ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ

ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Συνεχής Χωρική Παρεμβολή με Μεθόδους Μάθησης Συνόλου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Φίλιππος Σιοζόπουλος

Επιβλέπων : Ανδρέας-Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2016

ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΚΑΙ ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΤΕΧΝΟΛΟΓΙΑΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΥΠΟΛΟΓΙΣΤΩΝ

Συνεχής Χωρική Παρεμβολή με Μεθόδους Μάθησης Συνόλου

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Φίλιππος Σιοζόπουλος

Επιβλέπων : **Ανδρέας-Γεώργιος Σταφυλοπάτης**
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Ανδρέας-Γεώργιος Σταφυλοπάτης Καθηγητής Ε.Μ.Π.	Στέφανος Κόλλιας Καθηγητής Ε.Μ.Π.	Γεώργιος Στάμου Επίκουρος Καθηγητής Ε.Μ.Π.

Αθήνα, Φεβρουάριος 2016

……

Φίλιππος Σιοζόπουλος
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών Ε.Μ.Π.

Περίληψη

Ο σκοπός της διπλωματικής εργασίας ήταν αρχικά η μελέτη και έπειτα η σύγκριση μεταξύ διαφόρων τεχνικών που αφορούν το πρόβλημα της χωρικής παρεμβολής. Πιο συγκεκριμένα, μελετήθηκαν τόσο τεχνικές μηχανικής μάθησης (όπως τα πολυεπίπεδα perceptron) όσο και γεωστατιστικές τεχνικές (όπως η τεχνική Ordinary Kriging). Έμφαση δόθηκε στην μελέτη των μεθόδων μάθησης συνόλου, καθώς και στον συνδυασμό αυτών με γεωστατιστικές τεχνικές.

Τα σύνολα δεδομένων τα οποία χρησιμοποιήθηκαν προέρχονται απο τον επιστημονικό διαγωνισμό χωρικής παρεμβολής (Spatial Interpolation Comparison) τις χρονιές 1997 και 2004 (SIC97 και SIC2004). Μετά την ολοκλήρωση της αρχικής μελέτης των τεχνικών, χρησιμοποιήθηκε και ένα σύνολο δεδομένων που δημιουργήθηκε χρησιμοποιώντας δεδομένα προερχόμενα απο το επιστημονικό πρόγραμμα earthscope, το οποίο ασχολείται με την μορφή και την γεωλογική εξέλιξη της Βορειοαμερικάνικης ηπείρου.

Τα αποτελέσματα της σύγκρισης μεταξύ των διαφόρων τεχνικών επιβεβαιώνουν την υπεροχή των τεχνικών μηχανικής μάθησης όσον αφορά το πρόβλημα της χωρικής παρεμβολής στα μεγαλύτερα σύνολα δεδομένων, ενώ φαίνεται πως τα καλύτερα αποτελέσματα λαμβάνονται με τον συνδυασμό τεχνικών μάθησης συνόλου με γεωστατιστικές τεχνικές.

Λέξεις Κλειδία:

Μηχανική Μάθηση, Επιβλεπόμενη Μάθηση, Μάθηση Συνόλου, Τυχαία Δάση, Δέντρα Απόφασης, Χωρική Παρεμβολή, Kriging, Παλινδρόμηση, Γεωστατιστική, Earthscope, Spatial Inteprolation Comparison (SIC)

Abstract

The purpose of this diploma thesis was the study and comparison between several techniques concerning the problem of spatial interpolation. Specifically, the techniques studied belong to the domain of machine learning (such as multilayer perceptrons) and the domain of geostatistics (such as Ordinary Kriging). Machine Learning techniques based on Ensemble Learning, as well as the combination of these with Geostatitics, were also studied at greater detail.

The datasets used are sourced from the scientific spatial interpolation competition (SIC), which took place in 1997 and 2004 (SIC97 and SIC2004 respectively). After the completion of the first round of experiments using the aforementioned datasets, a novel dataset was created and used, drawing data sourced from the earthscope scientific program, the purpose of which is the study of the North American continent's geological development.

The results of the various experiments confirm the superiority of machine learning techniques in spatial interpolation when the relevant datasets are sufficiently large, while the combination of machine learning and geostatistics techniques achieve even better results.

Keywords:

Machine Learning, Supervised Learning, Ensemble Learning, Random Forests, Decision Trees, Spatial Interpolation, Kriging, Regression, Geostatistics, Earthscope, Spatial Inteprolation Comparison (SIC)

Δομή της Διπλωματικής Διατριβής

Στο 1ο Κεφάλαιο παρουσιάζεται αρχικά το πρόβλημα της παρεμβολής στη γενική του μορφή, καθώς και το πρόβλημα της χωρικής παρεμβολής συγκεκριμένα. Γίνεται επίσης αναφορά σε διάφορες κλασικές τεχνικές αντιμετώπισης των παραπάνω προβλημάτων. Στη συνέχεια παρουσιάζεται ο τρόπος με τον οποίο το πρόβλημα της χωρικής παρεμβολής προσαρμόζεται σε τεχνικές μηχανικής μάθησης. Τέλος, παρουσιάζονται συνοπτικά κάποιες τυπικές μέθοδοι μηχανικές μάθησης, οι οποίες χρησιμοποιούνται και στα πειράματα της διατριβής.

Στο 2ο Κεφάλαιο γίνεται μια αναλυτικότερη παρουσίαση του αλγορίθμου των δέντρων απόφασης. Αυτή η παρουσίαση στοχεύει στην καλύτερη κατανόηση του αλγορίθμου Τυχαίων Δασών (Random Forests), ο οποίος παρουσιάζεται αργότερα και χρησιμοποιεί τα δέντρα απόφασης ως βασική δομική του μονάδα.

Επίσης για την καλύτερη κατανόηση του αλγορίθμου Τυχαίων Δασών, στο 3ο Κεφάλαιο παρουσιάζεται η τεχνική της Μάθησης Συνόλου (Ensemble Learning), οι διάφορες πτυχές αυτής, καθώς και κάποιοι χαρακτηριστικοί αλγόριθμοι Μάθησης Συνόλου.

Στο 4ο Κεφάλαιο παρουσιάζεται ο αλγόριθμος Τυχαίων Δασών, όπως αυτός αναλύθηκε απο τον L. Breiman αλλά και κάποιες πληροφορίες που πλαισιώνουν την δομή αυτού του αλγορίθμου.

