόσεις είναι χαμηλότερες από τις αντίστοιχες επιδόσεις που επιτυγχάνονται με το MLP. Ιδιαίτερα για το πρόβλημα των 4 κατηγοριών η ακρίβεια είναι 10% χαμηλότερη από την αντίστοιχη του MLP. Η ακρίβεια μειώθηκε κατά πολύ περισσότερο για το Β' σώμα κειμένων, η οποία ήταν 50,6% για *k =60* και διαγώνιο πίνακα συμμεταβλητότητας.

Βάσει των παραπάνω συνάγεται ότι το σύστημα GMM–RBF αντενδείκνυται για το δεδομένο πρόβλημα. Πιθανά αίτια είναι η πολυπλοκότητα των δεδομένων και το μεγάλο μέγεθος του διανύσματος εισόδου. Είναι επίσης πολύ πιθανό να υπάρχουν περιοχές του χάρτη χωρίς καμία πληροφορία, στοιχείο που επιβαρύνει την ευστάθεια του αλγόριθμου επηρεάζοντας τους περιορισμούς για τον πίνακα συμμεταβλητότητας. Η στενή εξάρτηση του αλγόριθμου εκπαίδευσης των GMM από τις αρχικές συνθήκες καθιστά το μοντέλο ευάλωτο σε τοπικά ελάχιστα, τα οποία εμφανίζονται ιδιαιτέρως συχνά σε χώρους μεγάλων διαστάσεων (Curse Of Dimensionality) (Haykin S. 1999). Έτσι, εάν ο αλγόριθμος αρχικοποίησης είναι βάσει αρχής κάποιος πολύ απλός αλγόριθμος ομαδοποίησης (π.χ. *k-means*) και δε δίνει επαρκείς λύσεις, τότε δύσκολα το GMM θα εντοπίσει κάποιο καλό ελάχιστο. Ο αλγόριθμος αυτός «συμβάλλει» και στην αρχικοποίηση του ιδιαίτερα «ευαίσθητου» πί-

νακα συμμεταβλητότητας. Πιθανώς κάποια τεχνική προεπεξεργασίας των δεδομένων όπως είναι η Ανάλυση σε Πρωτεύουσες Συνιστώσες (PCA) (Duda R.O., Hart P.E. & Stork D.G. 2001) να βελτίωνε τα αποτελέσματα, ωστόσο περαιτέρω προσπάθειες βελτίωσης του συστήματος GMM-RBF δεν επιχειρήθηκαν στην παρούσα μελέτη.

5.3 Σύστημα οργάνωσης κειμένων βασισμένο σε HMM

Μία ακόμη εναλλακτική πρόταση στη δημιουργία ομάδων λέξεων αποτελεί η εφαρμογή των κρυφών Μαρκοβιανών μοντέλων (Hidden Markov Model - HMM). Το μοντέλο HMM χρησιμοποιείται στο πρώτο στάδιο του συστήματος της ενότητας 4.1 αντικαθιστώντας το μοντέλο SOM μαζί με το *k-means*.

5.3.1 Περιγραφή HMM

Τα κρυφά Μαρκοβιανά μοντέλα (HMM) είναι στατιστικά μοντέλα που βασίζονται στην αρχή του Bayes (Rabiner L.R. 1989). Η κυριότερη παραδοχή που γίνεται στα HMM είναι ότι η διεργασία που προσομοιάζεται είναι μία Μαρκοβιανή διεργασία (διαδικασία χωρίς μνήμη - memoryless) με μεταβάσεις μεταξύ κρυφών καταστάσεων, της οποίας σκοπός είναι η εύρεση των κρυφών παραμέτρων βάσει της ακολουθίας των δεδομένων που παρατηρούνται στην έξοδο. Ο περιορισμός των Μαρκοβιανών διεργασιών είναι ότι θεωρούν πως η κατάσταση του συστήματος δεν εξαρτάται από την εκάστοτε χρονική στιγμή, παρά μόνο από τις προηγούμενες καταστάσεις του συστήματος, το πλήθος των οποίων ορίζεται ως η τάξη της Μαρκοβιανής διεργασίας.

Στα συμβατικά Μαρκοβιανά μοντέλα οι καταστάσεις του συστήματος είναι άμεσα παρατηρήσιμες σε αντιδιαστολή με τα κρυφά Μαρκοβιανά μοντέλα (HMM) όπου παρατηρήσιμες είναι μόνο οι έξοδοι, οι οποίες επηρεάζονται από τις κρυφές καταστάσεις. Στα HMM σε κάθε κατάσταση ορίζεται μία στοχαστική κατανομή, η οποία εκφράζει την πιθανότητα να εμφανισθεί κάποιο σύμβολο στη δεδομένη κατάσταση. Αυτό έχει ως αποτέλεσμα η ακολουθία των παρατηρούμενων συμβόλων να παρέχει κάποια πληροφόρηση σχετικά με τις εναλλαγές καταστάσεων του συστήματος.

Τα HMM είναι πολύ αποτελεσματικά στην κωδικοποίηση ακολουθιακών φαινομένων, των οποίων η δημιουργία δεν μπορεί να αποδοθεί άμεσα σε κάποια αίτια και έχει στοχαστική φύση. Τέτοια ζητήματα προκύπτουν συχνά στον χώρο της επεξεργασίας φυσικής γλώσσας (natural language processing - NLP), όπου για παράδειγμα πολλοί γραμματικοί επισημειωτές (taggers) (Brants T. 1999) βασίζονται σε Μαρκοβιανά μοντέλα. Επιπλέον, τα HMM εφαρμόζονται στην αναγνώριση φωνής (speech recognition) και σε άλλα προβλήματα χρονικά μεταβαλλόμενων δεδομένων.

Τα HMM αποτελούνται από ένα σύνολο καταστάσεων, κάθε μία από τις οποίες αντιστοιχεί σε μία κατάσταση του συστήματος. Στα διακριτά HMM, τα οποία και χρησιμοποιήθηκαν στο προτεινόμενο σύστημα, σε κάθε κατάσταση κάποιο διακριτό σύμβολο μπορεί να παράγεται από ένα διαθέσιμο σύνολο συμβόλων. Το σύστημα μεταβάλλεται μετακινούμενο από μία κατάσταση σε μία άλλη με βάση μία διακριτή συνάρτηση πυκνότητας πιθανότητας. Για την αναπαράσταση αυτής της συνάρτησης χρησιμοποιείται ένας τετραγωνικός πίνακας που αναπαριστά την πιθανότητα μεταβολής του συστήματος από μία κατάσταση *i* σε μία κατάσταση *j* για οποιαδήποτε χρονική στιγμή. Αντίστοιχα, ένας άλλος πίνακας (διαφορετικός για κάθε κατάσταση) περιγράφει την πιθανότητα που έχει κάθε σύμβολο να «προβληθεί» από τη δεδομένη κατάσταση. Τέλος, ένας άλλος πίνακας περιγράφει την πιθανότητα το σύστημα να βρίσκεται αρχικά σε κάθε μία από τις καταστάσεις του. Οι τρεις αυτοί πίνακες είναι ανεξάρτητοι του χρόνου βάσει της Μαρκοβιανής υπόθεσης και αποτελούν τις παραμέτρους του HMM.

Με βάση τις παραμέτρους του HMM ορίζονται τρία βασικά προβλήματα προς επίλυση, όπως περιγράφονται από τον Rabiner (Rabiner L.R. 1989):

1. Η εύρεση της πιθανότητας να προήλθε μία ακολουθία συμβόλων από ένα δεδομένο μοντέλο.
2. Η εύρεση της ακολουθίας καταστάσεων του συστήματος που θα μπορούσε να παράγει μία δεδομένη ακολουθία συμβόλων με τη μεγαλύτερη πιθανότητα (δυναμικός προγραμματισμός - Viterbi).
3. Η εκτίμηση των παραμέτρων του μοντέλου, ώστε η δεδομένη συμβολοακολουθία να παράγεται με τη μεγαλύτερη πιθανότητα.

Η τρίτη διαδικασία επελέγη για το προτεινόμενο σύστημα. Η εκτίμηση των παραμέτρων του μοντέλου μπορεί να εκληφθεί ως ο αλγόριθμος εκπαίδευσης, αφού στοχεύει στην προσαρμογή του μοντέλου HMM στα δεδομένα. Ο αλγόριθμος αυτός, γνωστός ως Baum Welch, αποτελεί υλοποίηση του Expectation-Maximization (Dempster A., Laird N. & Rubin D., 1977). Βελτιστοποιεί τοπικά τις παραμέτρους και βασίζεται σε σημαντικό βαθμό στις αρχικές τιμές του συστήματος (greedy).

5.3.2 Περιγραφή συστήματος

Στο πρώτο στάδιο του συστήματος και αμέσως μετά την προεπεξεργασία το μοντέλο HMM χρησιμοποιείται ως μέθοδος ομαδοποίησης λέξεων ανάλογα με τη νοηματική τους συγγένεια. Η βασική παραδοχή και εδώ είναι το ότι οι λέξεις που εμφανίζονται στην ίδια περίοδο δεν μπορεί παρά να συνδέονται νοηματικά. Και στην προκείμενη περίπτωση, για την απλοποίηση του συστήματος οι λειτουργικές λέξεις έχουν αφαιρεθεί από τις περιόδους.

Στο σύστημα κάθε σύμβολο αντιστοιχίζεται σε μία λέξη (ή λήμμα για την ακρίβεια) και μπορεί να παράγεται από πολλές καταστάσεις ανάλογα με την εκάστοτε συνάρτηση πυκνότητας πιθανότητας. Αυτό το χαρακτηριστικό επιτρέπει στην κάθε λέξη να αντιστοιχίζεται σε περισσότερες της μίας ομάδες, επιτυγχάνοντας με αυτό τον τρόπο μία πιο εκφραστική απεικόνιση, η οποία θα μπορούσε να βοηθήσει σε προβλήματα αποσαφήνισης του νοήματος και αποτελεί πλεονέκτημα σε σχέση με το σύστημα SOM του κεφαλαίου 4. Η αντιστοίχηση του συστήματος HMM, σε αντίθεση με τους περισσότερους αλγόριθμους ομαδοποίησης που αντιστοιχίζουν κάθε πρότυπο σε μία μόνο ομάδα (hard-clustering), έχει ως αποτέλεσμα τη δημιουργία συναρτήσεων συμμετοχής (membership functions) με τρόπο όμοιο με τους αλγόριθμους ομαδοποίησης ασαφούς λογικής (fuzzy clustering algorithms) (MacKay D. 2003 και Duda R.O., Hart P.E. & Stork D.G. 2001). Επιπλέον, το μοντέλο HMM χειρίζεται τις λέξεις ως διακριτά σύμβολα και δεν απαιτεί κάποια διανυσματική αναπαράσταση, η οποία θα οδηγούσε σε απώλεια πληροφορίας καθώς θα απαιτούσε, για υπολογιστικούς λόγους περιορισμό των λέξεων προς θεώρηση (ώστε να περιοριστεί η διάσταση του πίνακα συνεμφανίσεων).

Η μοντελοποίηση που ακολουθήθηκε ώστε το HMM να παραγάγει φυσική γλώσσα ήταν η εξής: Κάθε περίοδος θεωρήθηκε ως μία ακολουθία παρατηρήσεων. Κάθε λέξη αντιμετωπίζεται ως ένα τέτοιο διακριτό σύμβολο, το οποίο μπορεί να παράγεται από κάποια κατάσταση του μοντέλου. Κάθε κατάσταση του HMM μπορεί να παράγει με μία πιθανότητα (ανάλογα με τη διακριτή συνάρτηση πυκνότητας πιθανότητας) κάποια από τις δυνατές λέξεις του σώματος κειμένων. Αφού το άθροισμα κάθε συνάρτησης πυκνότητας πιθανότητας είναι εξ ορισμού ίσο με 1, η κάθε λέξη αποδίδεται στην ομάδα που αντιστοιχεί στην κατάσταση που την παράγει με τη μέγιστη πιθανότητα.

Θα ήταν επιθυμητό το σύστημα να αντιστοιχεί τις λέξεις που συνεμφανίζονται συχνά στην ίδια περίοδο στις ίδιες ομάδες, αφού είναι πιο πιθανό να συνδέονται νοηματικά. Με βάση αυτή την παραδοχή, στο μοντέλο HMM η πιθανότητα της παραμονής του συστήματος στην

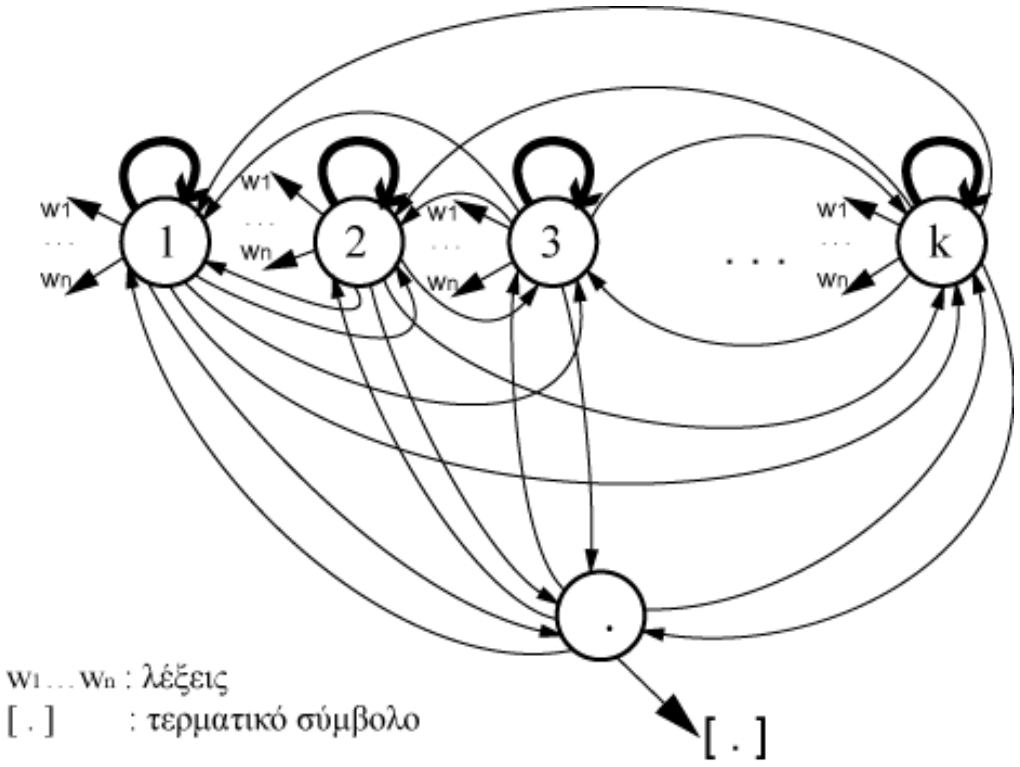
ίδια κατάσταση πρέπει να είναι σχετικά υψηλή, ώστε να μην υπάρχουν πολλές μεταβάσεις καταστάσεων για τις λέξεις της εκάστοτε περιόδου.

Ο περιορισμός αυτός μπορεί να τεθεί ως αρχική συνθήκη, ορίζοντας μία πολύ υψηλή τιμή στην πιθανότητα παραμονής στην ίδια κατάσταση στον πίνακα μετάβασης. Επειδή ο αλγόριθμος εκπαίδευσης Baum Welch βελτιστοποιεί τοπικά (greedy), πιθανότατα δε θα αλλοιώσει το συγκεκριμένο περιορισμό, ωστόσο η ικανοποίηση της συνθήκης αυτής πρέπει να ελέγχεται στο τέλος της εκπαίδευσης.

Για την καλύτερη εκπαίδευση του HMM είναι απαραίτητο να χρησιμοποιείται μεγάλος όγκος δεδομένων. Το σύνολο εκπαίδευσης περιλαμβάνει το σύνολο των προτάσεων των διαθέσιμων κειμενικών δεδομένων, για το οποίο σύνολο δεν απαιτείται καμία ιδιαίτερη επεξεργασία, στοιχείο που καθιστά την υλοποίηση ιδιαίτερα αποδοτική. Επειδή το HMM πρέπει να χειρίζεται πολλαπλές προτάσεις (ακολουθίες συμβόλων), η αρχιτεκτονική που θα επιλεγεί θα πρέπει να έχει προβλέψει αυτή την παράμετρο.

Η αρχιτεκτονική του HMM συνίσταται στο πλήθος των καταστάσεων αλλά και στις επιτρεπόμενες μεταβάσεις. Ειδικότερα, αν μία πιθανότητα μετάβασης τεθεί αρχικά στο 0, τότε ο αλγόριθμος Baum Welch δε θα μπορέσει ποτέ να αποδώσει σε αυτή τη μετάβαση μη μηδενική τιμή. Το μοντέλο θα πρέπει να έχει μία εικονική αρχική κατάσταση και να μεταβαίνει σε αυτή με το χειρισμό κάθε καινούριας πρότασης. Για να επιτευχθεί ο στόχος των πολλαπλών αρχικοποιήσεων δημιουργήθηκε μία νέα ειδική κατάσταση, η οποία παράγει το τερματικό σύμβολο της πρότασης (την τελεία «.») και δεν αντιστοιχίζεται σε ομάδα.