Το 5ο Κεφάλαιο αναλύει πιο συγκεκριμένα τα πειράματα που πραγματοποιήθηκαν, με παρουσίαση των συνόλων δεδομένων που χρησιμοποιήθηκαν καθώς και τον πρωτοκόλλων εκτίμησης και προσαρμογής των παραμέτρων αυτών.

Το 6ο Κεφάλαιο περιλαμβάνει τα αποτελέσματα των πειραμάτων στα σύνολα δεδομένων SIC καθώς και σύντομο σχολιασμό επι των αποτελεσμάτων αυτών.

Στο 7ο Κεφάλαιο παρουσιάζεται η προσπάθεια περεταίρω βελτίωσης των αποτελεσμάτων που πραγματοποιήθηκε, χρησιμοποιώντας τεχνικές απο τον τομέα της γεωστατιστικής σε συνδυασμό με συμπεράσματα των προηγούμενων πειραμάτων. Αναλύονται οι σχετικές τεχνικές που προστέθηκαν, καθώς και η δημιουργία του πρωτότυπου συνόλου δεδομένων που χρησιμοποιήθηκε, το οποίο βασίζεται σε δεδομένα από το επιστημονικό πρόγραμμα earthscope.

Το 8ο Κεφάλαιο περιλαμβάνει τα αποτελέσματα των πειραμάτων στο πρωτότυπο σύνολο δεδομένων earthscope καθώς και σύντομο σχολιασμό επι αυτών.

Τέλος, το 9ο Κεφάλαιο περιλαμβάνει πιθανές επεκτάσεις και βελτιστοποιήσεις στις τεχνικές που χρησιμοποιήθηκαν με στόχο την περεταίρω βελτίωση των αποτελεσμάτων.

Περιεχόμενα

1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζεται το πρόβλημα της παρεμβολής, καθώς και οι βασικότεροι τρόποι αντιμετώπισης του. Στη συνέχεια εξετάζεται το πρόβλημα της χωρικής παρεμβολής συγκεκριμένα, ενώ παρουσιάζονται επίσης τεχνικές που έχουν σχεδιαστεί με σκοπό την αντιμετώπιση αυτού συγκεκριμένα. Τέλος, εξετάζεται συνοπτικά και η αντιμετώπιση του προβλήματος της χωρικής παρεμβολής με χρήση τεχνικών μηχανικής μάθησης.

1.1 Παρεμβολή (Interpolation)

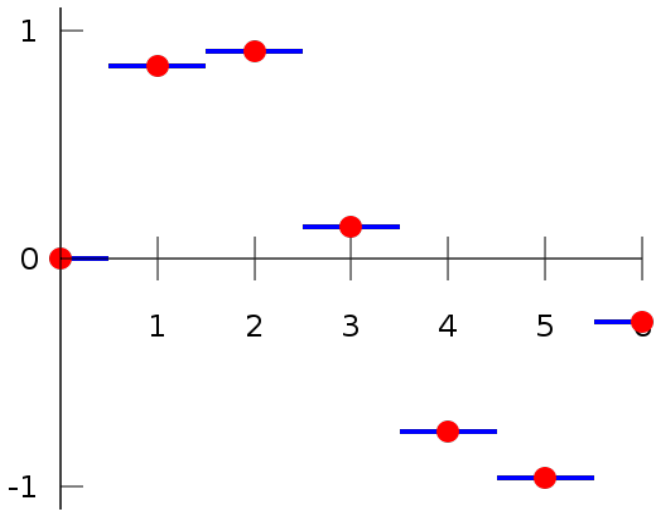
Η παρεμβολή (Interpolation) αποτελεί μια μέθοδο για τον υπολογισμό των τιμών σημείων που βρίσκονται εντός ενός συνόλου απο σημεία των οποίων οι τιμές είναι ήδη γνωστές.

Αλλιώς, δεδομένου ενός συνόλου $(x_i, g(x_i)) = (x_i, y_i) \in X\,x\,Y$ όπου X και Y είναι τα πεδία τιμών των x_i, y_i αντιστοίχως, σκοπός είναι ο υπολογισμός μιας *συνάρτησης παρεμβολής* $f(x)$ για την οποία ισχύει $f(x_i) = y_i$. Η συνάρτηση $g(x)$ που εκφράζει την σχέση μεταξύ των μεταβλητών x, y δεν θεωρείται γνώστη. Έτσι, είναι επιθυμητό η συνάρτηση παρεμβολής $f(x)$ να την προσεγγίζει κατά το βέλτιστο δυνατό τρόπο.
Η διαδικασία της παρεμβολής χρησιμοποιεί συνεπώς ένα σύνολο από διακριτές τιμές για τον υπολογισμό μιας συνεχούς συνάρτησης. Αυτή η διαδικασία αποκαλείται αλλιώς *data fitting* ή *curve fitting.*
Η παρεμβολή μπορεί να χρησιμοποιηθεί και για τον υπολογισμό των τιμών σημείων που βρίσκονται *εκτός* του συνόλου γνωστών σημείων. Έτσι, αν θεωρήσουμε οτι $x_i \in [a, b]$ για όλα τα γνωστά x_i , είναι δυνατός ο υπολογισμός της τιμής ενός $x_e \notin [a, b]$ με χρήση της συνάρτησης παρεμβολής. Ωστόσο τυπικά αυτή η διαδικασία αποκαλείται προεκβολή (Extrapolation).

Η παρεμβολή εμφανίζει ιδιαίτερο πρακτικό ενδιαφέρον σε περιπτώσεις που η πλήρης γνώση της συνάρτησης $g(x)$ είναι επιθυμητή ή απαραίτητη αλλά πρακτικά δύσκολη ή αδύνατη. Υπο αυτή την σκοπιά, το σύνολο γνωστών τιμών (x_i, y_i) μπορεί να αποτελεί τιμές που έχουν προκύψει απο δειγματοληψία (sampling) ή μέσω πειραμάτων.

Η απλούστερη μέθοδος παρεμβολής ονομάζεται Piecewise Constant Interpolation (ή Nearest-Neighbour Interpolation). Σύμφωνα με αυτή την μέθοδο, κάθε άγνωστο σημείο προς υπολογισμό λαμβάνει την τιμή του κοντινότερου σημείου απο το σύνολο γνωστών τιμών (x_i, y_i) .