Αρχικά, το σύστημα με πιθανότητα 1 στον πίνακα της αρχικής κατάστασης τίθεται στην ειδική αυτή κατάσταση. Στη συνέχεια, με το πέρας κάθε περιόδου, όταν «παράγεται» το τερματικό σύμβολο «.», το HMM επανέρχεται σε αυτή την ειδική αρχική κατάσταση. Ένα παράδειγμα της υλοποιημένης αρχιτεκτονικής παρουσιάζεται στο σχήμα 5.7.



Σχήμα 5.7. Αρχιτεκτονική του συστήματος HMM

Μία πολύ σημαντική παράμετρο της υλοποίησης του συστήματος συνιστά η μετατροπή των τιμών των πιθανοτήτων σε άλλη κλίμακα τιμών (scaling) (Rabiner L.R. 1989) (Rahimi A.). Αφού το σύστημα χειρίζεται εξαιρετικά εκτενείς ακολουθίες, είναι πολύ πιθανό ότι οι αριθμοί διπλής ακρίβειας (double) των 32-bit υπολογιστικών συστημάτων δε θα έχουν επαρκή ακρίβεια. Η μετατροπή προστατεύει το σύστημα από απώλειες της ακρίβειας όταν το εύρος των τιμών κινητής υποδιαστολής γίνεται πολύ μικρό (float / double underflow).

Ένας περιορισμός που υπάρχει στο σύστημα είναι το συνολικό μήκος των προτάσεων και οφείλεται στο γεγονός ότι η μνήμη που χειρίζονται τα 32-bit συστήματα είναι περιορισμένη στα 3~4 περίπου GB. Εξάλλου, το σύνολο των πιθανών συμβόλων είναι εξαιρετικά μεγάλο (π.χ. για το Α' σώμα κειμένων είναι περίπου 7600, όσα δηλαδή και τα λήμματα) και δεν επιτρέπει τη δημιουργία HMM με περισσότερες από 100 καταστάσεις. Για αυτό το λόγο, αρχικά χρησιμοποιήθηκε για εκπαίδευση μόνο ένα ποσοστό των διαθέσιμων προτάσεων. Η άρση, ωστόσο, αυτού του περιορισμού επετεύχθη με την παραλλαγή του αλγόριθμου εκπαίδευσης (Rabiner L.R. 1989) (Rahimi A.), ούτως ώστε να έχει τη δυνατότητα να χειρίζεται πολλαπλές ακολουθίες. Ο αλγόριθμος των πολλαπλών ακολουθιών αρχικά χρησιμοποιήθηκε για την εκπαίδευση HMM, των οποίων η αρχιτεκτονική δεν επέτρεπε μεταβάσεις σε προηγούμενες καταστάσεις.

Σύμφωνα με αυτή την παραλλαγή, το σύνολο των διαθέσιμων προτάσεων χωρίστηκε σε υποσύνολα λιγότερων προτάσεων και, ως εκ τούτου, μικρότερου συνολικού μήκους, τα οποία, όμως επιβάλλουν ο αριθμός των βημάτων εκπαίδευσης να ισούται με το πλήθος των υποακολουθιών που εξετάζονται. Παρ' όλα αυτά, η χρήση περισσότερων βημάτων περιορίζει τη μέγιστη απαιτούμενη μνήμη και καθιστά την εκπαίδευση του μοντέλου με περισσότερα δεδομένα εφικτή. Επιπρόσθετα, ο αλγόριθμος δύναται να εκτελεστεί παράλληλα σε ένα περιβάλλον πολλών επεξεργαστών και να επωφεληθεί από τις δυνατότητες των σύγχρονων υπολογιστικών συστημάτων. Η μόνη απαίτηση για συγχρονισμό θα ήταν η ενημέρωση των τιμών του μοντέλου με το πέρας της επεξεργασίας από κάθε νήμα εκτέλεσης (thread).

Το μοντέλο HMM αποδίδει κάθε λέξη σε μία ομάδα ανάλογα με την πιθανότητα παραγωγής της. Με τον τρόπο αυτό, όμοια με το μοντέλο SOM, δημιουργείται μία μέθοδος αναπαράστασης των κειμένων στο χώρο προτύπων πριν την εφαρμογή ενός κατηγοριοποιητή τύπου MLP. Για την εκτέλεση των πειραμάτων υλοποιήθηκε ένα σύστημα σε γλώσσα C++, ενώ ιδιαίτερα σημαντικές στην υλοποίηση ήταν οι υποδείξεις του (Rahimi A.), οι οποίες εκτός των άλλων διόρθωναν κάποια λάθη στους τύπους του (Rabiner L.R. 1989).

5.3.3 Πειράματα

Η αξιολόγηση του συστήματος HMM για τη δημιουργία των χαρτών πραγματοποιήθηκε μέσω ενός συνόλου πειραμάτων με τη χρήση και των δύο σωμάτων κειμένων. Οι κυριότερες παράμετροι του συστήματος κατηγοριοποίησης είναι το πλήθος των καταστάσεων του HMM, που είναι ίσο με το πλήθος των ομάδων συν μία για το τερματικό σύμβολο «.», και το μέγεθος του κρυφού επιπέδου του κατηγοριοποιητή MLP. Για το μέγεθος του κρυφού επιπέδου επιλέχτηκαν και εδώ τιμές από 5 έως 10 νευρώνες. Επιπλέον, επιλέχθηκε ο αλγόριθμος εκπαίδευσης ανάστροφης διάδοσης RPROP, ο οποίος και προτιμήθηκε από τον LMBR, αφού παρουσίασε καλύτερα αποτελέσματα στα πειράματα του συστήματος SOM της ενότητας 4.4.1.1. Μία επιπλέον παράμετρος με κάποιο ρόλο στο σύστημα είναι η αρχική τιμή της πιθανότητας της παραμονής του συστήματος στην ίδια κατάσταση, η οποία αναφέρεται ως «πιθανότητα παραμονής» και επιλέγεται να είναι ίδια για κάθε κατάσταση πλην της ειδικής κατάστασης που «εκπέμπει» το τερματικό σύμβολο, οπότε είναι μηδέν (αφού προτάσεις χωρίς καμία λέξη δεν είναι αποδεκτές). Η «πιθανότητα παραμονής» βάσει της αρχιτεκτονικής (σχήμα 5.7) βασίζεται και στο πλήθος των καταστάσεων του HMM και πρέπει να είναι πολύ μεγαλύτερη από κάθε άλλη πιθανότητα μετάβασης.

Το μέτρο των πιθανοτήτων μετάβασης είναι αντιστρόφως ανάλογο προς το πλήθος των δυνατών μεταβάσεων, καθώς το άθροισμα όλων πρέπει να ισούται με τη μονάδα. Έτσι, σε αρχιτεκτονικές με λιγότερες καταστάσεις πρέπει η «πιθανότητα παραμονής» να είναι υψηλότερη από αντίστοιχες αρχιτεκτονικές με περισσότερες καταστάσεις, και σε κάθε περίπτωση να είναι αρκετά μεγαλύτερη από τη μέση αναμενόμενη πιθανότητα μετάβασης. Στοιχείο μελέτης αποτέλεσε επίσης η αφαίρεση λέξεων με συχνότητα μεγαλύτερη από ένα κατώφλι, μιας και αυτές θα μπορούσαν να είναι λέξεις που δεν εισάγουν κάποια επιπλέον πληροφορία σε ό,τι αφορά το περιεχόμενο των κειμένων.

Η μέση ακρίβεια κατηγοριοποίησης (μέσω της κυκλικής (leave-one-out) αξιολόγησης που χρησιμοποιήθηκε σε όλα τα πειράματα κατηγοριοποίησης) για τα δεδομένα της Βουλής (Α' σώμα κειμένων) για διάφορες παραμέτρους του συστήματος παρουσιάζεται στον πίνακα 5.1. Στον πίνακα αυτό, μαζί με τις παραμέτρους του MLP που έδωσαν τα καλύτερα αποτελέσματα, παρουσιάζεται και η απόδοση όλων των συστημάτων για το ίδιο επιλεγμένο MLP (με 8 νευρώνες στο κρυφό επίπεδο), ώστε να δοθεί ένα μέτρο σύγκρισης μεταξύ των διαφορετικών παραμέτρων των μοντέλων HMM. Οι καλύτερες επιδόσεις είναι υπογραμμισμένες.

Πλήθος ομάδων	**Κατώφλι συχνότητας αφαίρεσης λέξεων**	**Πιθανότητα παραμονής**	**RPROP *h=8***	**RPROP με μεγαλύτερη ακρίβεια**
10	**Καμία**	0,7	64,2%	64,6% *(h=10)*
10	**>1000**	0,7	68,9%	69,6% *(h=9)*
10	**>300**	0,7	68,4%	69,2% *(h=10)*
15	**Καμία**	0,6	67,5%	67,8% *(h=9)*
15	**>1000**	0,6	68,6%	69,6% *(h=6)*
15	**>300**	0,6	69,3%	69,4% *(h=9)*
20	**Καμία**	0,7	71,1%	71,8% *(h=5)*
20	**>1000**	0,7	71,4%	72,2% *(h=10)*
20	**>300**	0,7	67,3%	68,0% *(h=10)*
50	**Καμία**	0,1	68,5%	70,1% *(h=9)*
50	**>1000**	0,1	70,7%	71,2% *(h=9)*
50	**>300**	0,1	65,6%	66,9% *(h=9)*
100	**Καμία**	0,1	72,3%	72,3% *(h=8)*
100	**>1000**	0,1	72,4%	73,7% *(h=10)*
100	**>300**	0,1	69,5%	69,5% *(h=8)*

Πίνακας 5.1.Αποτελέσματα του συστήματος HMM στο Α' σώμα κειμένων (κείμενα Βουλής)

Η βέλτιστη ακρίβεια κατηγοριοποίησης που παρατηρήθηκε είναι ίση με 73,7% για 100 ομάδες, όταν λέξεις με συχνότητα μεγαλύτερη του 1000 παραλείπονται. Το αποτέλεσμα αυτό είναι άμεσα συγκρίσιμο με το καλύτερο αποτέλεσμα που επιτυγχάνει το σύστημα SOM συνδυασμένο με το *k-means* (76,5%, όπως αναφέρθηκε στην ενότητα 4.4.1). Και εδώ φαίνεται ότι η μεταβολή του μεγέθους του MLP δεν αλλάζει την ακρίβεια περισσότερο από 2%, όπως παρατηρήθηκε και στην περίπτωση του SOM. Αξιοσημείωτη είναι η απόδοση του συστήματος με 20 ομάδες (72,2%), όταν οι λέξεις με περισσότερες από 300 εμφανίσεις παραλείπονται.

Το σύστημα HMM δοκιμάστηκε και στο Β' σώμα κειμένων (κείμενα εφημερίδας). Τα καλύτερα αποτελέσματα αξιολόγησης παρατηρήθηκαν για 50 ομάδες (ακρίβεια 84,6%), όταν η «πιθανότητα παραμονής» ήταν 0,5 και δεν αφαιρέθηκαν λέξεις. Η ακρίβεια αυτή πλησιάζει τη βέλτιστη ακρίβεια κατηγοριοποίησης του συνδυασμού SOM και *k-means*, που ήταν 85,3%. Λαμβάνοντας υπόψιν ότι ο αλγόριθμος εκπαίδευσης του HMM συγκλίνει ύστερα από ένα μικρό αριθμό βημάτων και βασίζεται στον επιτυχημένο αλγόριθμο ΕΜ, το μοντέλο αυτό μπορεί να κριθεί προτιμητέο υπό κάποιες προϋποθέσεις.

Σημαντικό ρόλο στο αποτέλεσμα που επιτεύχθηκε διαδραμάτισε και η παραλλαγή του αλγόριθμου εκπαίδευσης (Baum Welch) με τις πολλαπλές ακολουθίες. Ειδικότερα η ακρίβεια του συστήματος με την παραλλαγή αυτή ήταν 77,6%, μιας και το σύστημα μπορούσε να χειριστεί μόνο ένα μικρό μέρος (14%) του συνόλου των διαθέσιμων προτάσεων. Τα αποτελέσματα αυτά στο μεγάλο σύνολο δεδομένων δείχνουν ότι το σύστημα HMM μπορεί να εφαρμοσθεί αποδοτικά σε πραγματικά προβλήματα και να αποτελέσει εναλλακτική μέθοδο του συστήματος SOM. Επιπλέον, ο τρόπος δημιουργίας του χάρτη λέξεων με το HMM δεν απαιτεί διανυσματική αναπαράσταση, η οποία και έχει μεγάλες απαιτήσεις σε μνήμη ενώ συνεπάγεται και απώλεια πληροφορίας. Η ομαδοποίηση που παρέχουν τα HMM είναι αντίστοιχη των ασαφών μεθόδων με συναρτήσεις συμμετοχής και είναι πιο παραστατική, έχοντας έτσι τη δυνατότητα να λειτουργήσει ως υπόβαθρο για τη μελέτη αμφισημιών.

5.4 Αλγόριθμος αναφοράς SVM-TF IDF ως μέτρο σύγκρισης των συστημάτων (State-of-the-Art)

Για την εκτενέστερη αξιολόγηση της απόδοσης των συστημάτων οργάνωσης κειμένων που προτάθηκαν, εξετάσθηκε η απόδοση ευρύτερα χρησιμοποιούμενων τεχνικών όπως οι κατηγοριοποιητές SVM σε συνδυασμό με διαφορετικού είδους χαρακτηριστικά (TF-IDF), τα οποία έχουν εφαρμοσθεί με επιτυχία στο παρελθόν (Diederich J., Kindermann J., Leopold E. & Paass G. 2003 και Drucker, H., Wu, D. & Vapnik, V. 1999). Σε αντιστοιχία με τα προηγούμενα πειράματα και για το σύστημα αναφοράς χρησιμοποιήθηκε η κυκλική αξιολόγηση (leave-one-out), ενώ η απόδοση του συστήματος αναφέρεται πάντα στη μέση τιμή της απόδοσης σε άγνωστα για τη φάση εκπαίδευσης δεδομένα.

5.4.1 Περιγραφή μεθόδου σύγκρισης

Το μοντέλο που χρησιμοποιήθηκε και εδώ (όπως και στην ενότητα 3.5) για την κατηγοριοποίηση των δεδομένων είναι το SVM, το οποίο παρουσιάσθηκε εκτενέστερα στην ενότητα 3.2.5. Οι είσοδοι στο μοντέλο είναι οι μετρήσεις TF-IDF, οι οποίες αποτελούν εναλλακτική των προτεινόμενων μεθόδων αναπαράστασης των κειμένων στο χώρο προτύπων (π.χ. SOM με *k-means*, HMM) και περιγράφηκαν στην ενότητα 3.5. Το μεγάλο πλήθος των λημμάτων στο σύνολο των δεδομένων επιβάλλει την επιλογή του πλήθους των χαρακτηριστικών TF-IDF που θα χρησιμοποιηθούν (σε κάθε λήμμα δύναται να αντιστοιχεί μία μέτρηση TF-IDF), αφού το πλήθος αυτό καθορίζει άμεσα το μέγεθος του διανύσματος εισόδου. Για αυτή τη διαδικασία αξιολόγησης ακολουθήθηκε μία πιο εξελιγμένη μέθοδος επιλογής των κατάλληλων TF-IDF, η οποία βασίζεται σε αρχές της θεωρίας πληροφορίας (information theory).

Πιο συγκεκριμένα, από ένα πολύ ευρύτερο σύνολο μετρήσεων λημμάτων επιλέχθηκε ένας πιο περιορισμένος αριθμός βάσει ενός κριτηρίου μέτρησης της πληροφορίας που κάκαθένα από αυτά συνεισφέρει. Το κριτήριο αυτό ονομάζεται «κέρδος πληροφορίας» (IG - Information Gain) και έχει εφαρμοσθεί με επιτυχία σε αντίστοιχα προβλήματα (Yang Y. & Pedersen J.O., 1997), ενώ δίνει καλύτερα αποτελέσματα σε σχέση με άλλα αντίστοιχου σκοπού κριτήρια.