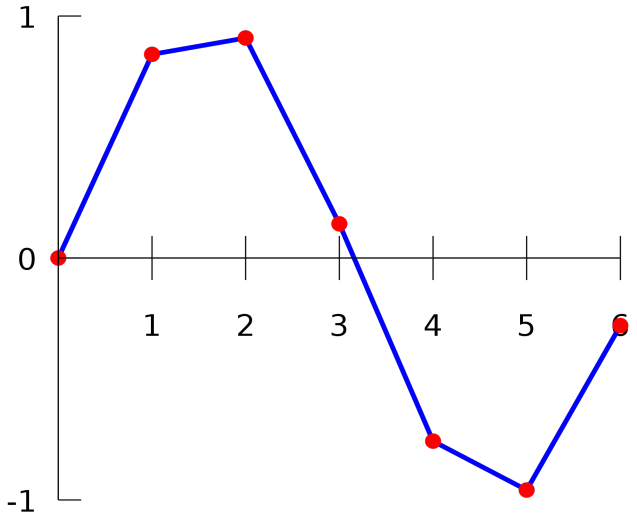
Έτσι η συνάρτηση παρεμβολής διαμορφώνεται ποιοτικά όπως φαίνεται στην παρακάτω γραφική παράσταση:



Piecewise Constant Interpolation

Οπού τα σημεία αντιπροσωπεύουν τις γνωστές τιμές.
Η εν λόγω μέθοδος δεν καταφέρνει εν γένει να προσεγγίσει την συνάρτηση $g(x)$ που θεωρητικά συνδέει τις μεταβλητές x, y . Παρ' όλα αυτά βρίσκει εφαρμογή σε περιπτώσεις όπου η διαστατικότητα του προβλήματος είναι αυξημένη, καθώς αποτελεί απλή και γρήγορη μέθοδο παρεμβολής.

Η αμέσως επόμενη (από άποψη πολυπλοκότητας) μέθοδος παρεμβολής είναι η Γραμμική Παρεμβολή (Linear Interpolation). Σύμφωνα με την γραμμική παρεμβολή, η υπολογιζόμενη συνάρτηση παρεμβολής $f(x)$ σχηματίζεται ενώνοντας τα γειτονικά σημεία του γνωστού συνόλου (x_i, y_i) με ευθύγραμμα τμήματα, ωστέ τελικά η συνάρτηση αποκτά την εξής μορφή:



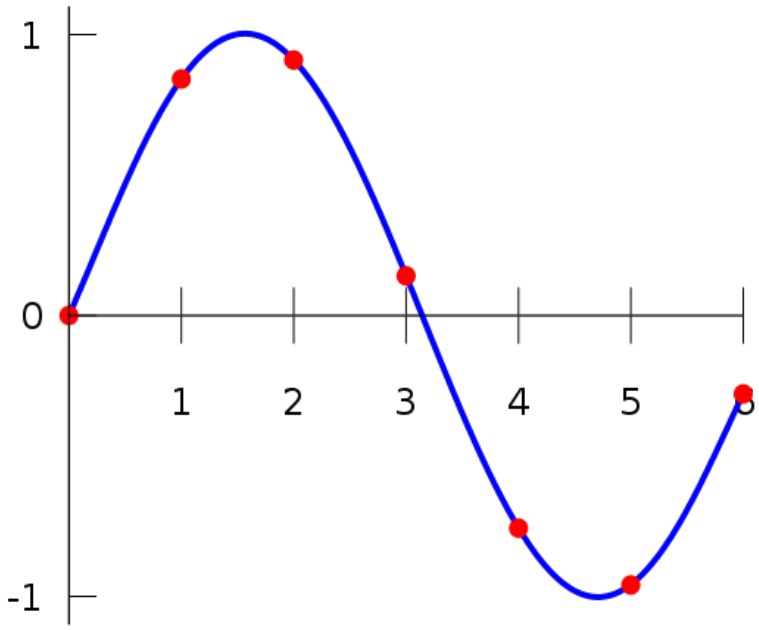
Γραμμική παρεμβολή

Μαθηματικά, αν θεωρήσουμε δύο γνωστά γειτονικά σημεία (x_a, y_a) και (x_b, y_b) , το μεταξύ τους ευθύγραμμο τμήμα προκύπτει ως $y = y_a + (y_b - y_a)\frac{x - x_a}{x_b - x_a}$. Η μέθοδος αυτή, αν και δίνει μια καλύτερη προσέγγιση από την προηγούμενη, εξακολουθεί να μην είναι ιδιαίτερα ακριβής ενώ αν και συνεχής, η υπολογιζόμενη συνάρτηση $f(x)$ δεν είναι διαφορίσιμη.

Στην προσπάθεια αντιμετώπισης της μη-διαφορισιμότητας, εφαρμόζεται η Πολυωνυμική Παρεμβολή (Polynomial Interpolation). Σε αυτή την περίπτωση η γραμμική συνάρτηση που παρουσιάστηκε στην γραμμική παρεμβολή αντικαθίσταται με ένα πολύωνυμο υψηλότερου βαθμού.

Η μέθοδος αυτή βασίζεται στο Θεώρημα Παρεμβολής σύμφωνα με το οποίο για *n+1* διαφορετικά μεταξύ τους γνωστά σημεία (x_i, y_i) υπάρχει μοναδικό πολυώνυμο βαθμού το πολύ *n* το οποίο διέρχεται απο όλα αυτά τα σημεία.