Ειδικότερα, το «κέρδος πληροφορίας» προσπαθεί να εντοπίσει λήμματα, των οποίων η έντονη παρουσία ή η πλήρης απουσία μπορεί να αποτελέσει ισχυρή ένδειξη για το ότι ένα κείμενο ανήκει σε συγκεκριμένη θεματική κατηγορία. Το μέτρο του κριτηρίου αυτού δίνεται από τον τύπο (5-16):

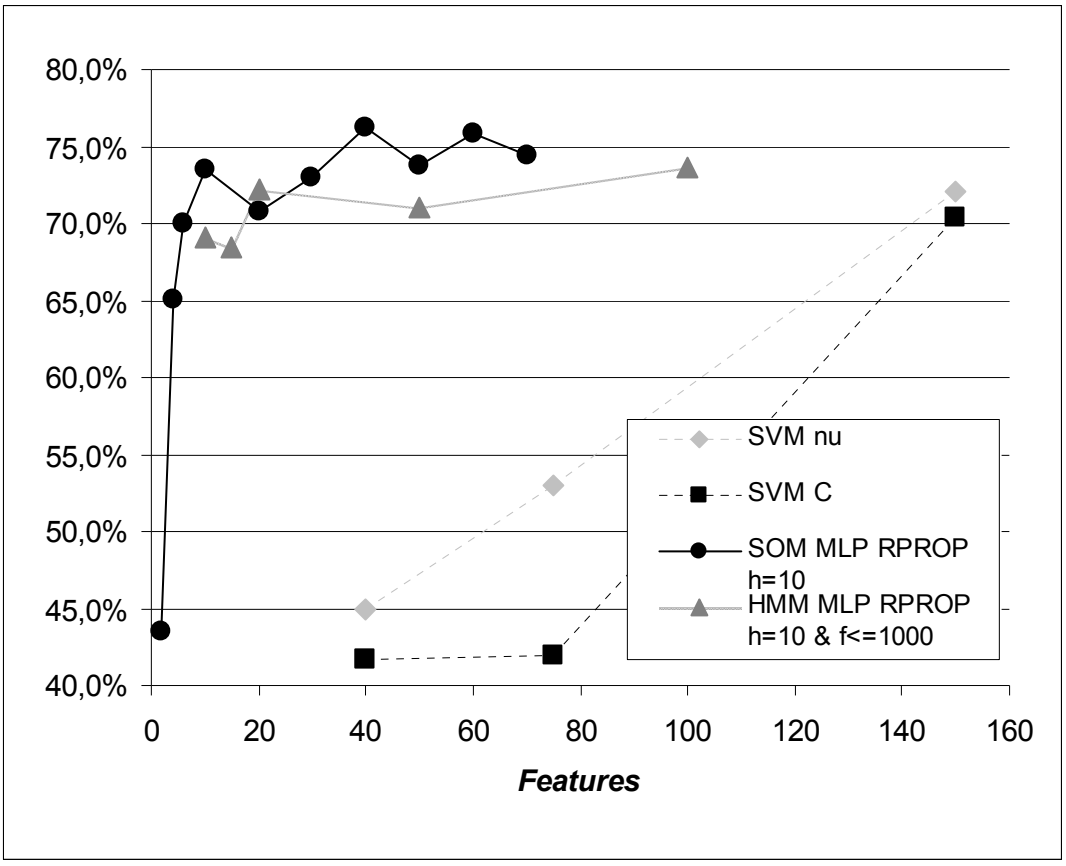
$$IG(t) = -\sum_{i=1}^{m} P(c_i) \log P(c_i) + \\ + P(t) \sum_{i=1}^{m} P(c_i \mid t) \log P(c_i \mid t) + P(\bar{t}) \sum_{i=1}^{m} P(c_i \mid \bar{t}) \log P(c_i \mid \bar{t}) \quad (5\text{-}16)$$

όπου το *t* αναφέρεται στην παρουσία ενός λήμματος (ή όρου στη γενικότερη περίπτωση) σε αντίθεση με το $\bar{t}$ που δηλώνει την απουσία του, ενώ το *m* ισούται με το πλήθος των κατηγοριών που υπάρχουν στο σύνολο δεδομένων. Ο όρος $P(c_i)$ εκφράζει την πιθανότητα ένα τυχαίο κείμενο να ανήκει στην κατηγορία c_i και ισούται με το πλήθος των κειμένων της κατηγορίας c_i ως προς το συνολικό πλήθος των κειμένων. Ο όρος $P(c_i \mid t)$ εκφράζει το ποσοστό των κειμένων που περιέχουν τον όρο *t* και ανήκουν στη δεδομένη κατηγορία c_i ως προς το πλήθος όλων των κειμένων που περιέχουν τον όρο αυτό έστω και μία φορά. Αντιστοίχως, ο όρος $P(c_i \mid \bar{t})$ εκφράζει το ποσοστό των κειμένων που **δεν** περιέχουν τον όρο *t* και ανήκουν στη δεδομένη κατηγορία c_i ως προς το πλήθος όλων των κειμένων που δεν περιέχουν τον όρο αυτό. Τέλος, οι όροι $P(t)$ και $P(\bar{t})$ υπολογίζονται από τη συχνότητα εμφάνισης του όρου *t* ή απουσίας του (για το $\bar{t}$) ως προς το συνολικό πλήθος όλων των λημμάτων που υπάρχουν στο σύνολο δεδομένων.

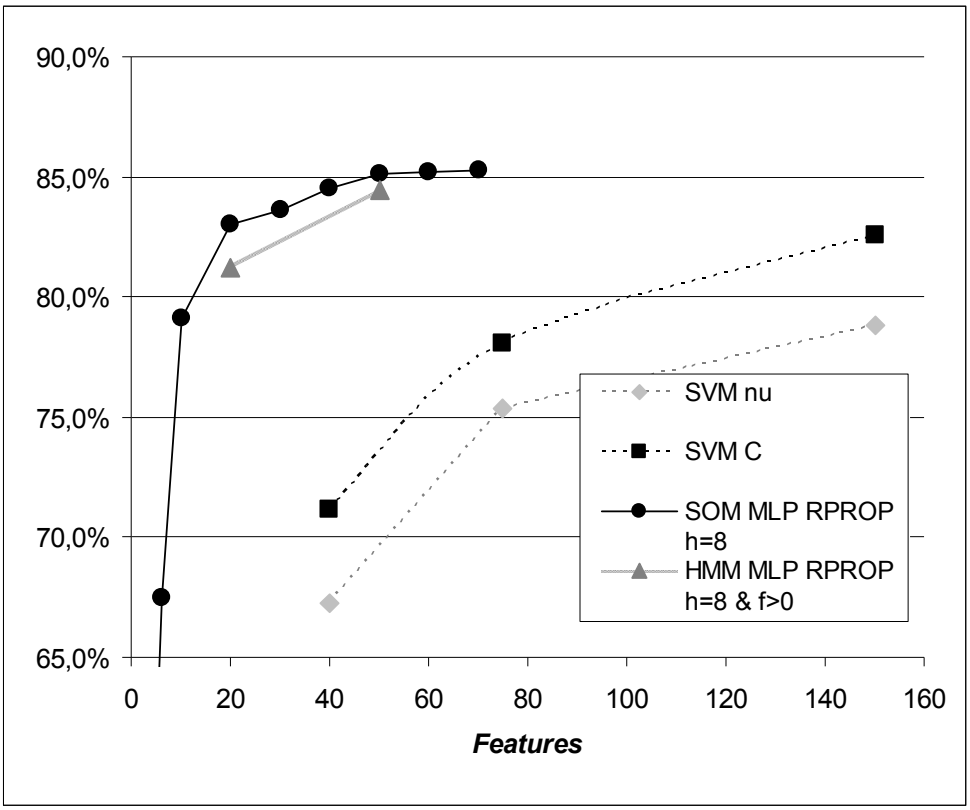
5.4.2 Αποτελέσματα σύγκρισης προτεινόμενων συστημάτων

Στην παρούσα ενότητα παρουσιάζονται τα αποτελέσματα του συστήματος αναφοράς (SVM με TF-IDF και IG) συγκριτικά με τα συστήματα που προτάθηκαν σε προηγούμενες ενότητες. Η σύγκριση περιλαμβάνει τα συστήματα που έδωσαν τα καλύτερα αποτελέσματα και βασίζονται στα μοντέλα SOM και HMM σε συνδυασμό με το MLP. Η πιο σημαντική παράμετρος του συστήματος αναφοράς είναι το πλήθος των χαρακτηριστικών TF-IDF που χρησιμοποιήθηκαν. Για την αποτελεσματικότερη σύγκριση των συστημάτων, το πλήθος των χαρακτηριστικών TF-IDF που επιλέχθηκε ήταν κοντά σε αυτό που χρησιμοποιούσαν τα συστήματα SOM (από 2 έως 70) και HMM (από 10 έως 100). Συγκεκριμένα, για το σύστημα αναφοράς ήταν 40, 75 και 150, εκ των οποίων το τελευταίο ήταν πολύ μεγαλύτερο από τα αντίστοιχα προτεινόμενα συστήματα.

Για την προσομοίωση των πειραμάτων έγινε χρήση του λογισμικού LIBSVM (Chang, C,-C. & Lin, C.-J. 2001). Χρησιμοποιήθηκε η λειτουργικότητα του πακέτου για πολυταξικά προβλήματα, ενώ επιλέχθηκαν Γκαουσσιανές συναρτήσεις πυρήνα (Gaussian Kernel). Επιπλέον, δοκιμάσθηκαν και οι δύο παραλλαγές του SVM, nu (Chen P.-H., Lin C.-J., & Schölkopf B. 2005) και C, που είναι ο αλγόριθμος εκπαίδευσης του SVM (Vapnik V. 1998), οι οποίες σχετίζονται με τον τρόπο χειρισμού των περιορισμών, ώστε να υπάρχει ανοχή σε δεδομένα εκπαίδευσης που βάσει της ομοιότητας των χαρακτηριστικών τους θα έπρεπε να ανήκουν σε άλλη τάξη (outliers). Τα πειραματικά αποτελέσματα ως προς το πλήθος των χαρακτηριστικών που κάθε κατηγοριοποιητής δέχεται ως είσοδο (άξονας x) και για τα δύο σώματα κειμένων παρουσιάζονται στα ακόλουθα σχήματα (5.8 και 5.9).



Σχήμα 5.8. Συγκριτική παρουσίαση της απόδοσης των συστημάτων για το Α' σώμα κειμένων (κείμενα Βουλής)



Σχήμα 5.9. Συγκριτική παρουσίαση της απόδοσης των συστημάτων για το Β' σώμα κειμένων (κείμενα εφημερίδων)

Η ακρίβεια κατηγοριοποίησης για το Α' σώμα κειμένων είναι σχετικά χαμηλή, αφού ξεκινάει από 45% και φτάνει το 72,1%, όταν χρησιμοποιούνται 150 χαρακτηριστικά. Αντίστοιχα, για το Β' σώμα κειμένων, η ακρίβεια κυμαίνεται από 67,2% έως 82,5%, εξακολουθώντας να είναι μικρότερη και από την ακρίβεια του SOM (85,3%), αλλά και του HMM (84,6%). Η αύξηση του πλήθους των χαρακτηριστικών TF-IDF βελτιώνει την απόδοση του συστήματος αναφοράς, ωστόσο, ακόμη και όταν τα χαρακτηριστικά αυτά είναι αρκετά περισσότερα από τα συστήματα SOM και HMM που παρουσιάσθηκαν στη μελέτη αυτή, το σύστημα αναφοράς SVM εξακολουθεί να υστερεί.

Επιπλέον, η εκπαίδευση των SVM απαιτεί ένα σύνολο παραμέτρων (αναφέρονται ως παράμετροι *a* στην ενότητα 3.2.5), οι οποίες βελτιστοποιούν κάποιο κριτήριο και το πλήθος των οποίων ισούται με το πλήθος των δεδομένων εκπαίδευσης. Αυτό έχει ως αποτέλεσμα, με την αύξηση του πλήθους των διαθέσιμων δεδομένων, να καθυστερεί πολύ περισσότερο η εκπαίδευση (βελτιστοποίηση των παραμέτρων) του SVM. Συνεπώς, η εκπαίδευση του SVM είναι μεν γρηγορότερη από εκείνη του MLP στα λιγότερα δεδομένα του Α' σώματος κειμένων, αλλά σημαντικά βραδύτερη στο μεγαλύτερο Β' σώμα. Άρα, ο αλγόριθμος εκπαίδευσης του SVM έχει πολύ μεγαλύτερο κόστος, όταν το σύνολο των δεδομέ-

νων εκπαίδευσης αυξάνει. Αντίθετα, η ταχύτητα εκπαίδευσης του MLP εξαρτάται από το πλήθος των παραμέτρων του (δηλαδή το μέγεθος του κρυφού επιπέδου) και επηρεάζεται σαφώς λιγότερο από την αύξηση των δεδομένων εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 6ο

6.1 Συμπεράσματα

Στην παρούσα εργασία προτάθηκαν συστήματα αυτόματης οργάνωσης κειμένων σε κατηγορίες, με βάση (α) το θεματικό τους περιεχόμενο και (β) το ύφος για την Ελληνική γλώσσα. Συγκεκριμένα, υλοποιήθηκαν αντίστοιχα δύο διακριτά συστήματα, τα οποία χρησιμοποιούν διαφορετικό τρόπο απεικόνισης των κειμένων στο χώρο προτύπων.

Το σύστημα αναγνώρισης ύφους βασίζεται κατά κύριο λόγο σε γλωσσικά χαρακτηριστικά, τα οποία αντικατοπτρίζουν τον ιδιαίτερο τρόπο με τον οποίο ο κάθε συγγραφέας χρησιμοποιεί το λόγο. Αντίθετα, το σύστημα αναγνώρισης θεματικών κατηγοριών βασίζεται αρχικά σε μετρήσεις λημμάτων και εν συνεχεία σε χάρτες λέξεων, οι οποίοι προβάλλουν τις νοηματικές συσχετίσεις μεταξύ λέξεων σε ένα χώρο δύο διαστάσεων στηριζόμενοι στο μοντέλο SOM. Οι χάρτες λέξεων δημιουργούνται χωρίς επίβλεψη και μπορούν να δημιουργηθούν για πολύ μεγάλο σύνολο δεδομένων χωρίς καμία ανθρώπινη παρέμβαση. Ο τρόπος επιλογής των χαρακτηριστικών για τη δημιουργία των χαρτών είναι καινοτόμος και βασίζεται στη χρήση του κανόνα ABC. Τα πειράματα που διεξήχθησαν αφορούσαν αποκλειστικά την Ελληνική γλώσσα, η οποία είναι αρκετά πιο πλούσια και περίπλοκη στο μορφολογικό επίπεδο σε σχέση με την ευρέως διαδεδομένη Αγγλική, ενώ έχει μελετηθεί ελάχιστα αναφορικά προς τέτοιου είδους προβλήματα.

Και τα δύο συστήματα αποδίδουν ικανοποιητικά δίνοντας υψηλή ακρίβεια διαχωρισμού, ενώ παρουσιάζουν βελτιωμένη απόδοση συγκρινόμενα με άλλες ευρέως διαδεδομένες τεχνικές, οι οποίες θεωρούνται και τεχνικές αναφοράς (state-of-the-art), όπως είναι ο συνδυασμός χαρακτηριστικών TF-IDF με το SVM. Ιδιαίτερα δε το σύστημα οργάνωσης βάσει περιεχομένου φαίνεται ότι μπορεί να ανταποκριθεί σε πολύ δύσκολα προβλήματα και πραγματικές συνθήκες, όπως αποδεικνύεται από την ακρίβεια που πέτυχε στο πολύ μεγάλο σώμα κειμένων προερχόμενων από αρχείο εφημερίδας που κάλυπτε πέντε έτη. Επιπλέον, με την αξιοποίηση των ιδιοτήτων των τεχνητών νευρωνικών δικτύων MLP, έγιναν προσπάθειες περαιτέρω βελτιστοποίησης των συστημάτων και αξιολόγησης της απεικόνισης που αυτά παρέχουν για τα κείμενα στο χώρο προτύπων με τη μείωση της πολυπλοκότητας τους.

Για το σύστημα αναγνώρισης περιεχομένου έγιναν προσπάθειες ενδιάμεσης αξιολόγησης του χάρτη λέξεων χρησιμοποιώντας ανεξάρτητα σύνολα λέξεων-κλειδιών αλλά και ένα θησαυρό όρων. Τα αποτελέσματα της αξιολόγησης επιβεβαίωσαν τα θετικά αποτελέσματα του συστήματος, δεν είχαν ωστόσο ικανή διακριτική ικανότητα ώστε να παρέχουν δυνατότητα επιλογής των καλύτερων παραμέτρων του συστήματος. Παράλληλα με το σύστημα που αναπτύχθηκε βάσει του μοντέλου SOM, δοκιμάσθηκαν εναλλακτικά συστήματα που εφάρμοζαν τεχνικές βελτιστοποίησης σμήνους (PSO), μίγματος Γκαουσσιανών κατανομών με δίκτυα ακτινικών συναρτήσεων (GMM/RBF) και κρυφά Μαρκοβιανά μοντέλα (HMM). Τα συστήματα PSO και HMM αποτελούν καινοτόμες προτάσεις, αφού τα μοντέλα αυτά δεν έχουν δοκιμασθεί σε προβλήματα κατηγοριοποίησης κειμένων. Τα συστήματα PSO και GMM/RBF δεν κατέληξαν σε εξίσου επιτυχημένα αποτελέσματα σε αντίθεση με το HMM, το οποίο κατέστη ικανό να αντικαταστήσει το σύστημα SOM με ελάχιστη απώλεια της ακρίβειας κατηγοριοποίησης, αλλά με κέρδος όσον αφορά στους υπολογιστικούς πόρους.