Με εφαρμογή Πολυωνυμικής παρεμβολής η συνάρτηση $f(x)$ λαμβάνει την εξής μορφή:



Πολυωνυμική παρεμβολή

Η πολυωνυμική παρεμβολή, αν και αντιμετωπίζει τα βασικά μειονεκτήματα της γραμμικής παρεμβολής, εμφανίζει κάποια αρνητικά στοιχεία σε σχέση με τις προηγούμενες μεθόδους, με το πιο χαρακτηριστικό να είναι η αυξημένη υπολογιστική πολυπλοκότητα. Για την αντιμετώπιση αυτών εφαρμόζονται παραλλαγές της βασικής πολυωνυμικής παρεμβολής όπως ενδεικτικά:

- Παρεμβολή με Splines (Spline Interpolation), οπού η $f(x)$ θεωρείται πολύκλαδη συνάρτηση κατά τρόπο αντίστοιχο με την γραμμική παρεμβολή, ωστόσο κάθε κλάδος είναι πολυώνυμο μεγαλύτερου βαθμού (και αποκαλείται Spline), ενώ εξασφαλίζεται η ομαλότητα στα σημεία αλλαγής κλάδου, σε αντίθεση με την απλή γραμμική παρεμβολή.
- Τριγωνομετρική Παρεμβολή (Trigonometric Interpolation), οπού τα πολυώνυμα αντικαθίστονται με τριγωνομετρικά πολύωνυμα. Η μέθοδος αυτή είναι ιδιαίτερα χρήσιμη στις περιπτώσεις παρεμβολής περιοδικών συναρτήσεων.

Με χρήση της πολυωνυμικής παρεμβολής καθώς και των παραλλαγών της γίνεται εφικτή η προσέγγιση αρκετά πολύπλοκων καμπύλων. Χαρακτηριστικό παράδειγμα είναι η παρεμβολή γνωστών αλλα υπολογιστικά πολύπλοκων συναρτήσεων, όπως ο φυσικός λογάριθμος. Στην συγκεκριμένη εφαρμογή υπολογίζεται πρώτα αναλυτικά η τιμή της συνάρτησης σε κατάλληλα σημεία (δειγματοληψία) κατασκευάζεται πίνακας τιμών (lookup table) και τελικά με χρήση παρεμβολής υπολογίζεται η ζητούμενη – άγνωστη τιμή χωρίς να χρειάζεται ο υπολογιστικά πολύπλοκος αναλυτικός υπολογισμός της.

1.2 Χωρική Παρεμβολή (Spatial Interpolation)

Μια ειδική κατηγορία της παρεμβολής είναι η Χωρική Παρεμβολή (Spatial Interpolation). Η χωρική παρεμβολή είναι στην ουσία παρεμβολή εφαρμοσμένη σε συναρτήσεις πολλών μεταβλητών. Έτσι, σε αυτή την περίπτωση έχουμε γνωστές τιμές της συνάρτησης σε ένα σύνολο σημείων $(x_i, y_i, z_i, ...)$ και το ζητούμενο είναι ο υπολογισμός των τιμών της συνάρτησης σε διαφορετικά σημεία $(x, y, z, ...)$.

Υπάρχει πλήθος τεχνικών που αφορούν την Χωρική Παρεμβολή, ένα μέρος των οποίων θα παρουσιαστεί αναλυτικότερα στη συνέχεια. Ωστόσο, αρχικά είναι ενδιαφέρον να σημειωθούν κάποιες γενικές διακρίσεις μεταξύ των διαφόρων τεχνικών. Έτσι, οι διάφορες τεχνικές χωρικής παρεμβολής μπορούν να χαρακτηριστούν:

- Ακριβείς (Exact) ή μη Ακριβείς (Inexact). Οι ακριβείς μέθοδοι χωρικής παρεμβολής εξασφαλίζουν πως οι προβλεπόμενες τιμές στα σημεία των οποίων η τιμή είναι γνωστή εξ' αρχής θα είναι ίσες με αυτές τις γνωστές τιμές. Αντίθετα, οι μη ακριβείς μέθοδοι εν γένει έχουν διαφορετική τιμή από την αρχικά γνωστή σε αυτά τα σημεία. Αυτή η πρακτική μπορεί να βοηθήσει στην αποφυγή απότομων κορυφών ή κοιλοτήτων στην προβλεπόμενη συνάρτηση παρεμβολής $f(x, y, z, ...)$.
- Regular Grid ή Irregular Grid. Στην πρώτη περίπτωση οι γνωστές τιμές έχουν μια προκαθορισμένη, γνωστή εξαρχής τοποθέτηση στο χώρο. Αντίθετα στη δέυτερη περίπτωση οι τιμές μπορεί να έχουν δειγματοληπτηθεί κατά τρόπο ακανόνιστο από διάφορα σημεία του χώρου.

Η Χωρική Παρεμβολή παρουσίαζεται ιδιαίτερα χρήσιμη στον τομέα της γεωστατιστικής. Σε αυτή την περίπτωση οι γνωστές τιμές προέρχονται από ενα Irregular Grid, το οποίο συνήθως ταυτίζεται με τις θέσεις σταθμών μέτρησης της σχετικής με το πρόβλημα μεταβλητής. Για παράδειγμα, τυπικές μετρήσεις αυτού του είδους είναι το ύψος της βροχόπτωσης ή η ένταση της ακτινοβολίας γάμμα σε διάφορα σημεία μιας περιοχής.
Αξίζει να σημειωθεί πως, αν και οι χωρικές συντεταγμένες (π.χ. απόκλιση ως προς τους άξονες x,y,z) αποτελούν μια πολύ λογική πρώτη προσέγγιση για την επιλογή των μεταβλητών εισόδου, σε πολλές περιπτώσεις αυτή η πληροφορία μπορεί – ή ακόμα χρειάζεται – να εμπλουτιστεί με περεταίρω πληροφορίες για την σωστή χωρική παρεμβολή και τελική προσέγγιση του υπό μελέτη φαινομένου. Για παράδειγμα, ενδεικτικές επιλογές για επέκταση των μεταβλητών εισόδου είναι η θερμοκρασία ή η υγρασία σε κάθε σημείο.
Όπως είναι λογικό, μετρήσεις αυτού του είδους είναι ιδιαίτερα δύσκολο και κοστοβόρο να γίνουν για όλη την περιοχή ενδιαφέροντος. Εντούτοις ακόμα και στις περιπτώσεις οπού υπάρχει πλήθος μετρήσεων δημιουργούνται δυσκολίες στον υπολογισμό. Αυτές οφείλονται σε ασυμφωνίες μεταξύ μετρήσεων των διαφόρων σταθμών όσον αφορά την μέθοδο με την οποία έχουν παρθεί οι σχετικές μετρήσεις ή ακόμα και την ψηφιακή αναπαράσταση αυτών. Έτσι, συχνά εμφανίζονται ετερογενή σύνολα δεδομένων τα οποία είναι αναγκαίο να συνδυαστούν με κατάλληλο τρόπο για την δημιουργία ενός αξιόπιστου χωρικού μοντέλου του υπο μελέτη φαινομένου.

Επιπλέον, η φύση των μετρήσεων αυτού του είδους συχνά προκαλεί προβλήματα σχετικά με την ακρίβεια αυτών. Έτσι, οι μετρήσεις μπορεί να έχουν παραμορφωθεί από την ύπαρξη θορύβου ή να έχουν λανθασμένα ακραίες τιμές που τελικά επηρεάζουν την ακρίβεια του μοντέλου. Για την αντιμετώπιση φαινομένων αυτού του είδους είναι συχνά απαραίτητη η αρχική μελέτη και ανάλυση του συνόλου δεδομένων (Exploratory Data Analysis – EDA), η οποία θα αναλυθεί σε επόμενο κεφάλαιο.

Ενδεικτική της σημασίας της Χωρικής Παρεμβολής είναι επίσης η ύπαρξη μιας εξειδικευμένης κατηγορίας υπολογιστικών πακέτων τα οποία ονομάζονται Geographic Information Systems (GIS), τα οποία έχουν ως σκοπό την διευκόλυνση των υπολογισμών αυτού του είδους. Στην παρούσα μελέτη δεν χρησιμοποιήθηκε κάποιο GIS.

1.3 Τεχνικές Χωρικής Παρεμβολής

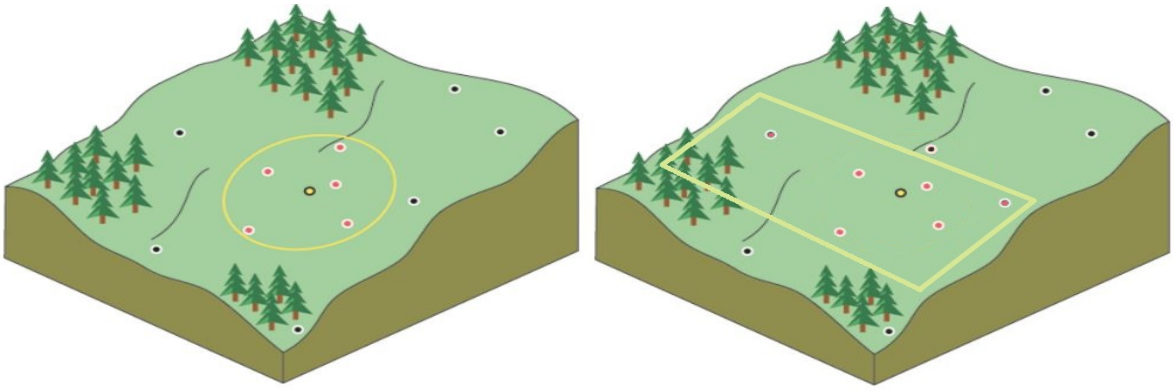
Σε αυτό το σημείο θα παρουσιαστούν κάποιες τυπικές τεχνικές χωρικής παρεμβολής που χρησιμοποιούνται. Προτού προβούμε σε αυτή την ανάλυση, αξίζει να σημειωθεί πως, παρά το γεγονός πως κάποιες απο αυτές αντιμετωπίζουν το πρόβλημα της χωρικής παρεμβολής με σχετικά απλοϊκό τρόπο ενώ άλλες με μαθηματικά πολύπλοκες τεχνικές, καμία δεν θεωρείται καλύτερη για όλες τις εφαρμογές. Στην πράξη, βασικής σημασίας είναι η αρχική ανάλυση των διαθέσιμων δεδομένων ώστε να επιλεγεί η καλύτερη για τις ανάγκες του υπο μελέτη φαινομένου τεχνική.

Local Neighbourhood Approach

Η απλούστερη ως προς την αρχική κατανόηση της λογικής είναι η προσέγγιση της τοπικής γειτονιάς (Local Neighbourhood Approach). Στην βάση αυτής της προσέγγισης βρίσκεται η υπόθεση πως καθένα από τα γνωστά σημεία επηρεάζει – η αλλιώς, φέρει πληροφορίες για – μια περιορισμένη περιοχή γύρω απο αυτό. Υπο αυτή τη σκοπιά, η προσέγγιση αυτή έχει αρκετές ομοιότητες με τις τεχνικές που παρουσιάστηκαν στο γενικό πρόβλημα της Παρεμβολής.

Η βασικότερη τεχνική τοπικής γειτονιάς είναι η παρεμβολή με στάθμιση αντίστροφης απόστασης (Inverse Distance Weighted Interpolation – IDW). Σύμφωνα με αυτή την τεχνική, η τιμή ενός άγνωστου σημείου υπολογίζεται ως ο σταθμισμένος μέσος όρος των *m* κοντινότερων γνωστών σημείων σε αυτό. Το βάρος καθενός απο αυτά τα σημεία στον υπολογισμό του μέσου όρου είναι η απόσταση αυτού απο το άγνωστο σημείο, υψωμένη σε κάποια δύναμη *p.* Τυπική τιμή αυτής της παράμετρου είναι *p=2* οπότε η τεχνική ονομάζεται Παρεμβολή Αντίστροφου Τετραγώνου Απόστασης (Inverse Distance Squared – IDS).

Η χρήση των *m* κοντινότερων σημείων για τον υπολογισμό του μέσου όρου αποτελεί μία παράμετρο της τεχνικής η οποία μπορεί να μεταβληθεί ώστε να επιτευχθεί καλύτερη ακρίβεια. Η μεταβολή αυτή μπορεί να είναι φυσικά η αυξομειώση της τιμής, εφόσον υπάρχει γνώση πως η περιοχή επιρροής κάθε γνωστού σημείου είναι μεγαλύτερη ή μικρότερη αντίστοιχα.
Εντούτοις η βασική ιδέα αυτής της παραμέτρου είναι η σκιαγράφηση της γειτονιάς επιρροής. Έτσι, στην περίπτωση που το υπό μελέτη φαινόμενο μεταβάλλεται διαφορετικά με βάση την κατεύθυνση, η γειτονιά αυτή μπορεί να λάβει αντίστοιχο σχήμα για την επίτευξη καλύτερης ακρίβειας.