6.2 Μελλοντικές Κατευθύνσεις

Μία πιθανή μελλοντική κατεύθυνση των προτάσεων που παρουσιάστηκαν στην παρούσα εργασία θα ήταν η περαιτέρω αξιοποίηση και εφαρμογή του συστήματος σε πραγματικά προβλήματα. Συνήθως σε τέτοιες συνθήκες, εκτός από την κατηγοριοποίηση με επίβλεψη σε τάξεις, θα ήταν επιθυμητός, με χρήση των καλύτερων χαρακτηριστικών, ένας αλγόριθμος απεικόνισης δεδομένων χωρίς επίβλεψη (όπως για παράδειγμα το SOM), πάνω στον οποίο θα προβάλλονταν οι «περιοχές» που καταλαμβάνει κάθε θεματική κατηγορία κειμένων και τα σύνορά της. Επιπλέον, ανάλογα με τις απαιτήσεις η περαιτέρω έρευνα πάνω σε ιεραρχικές δομές οργάνωσης (δενδρογράμματα) θα ήταν ενδιαφέρουσα, καθώς μπορεί να παράσχει ένα γρηγορότερο και πιο ελκυστικό στο χρήστη τρόπο αναζήτησης της πληροφορίας. Πολύ σημαντική για ένα τέτοιο σύστημα είναι η ανεξάρτητη αξιολόγησή του από τους χρήστες, ώστε να εκτιμηθεί η πραγματική χρησιμότητά του. Εξίσου σημαντική θα ήταν η ενσωμάτωση των αντιδράσεων των χρηστών στο σύστημα. Εν συνεχεία, ο ενδιάμεσος χάρτης θα μπορούσε να αποτελέσει το πρώτο στάδιο ενός αυτόματου τρόπου δημιουργίας θησαυρού και θα μπορούσε να αξιοποιηθεί με κάποιες παραλλαγές στη δημιουργία συνωνύμων ή συνόλων λέξεων-κλειδιών για την περιεκτική περιγραφή των κειμενικών δεδομένων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

BayesNetToolbox, available at:

http://www.cs.ubc.ca/~murphyk/Software/BNT/bnt.html

Bell T. C., Cleary J. G., & Witten I. H., 1990. “Text Compression”. Prentice Hall.

Bilmes J. A., 1997. “A gentle tutorial on the EM algorithm and its application to parameter estimation for gaussian mixture and hidden markov models”, U.C. Berkely, TR-97-02.

Brants T. 1999, “Tagging and Parsing with Cascaded Markov Models - Automation of Corpus Annotation”. Saarbrucken Dissertations in Computational Linguistics and Language Technology, Vol. 6. German Research Center for Artificial Intelligence and Saarland University, Saarbrócken, Germany.

Cam, R. and Gilles, L. 2002. “A new model of the Pareto Effect (80:20 rule) at the brand level”. ANZMAC 2002 Conference Proceedings, pp. 1431-1436.

Chang C.-C. & Lin C.-J. 2001. “LIBSVM: a library for support vector machines”. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen P.-H., Lin C.-J., & Scholkopf B. 2005. “A tutorial on nu-support vector machines”. Applied Stochastic Models in Business and Industry, Vol. 21, pp. 111-136.

Deerwester S., Dumais S.T., Furnas G.W., Landauer T.K. & Hashman R. 1990. “Indexing by latent semantic Analysis”, Journal of the American Society for Information Science ,Vol. 41, No. 6.

Dempster A., Laird N. & Rubin D., 1977. ”Likelihood from incomplete data via the EM algorithm”, Journal of the Royal Statistical Society, Series B, Vol. 39, No. 1, pp.1–38.

Diederich J., Kindermann J., Leopold E. & Paass G. 2003. “Authorship attribution with support vector machines”. Applied Intelligence, Vol 19, pp. 109 – 123.

Dijkstra E. W., 1959. “A note on two problems in connexion with graphs”, Numerische Mathematik, Vol. 1, pp. 269–271.

Dorigo M. & Gambardella M., 1997. "Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem", IEEE Transactions on Evolutionary Computation, Vol. 1, No. 1, pp. 53–66.

Drucker, H., Wu, D. & Vapnik, V. 1999. “Support Vector Machines for Spam categorization”. IEEE Trans. on Neural Networks, Vol.10, No. 5, 1048-1054.

Duda R.O., Hart P.E. & Stork D.G. 2001. “Pattern Classification” (2nd edition), Wiley, New York.

Eurovoc, available at http://europa.eu/eurovoc/

Everitt B. S., Landau S. & Leese M. 2001. “Cluster Analysis”. Hodder Arnold Publication.

Figueiredo M.A.T. & Jain A.K., 2002. “Unsupervised learning of finite mixture models”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 24, No. 3.

Foresee F. & Hagan, T. 1997. “Gauss Newton Approximation To Bayesian Learning” Proceedings of the 1997 International Joint Conference on Neural Networks, Vol. 3, pp. 1930-1935.

Freeman R.T. & Yin. H. 2005. “Web content management by self-organization”, IEEE Transactions on Neural Networks, Vol. 16, No 5, pp. 1256-1268.

Georgakis A., Kotropoulos C., Xafopoulos A., & Pitas I., 2004. “Marginal median SOM for document organization and retrieval”, Neural Networks, Vol. 17, No. 3, pp. 365-377.

Goldberg D. E., 1989. “Genetic Algorithms in Search, Optimization and Machine Learning”, 1st edition, Addison-Wesley.

Gurney P. & Gurney L. 1998a. “Authorship Attribution οφ the Scriptores Historiae Augustae”. Literary and Linguistic Computing, Vol 13, No 3, pp. 119 – 131.

Gurney P. & Gurney L. 1998b. “Subsets and Homogeneity: Authorship Attribution in the Scriptores Historiae Augustae”. Literary and Linguistic Computing, Vol 13, No 3, pp. 133 – 140.

Hagan, M.T & Menhaj, M.B. 1994. “Training feedforward networks with the Marquardt algorithm”, IEEE Transactions on Neural Networks, Vol. 5, No. 6, pp 989-993.

Haykin S. 1999. “NEURAL NETWORKS: A Comprehensive foundation” Second Edition, Prentice Hall.

Holland, J. H. 1975. “Adaptation in Natural and Artificial Systems” Ann Arbor, MI: Univ. of Michigan Press

Hromkovic J., 2002. “Algorithmics for Hard Problems”, Springer.

Igel C. & Huesken. M., 2003. “Empirical evaluation of the improved Rprop learning algorithm”, Neurocomputing, Vol. 50, pp. 105-123.

Jain A.K., Duin R. P. & Mao J. 2000. "Statistical pattern recognition: a review" IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, no. 1, pp. 4-37.

Kaski S., 1998. “Dimensionality Reduction by Random mapping: Fast Similarity Computation for Clustering.” In Proceedings of IJCNN’98, International Joint Conference on Neural Networks, Vol. 1, pp. 413-418.

Kennedy J. & Eberhart R. C. 2001. “Swarm Intelligence”. Morgan Kaufmann.

Kohonen T. & Somervuo P., 1998. “Self-organizing maps of symbol strings”, Neurocomputing, Vol. 21, No. 1-3, pp. 19-30.

Kohonen T., 1982. “Self-Organized formation of topologically correct feature maps”, Biological Cybernetics, Vol. 43, pp. 59-69.

Kohonen T., 1985. “Median Strings”, Pattern Recognition Letters, Vol.3, pp. 309-313.

Kohonen T., 1997. “Self-Organizing Map” (2nd edition), Springer Verlag, Berlin, Germany.

Kohonen T., Kaski S., Lagus K., Salojarvi, Honkela J., Patero V. & Saarela A., 2000. “Self-Organisation of a Massive Document Collection.” IEEE Transactions on Neural Networks, Vol. 11, No. 3, pp. 574-585

Lagus K., Kaski S. & Kohonen T., 2004. “Mining Massive Document Collections by the WEBSOM Method.” Information Sciences, Vol. 163, No. 1-3, pp.135-156.

Lessing L., Dumitrescu I. & Stutzle T, 2004. “A Comparison Between ACO Algorithms for the Set Covering Problem”, Lecture Notes in Computer Science, Vol. 3172, pp. 1-12.

Likas A., Vlassis N. & Verbeek J., 2003. “The Global K-Means Clustering Algorithm” , Pattern Recognition , vol. 36, pp. 451-461.

Linden K. 2004. “Evaluation of Linguistic Features for Word Sense Disambiguation with Self-Organised Document Maps”. Computers and the Humanities, Vol. 38, pp. 417-435.

Lowe D. & Matthews R. 1995. “Shakespeare Vs. Fletcher: A Stylometric Analysis by Radial Basis Functions”. Computers and Humanities, Vol 29, pp. 449 – 461.

MacKay D. 2003. “Information Theory, Inference, and Learning Algorithms”. Cambridge University Press.

Mackay, D. 1992. “A practical Bayesian framework for backpropagation networks” Neural Computation, Vol. 4, No. 3, pp. 448–472.

Manning C. D. & Schutze H. 1999. “Foundations of Statistical Natural Language Processing”, The MIT Press, Cambridge, Massachusetts

Manning C. D. & Schutze H. 1999. “Foundations of Statistical Natural Language Processing”, The MIT Press, Cambridge, Massachusetts.