Δυο διαφορετικές προσεγγίσεις ως προς την τοπική γειτονιά, για την ίδια περιοχή

Για παράδειγμα, στο παραπάνω σχήμα, μπορεί να θεωρηθεί πως μια ορθογωνική περιοχή κάθετη στην κατεύθυνση της πλαγιάς προσεγγίζει καλύτερα τη γνώση για το υπό μελέτη φαινόμενο και άρα θα δώσει καλύτερη ακρίβεια στον τελικό υπολογισμό.

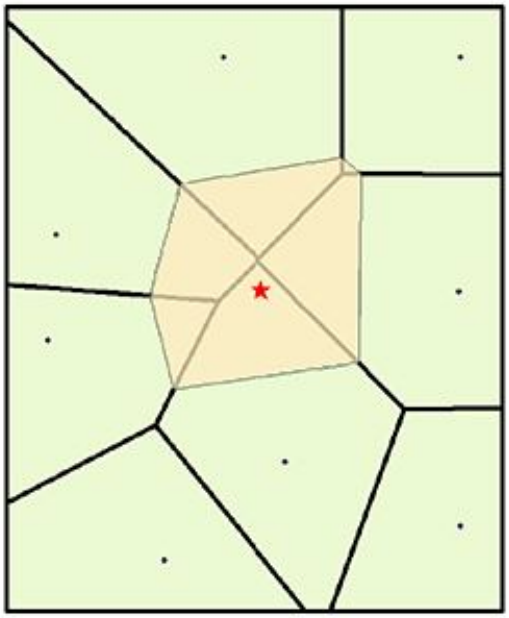
Η τεχνική IDW αποτελεί ντετερμινιστικό υπολογισμό, ενώ δεν λαμβάνει υπ' όψην την οργάνωση των γνωστών σημείων στο χώρο, περαν της απλής απόστασης αυτών από το υπολογιζόμενο σημείο. Αυτό έχει ως αποτέλεσμα την άμεση και ισχυρή συσχέτιση μεταξύ της πυκνότητας των γνωστών σημείων και το τελικό αποτέλεσμα. Επιπλέον, εφόσον η τεχνική βασίζεται στον υπολογισμό μέσου όρου, οι υπολογιζόμενες τιμές δεν μπορούν να είναι μεγαλύτερες της μέγιστης και μικρότερες της ελάχιστης γνωστής τιμής. Πρακτικά, αυτό σημαίνει πως αν δεν υπάρχουν σημεία χαρακτηριστικά των κορυφών και των κοιλοτήτων στο σύνολο γνωστών σημείων, τα αποτελέσματα μπορεί να είναι ιδιαίτερα ανακριβή.

Έτσι, αν και απλή στην κατανόηση και υλοποίηση, η τεχνική IDW είναι αρκετά ευαίσθητη στα δεδομένα εισόδου και ως αποτέλεσμα στον θόρυβο και τα σφάλματα μετρήσεων.

Μια επέκταση των τεχνικών τοπικής γειτονιάς βασίζεται στην χρήση φυσικών γειτόνων (Natural Neighbours). Σε αυτή την τεχνική αρχικά κατασκευάζεται το διάγραμμα Voronoi των γνωστών σημείων. Στη συνέχεια, για τον υπολογισμό της τιμής ενός άγνωστου σημείου, σχηματίζεται ένα μόνο πολύγωνο Voronoi για το άγνωστο σημείο.

Η τελικά υπολογιζόμενη τιμή για το άγνωστο σημείο προκύπτει ως ο σταθμισμένος μέσος όρος των γνωστών τιμών των γειτόνων αυτού, όπου το βάρος κάθε γείτονα είναι ανάλογο της επικάλυψης μεταξύ πολύγωνου Voronoi του άγνωστου σημείου και του αρχικού διαγράμματος Voronoi.

Με αυτή την παραλλαγή δίνεται μεγαλύτερη σημασία στην πυκνότητα και την τοποθεσία των γνωστών τιμών. Εξακολουθούν όμως να υφίστανται τα προβλήματα που σχετίζονται με την χρήση μέσου όρου ως τον υπολογιστικό πυρήνα.



Διάγραμμα Voronoi. Με πράσινο φαίνεται το αρχικό διάγραμμα ενώ με μπεζ το πολύγωνο της άγνωστης τιμής

Geostatistics

Οι τεχνικές που αναφέρθηκαν παραπάνω αποτελούν ντετερμινιστικές προσεγγίσεις στο πρόβλημα της χωρικής παρεμβολής.

Αντίθετα, στον τομέα της γεωστατιστικής, οι τεχνικές βασίζονται στην χρήση πιθανοτικών μοντέλων, μέσω των οποίων εκφράζεται η αβεβαιότητα σχετικά με τις υπολογιζόμενες τιμές στα άγνωστα σημεία. Σημειώνεται πως η αβεβαιότητα αυτή δεν αποτελεί εγγενές χαρακτηριστικό των υπό μελέτη φαινομένων, αλλά είναι αποτέλεσμα μη πλήρους γνώσης από την πλευρά του παρατηρητή.

Στη βάση αυτής της προσέγγισης είναι η θεώρηση πως, παρά την έλλειψη πλήρους γνώσης σχετικά με αυτό, το υπό μελέτη φαινόμενο εκφράζεται από ένα σύνολο χωρικά συσχετιζόμενων (spatially correlated) τυχαίων μεταβλητών. Έτσι, λαμβάνεται υπ' όψην και το γεγονός πως σημεία τα οποία βρίσκονται κοντά στο χώρο έχουν μεγαλύτερη πιθανότητα να έχουν και κοντινές τιμές. Με άλλα λόγια, η τυχαιότητα που εισάγεται δεν υπονοεί και ανεξαρτησία μεταξύ των διαφόρων άγνωστων σημείων.

Τυπικά, αν ορίσουμε ως $Z(x)$ την τιμή της μεταβλητής που αφορά το υπό μελέτη φαινόμενο στη θέση x, η γεωστατιστική προσέγγιση ορίζει πως, αν και η $Z(x)$ αποτελεί τυχαία μεταβλητή, η πιθανή τιμή αυτής περιορίζεται με βάση την γνώση που διαθέτουμε για το φαινόμενο. Ο περιορισμός αυτός μπορεί για παράδειγμα να εκφράζεται από την ύπαρξη γνωστής τιμής $Z(x_{known})$ σε σημείο x_{known} πολύ κοντινό της θέσης x, γνώση η οποία ενισχύει την υπόθεση πως η τιμή $Z(x)$ δεν θα απέχει από την γνωστή τιμή $Z(x_{known})$. Ωστόσο, σε αντίθεση με τις προηγούμενες ντετερμινιστικές τεχνικές, αυτή η υπόθεση απλά περιορίζει την τυχαιότητα της $Z(x)$ χωρίς να την ορίζει μονοσήμαντα.

Η χωρική συσχέτιση της τυχαίας μεταβλητής $Z(x)$ περιγράφεται εν γένει με μοντέλα χωρικής συνέχειας (spatial continuity models), όπως το variogram, το οποίο θα αναλυθεί στη συνέχεια. Αυτά τα μοντέλα δίνουν την δυνατότητα να συμπεριληφθούν πιο πολύπλοκες πληροφορίες σχετικά με την χωρική συμπεριφορά του υπό μελέτη φαινομένου στον υπολογισμό.

Variogram

Προτού προβούμε σε ανάλυση τεχνικών της γεωστατιστικής, κρίνεται σκόπιμο να παρουσιαστεί το μοντέλο variogram, καθώς αποτελεί βασικό εργαλείο στην πλειονότητα των τεχνικών αυτών.

Το variogram αποτελεί μια συνάρτηση – μοντέλο που περιγράφει τον βαθμό χωρικής εξάρτησης ενός τυχαίου πεδίου η στοχαστικής συνάρτησης.

Ας θεωρήσουμε δύο σημεία με διανύσματα θέσης u, $u+h$ και τιμές $Z(u)$, $Z(u+h)$ αντίστοιχα. Το διάνυσμα h αποκαλείται lag vector.

Ο τυπικός ορισμός του variogram με πιθανότικους όρους είναι ο εξής:

$$2\gamma(h)=E\{[Z(u)-Z(u+h)]^2\}$$

Ενώ μπορεί να εκφραστεί και ως:

$$2\gamma(h)=\frac{1}{N(h)}\sum_{N(h)}[Z(u)-Z(u+h)]^2$$

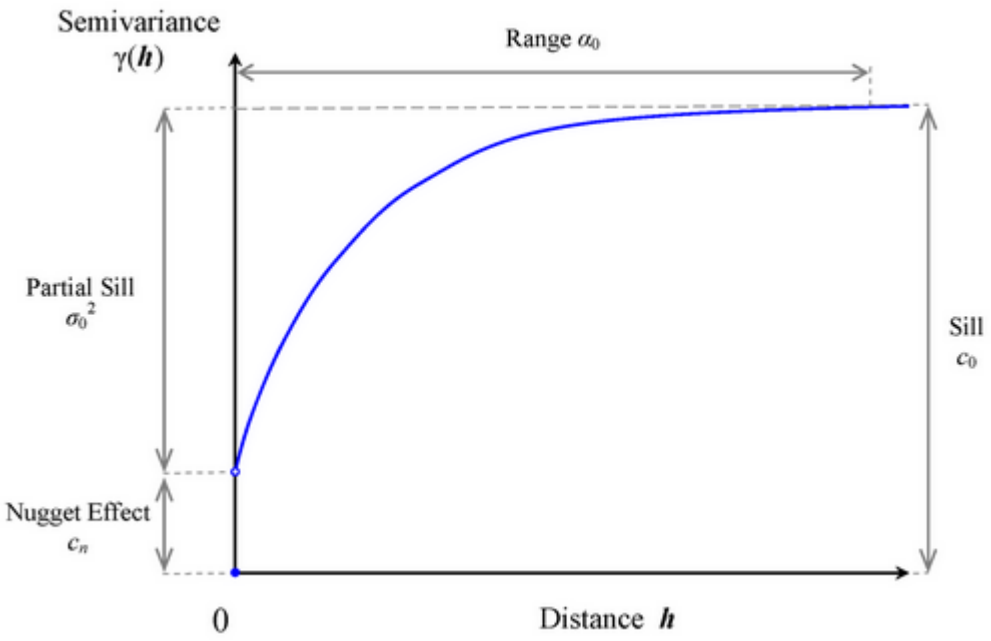
Έτσι, το variogram υπολογίζει τον μέσο όρο της διαφοράς μεταξύ όλων των τιμών που απέχουν μεταξύ τους κατα h. Με αυτό τον τρόπο, το variogram αποτελεί έναν αριθμητικό τρόπο έκφρασης της χωρικής συσχέτισης μεταξύ των διαφόρων γνωστών σημείων $(u,Z(u))$.

Η “αυστηρότητα” της μεταβλητής *h* στην παραπάνω συνάρτηση αποτελεί ενα πρόβλημα που εμφανίζεται πρακτικά κατά τον υπολογισμό του variogram. Αυτό συμβαίνει διότι στην πράξη, δεδομένου του ότι χρησιμοποιούνται δεδομένα τα οποία έχουν ληφθεί κατά τρόπο ακανόνιστο (Irregular Grid), συγκεκριμένες τιμές *h* συνήθως συνδέεουν το πολύ ένα ζευγάρι σημείων.

Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιούνται διαστήματα αποστάσεων (“bins”) κατά τον υπολογισμό αντί για αυστηρές τιμές *h.* Κατ' αυτό τον τρόπο ομαδοποιούνται ζεύγη σημείων με παραπλήσια απόσταση.

Το σύνολο αυτών των υπολογισμών αποτελεί το Εμπειρικό Variogram (Empirical Variogram), με βάση το οποίο σκιαγραφείται το τελικά χρησιμοποιούμενο Variogram, συνήθως με την μέθοδο ελαχίστων τετραγώνων.

Η μορφή ενός τυπικού Variogram φαίνεται παρακάτω:



Τυπική μορφή εκθετικού variogram

Πρόκειται για την περίπτωση του Εκθετικού (Exponential) Variogram, σύμφωνα με το οποίο η χωρική συσχέτιση μεταξύ των σημείων μειώνεται εκθετικά με την απόσταση (υπενθυμίζεται οτι το variogram σκιαγραφεί τον *μέσο όρο της διαφοράς* μεταξύ των τιμών κάθε ζεύγους σημείων).