Martens D., De Backer M., Haesen R., Baesens, B., Mues C. & Vanthienen, J, 2006. “Ant-Based Approach to the Knowledge Fusion Problem”, Lecture Notes in Computer Science, Vol. 4150, pp. 84-95.

Matthews R. & Merriam T. 1993. “Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher”. Literary and Linguistic Computing, Vol 8, No 4, pp. 203 – 209.

Merriam T. & Matthews R. 1994. “Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe”. Literary and Linguistic Computing, Vol 9, No 1, pp. 1 – 6.

Murphy K., 2001. “The Bayes Net Toolbox for Matlab”, Computing Science and Statistics, Vol. 33.

Netlab, available at: http://www.ncrg.aston.ac.uk/netlab/index.php

Nguyen D. & Widrow B. 1990. “Improving the learning speed of 2-layer neural networks by choosing initial values of adaptive weights.”, Proceedings of the International Joint Conference on Neural Networks, Vol. 3, pp. 21-26.

Papageorgiou H., Prokopidis P., Giouli V. & Piperidis S. 2000. “A Unified PoS Tagging Architecture and its Application to Greek”. Second International Conference on Language Resources and Evaluation Proceedings, Athens, Greece, Vol. 3, pp. 1455 – 1462.

Pernkopf F. & Bouchaffra D., 2005. “Genetic-Based EM Algorithm for Learning Gaussian Mixture Models” , IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 27, No. 8.

Pullwitt D., 2002. “Integrating Contextual Information to Enhance SOM-based Text Document Clustering.” Neural Networks, Vol. 15, pp. 1099-1106.

Quinlan J. R. 1993. “C4.5: Programs for Machine Learning”. Morgan Kaufmann Publishers.

Rabiner L.R. 1989. "A tutorial on HMM and selected applications in speech recognition". In Proc. IEEE, Vol. 77, No. 2, pp. 257-286, Feb.

Rahimi A., “An Erratum for ‘A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition’ “, available at http://alumni.media.mit.edu/~rahimi/rabiner/rabiner-errata/rabiner-errata.html

Rauber A., Merkl D., Dittenbach M., 2002. “The Growing Hierarchical Self-Organizing Map: Exploratory Analysis of High-Dimensional Data”. IEEE Transactions on Neural Networks, Vol 13, No 6, pp. 1331 – 1341.

Riedmiller M. & Braun H., 1993. “A direct adaptive method for faster backpropagation learning: the RPROP algorithm”. Proceedings of the IEEE International Conference on Neural Networks, San Francisco, CA, pp. 586 – 591.

Rumelhart D.E., Hinton† G.E. & Williams R.J. 1986a. “Learning representations by back-propagating errors”. Nature, Vol. 323, pp. 533 - 536.

Rumelhart D.E., Hinton† G.E. & Williams R.J. 1986b. “Learning internal representations by backpropagation”. Parallel Distributed Processing, Vol.1, Chapter 8.

Salton, G. & Buckley, C. 1988. “Term-weighting approaches in automatic text retrieval”. Information Processing & Management, Vol. 24, No. 5, pp. 513–523.

Stamatatos E., Fakotakis N., & Kokkinakis G. 2001. “Computer-Based Authorship Attribution Without Lexical Measures”. Computers and the Humanities, Vol 35, pp. 193 – 214.

Stamatatos, Fakotakis Kokkinakis. 2001. “Automatic Text Categorization in Terms of Genre and Author”. Association for Computational Linguistics, Vol 26, No 4, pp. 471 – 494.

Stutzle T. & Hoos H., 2000. “MAX-MIN Ant System”, Future Generation Computer Systems, Vol. 16, pp. 889-914.

Tambouratzis G. 2006, “Assessing the effectiveness of Feature Groups in Author Recognition Tasks with the SOM Model.” IEEE Transactions on Systems, Man & Cybernetics, Vol. 36, No. 2, pp. 249-259.

Tambouratzis G., Hairetakis N., Markantonatou S. & Carayannis G. 2003. “Applying the SOM model to text classification according to register and stylistic content”. International Journal on Neural Systems, Vol. 13, No. 1, pp. 1 – 11.

Tambouratzis G., Markantonatou S., Hairetakis N. , Vassiliou M., Tambouratzis D. & Carayannis G. 2000. “Discriminating The Registers And Styles In The Modern Greek Language”. ACL 2000, Proceedings of the workshop on Comparing Corpora, pp. 34 – 43.

Tambouratzis, G., Markantonatou, S., Hairetakis, N., Vassiliou, M., Tambouratzis, D. & Carayannis, G. 2004. “Discriminating the Registers and Styles in the Modern Greek Language – Part 2: Extending the Feature Vector to Optimise Author Discrimination”. Literary & Linguistic Computing, Vol. 19, No. 2, 221-242.

Tanwari A., Lakhiar, Q.A., and Shaikh G.Y. 2000. “ABC Analysis as an Inventory Control Technique”. Research Journal of Engineering Science and Technology, Vol. 1, No 1.

Tweedie F., Singh S. & Holmes D. 1996. “An Introduction to Neural Networks in Stylometry”. Research in Humanities Computing, Vol 5, pp. 249 – 263.

Tweedie F., Singh S. & Holmes D. 1996. “Neural Networks Applications in Stylometry: The Federalist Papers”. Computers and Humanities, Vol 30, pp. 1 – 10.

Vapnik V. 1998. “Statistical Learning Theory”. Wiley Interscience, New York.

Vesanto J. & Alhoniemi E., 2000. “Clustering of the Self-Organising Map”. IEEE Transactions on Neural Networks, Vol. 11, No. 3, pp. 586-600.

Vesanto J., 2000. “Neural Network Tool for Data Mining: SOM Toolbox”. Proceedings of Symposium on Tool Environments and Development Methods for Intelligent Systems (TOOLMET2000), Finland, pp. 184 – 196.

Yang Y. and Pedersen J.O., 1997. “A comparative study on feature selection in text categorization”. In Proceedings of the Fourteenth International Conference on Machine Learning, pp. 412 – 420.

Yiming Yang 1999. “An Evaluation of Statistical Approaches to Text Categorization”. In Journals of Information Retrieval, Vol 1, No 1/2, pp. 67 – 88.

Καραγιάννης Γ., Σταϊνχάουερ Γ. 2001. “Μάθηση μηχανών και αναγνώριση προτύπων”. Σημειώσεις για το μάθημα «Αναγνώριση Προτύπων με Έμφαση στην Αναγνώριση Φωνής» ΕΜΠ, Αθήνα.

Τζαφέστας Σ. 2002. “Υπολογιστική νοημοσύνη Τόμος Α: ΜΕΘΟΔΟΛΟΓΙΕΣ”. ΕΜΠ, Αθήνα.

ΔΗΜΟΣΙΕΥΣΕΙΣ

Συνέδρια

Τσιμπουκάκης Ν., Ταμπουρατζής Γ., Καραγιάννης Γ., 2007. «Αναγνώριση συγγραφέων από κείμενα με χρήση Τεχνητών Νευρωνικών Δικτύων MLP», Πανελλήνιο Συνέδριο Φοιτητών ΗΜΜΥ

Tsimboukakis N. and Tambouratzis G., 2007. "Neural Networks for Author Attribution", FUZZ-IEEE 2007 IEEE International Conference on Fuzzy Systems

Tsimboukakis N. and Tambouratzis G., 2007. "Self-Organizing Word Map for Context-Based Document Classification", WSOM 2007 International Workshop on Self-Organizing Maps

Tsimboukakis N. and Tambouratzis G., 2008. "Document classification system based on HMM word map", CSTST 08, 5th International Conference on Soft Computing as Transdisciplinary Science and Technology, Paris, France

Περιοδικά

Tsimboukakis N. and Tambouratzis G., 2009. "A comparative study on authorship attribution classification tasks using both neural network and statistical methods", Accepted for Publication on Neural Computing and Applications, Reference Number NCA-464R1, DOI: 10.1007/s00521-009-0314-7

Tsimboukakis N. and Tambouratzis G., 2009. "Word map systems for document classification tasks based on content", under consideration for publication on IEEE Transactions on Systems, Man and Cybernetics Part C.

Βιβλία

Nikos Tsimboukakis and George Tambouratzis., 2009. "ACO Hybrid Algorithm for document classification system", Swarm Intelligence for Knowledge-Based Systems, Edited by L.C. Jain, S. Dehuri and CP Lim, Springer-Verlag, Germany

ΑΝΑΦΟΡΑ στο ΠΕΝΕΔ

Η παρούσα ερευνητική εργασία χρηματοδοτήθηκε από τη Γενική Γραμματεία Έρευνας και Τεχνολογίας στα πλαίσια του προγράμματος ΠΕΝΕΔ 03ΕΔ97.