Άλλα τυπικά μοντέλα variogram είναι το Γραμμικό (Linear) ή το Σφαρικό (Spherical), καθένα απο τα οποία προσδίδει διαφορετική πληροφορία για την χωρική συσχέτιση μεταξύ των σημείων.

Βασικές πληροφορίες που προκύπτουν απο ένα Variogram είναι τα Nugget, Sill και Range.

- Nugget ή αλλιώς Nugget Effect: Όπως φαίνεται απο τον τυπικό ορισμό του Variogram, για *h = 0* προκύπτει $\gamma(0)=0$. Ωστόσο, για πολύ μικρές τιμές *h* συχνά παρατηρείται ένα άλμα ασυνέχειας, όπως φαίνεται και στην παραπάνω γραφική παράσταση. Η ύπαρξη αυτού του σφάλματος συχνά αποδίδεται στην ύπαρξη θορύβου ή σφαλμάτων κατά τις μετρήσεις.
- Sill: Το όριο του Variogram για αποστάσεις που τείνουν στο άπειρο. Η τιμή αυτή αντιπροσωπεύει την διακύμανση (variance) της εξεταζόμενης μεταβλητής.
- Range: Η ελάχιστη απόσταση στην οποία εμφανίζεται η τιμή Sill. Η τιμή αυτή αντιπροσωπεύει την απόσταση έπειτα από την οποία ένα ζεύγος σημείων δεν θεωρούνται πλέον συσχετιζόμενα (correlated).

Kriging

Έχοντας εξετάσει το Variogram, μπορούμε πλέον να παρουσιάσουμε την τεχνική Kriging, που αποτελεί ιδιαίτερα διαδεδομένο εργαλείο γεωστατιστικής, και στον πυρήνα του οποίου χρησιμοποιείται το Variogram.

Η βασική ιδέα του Kriging είναι αντίστοιχη αυτής των ντετερμινιστικών τεχνικών τοπικής γειτονιάς που παρουσιάστηκαν προηγουμένως: Οι άγνωστες τιμές προκύπτουν ως σταθμισμένος μέσος όρος γνωστών τιμών σε μια περιοχή – γειτονιά γύρω από το άγνωστο σημείο. Ωστόσο, όπως αναφέρθηκε στην γενική περιγραφή των τεχνικών γεωστατιστικής, σε αυτή την περίπτωση εισάγεται και ένας βαθμός αβεβαιότητας στον υπολογισμό.

Για να επιτευχθεί αυτό, οι προς υπολογισμό τιμές μοντελοποιούνται με χρήση Gaussian Process. Έτσι ορίζεται πως, αν και τυχαία μεταβλητή, η προς υπολογισμό τιμή σε κάθε άγνωστο σημείο ακολουθεί κανονική κατανομή ενώ κάθε πεπερασμένο σύνολο αυτών των τυχαίων μεταβλητών αποτελεί Πολυμεταβλητή Κανονική Κατανομή (Multivariate Normal Distribution).

Αυτή η υπόθεση στην ουσία εκφράζει την λογική πως ενώ δεν γνωρίζουμε πλήρως τις άγνωστες τιμές (τυχαίες μεταβλητές) γνωρίζουμε οτι υπάρχει χωρική συσχέτιση μεταξύ τους (πολυμεταβλητή κανονική κατανομή) ενώ υποθέτουμε πως δεν απέχουν πολύ απο τις γνωστές τιμές (τυχαίες μεταβλητές κανονικής κατανομής). Στη συνέχεια, παρουσιάζονται οι σχέσεις και η θεωρία πίσω απο τεχνική του Kriging, ακολουθώντας τη μεθοδολογία και τον συμβολισμό του Deutsch. [1]

Η βασική σχέση που περιγράφει όλες τις παραλλαγές Kriging είναι η εξής:

$$\hat{Z}(u)-m(u)=\sum_{\alpha=1}^{n(u)} \lambda_\alpha [Z(u_\alpha)-m(u_\alpha)]$$

Οπού

- $\hat{Z}(u)$ είναι η τιμή προς πρόβλεψη,
- u, u_α είναι τα διανύσματα θέσης της τιμής προς πρόβλεψης και των γειτονικών γνωστών σημείων αντίστοιχα
- $n(u)$ είναι ο αριθμός των γνωστών σημείων που ανήκουν στην τοπική γειτονία που χρησιμοποείται για τον υπολογισμό της άγνωστης τιμής
- $m(u), m(u_\alpha)$ είναι οι αναμενόμενες τιμές των $Z(u), Z(u_\alpha)$ αντίστοιχα
- $\lambda_\alpha(u)$ είναι τα βάρη που ορίζονται με βάση την διαδικασία Kriging για κάθε γνωστό σημείο της τοπικής γειτονιάς προς υπολογισμό της άγνωστης τιμής στην θέση *u.* Τα βάρη αυτά θα είναι διαφορετικά για τον υπολογισμό άλλου σημείου *u'.*

Ο στόχος είναι ο προσδιορισμός των βαρών λ_α τα οποία ελαχιστοποιούν την διακύμανση $\sigma_E^2(u)=Var\{\hat{Z}(u)-Z(u)\}$ υπό τον περιορισμό $E\{\hat{Z}(u)-Z(u)\}=0$ ο οποίος εξασφαλίζει οτι η τεχνική Kriging είναι ακριβής (Exact).

Για τον προσδιορισμό των βαρών λ_α τελικά χρησιμοποιείται η συνδιακύμανση $C(h)$ η οποία ορίζεται ως $C(h)=E\{Y(u)\cdot Y(u+h)\}$, οπού $Y(u)=Z(u)-m(u)$. Σε αυτό το σημείο εμφανίζεται και η χρησιμότητα του Variogram, το οποίο επιτρέπει τον εύκολο υπολογισμό της τιμής $C(h)$ ως $C(h)=C(0)-\gamma(h)$.

Οι παραλλαγές στην μέθοδο Kriging ασχολούνται κυρίως με την παράμετρο $m(u)$ της παραπάνω βασικής σχέσης, η οποία μπορεί να θεωρηθεί σταθερή και γνωστή, τοπικά μεταβαλλόμενη ή και σταθερή αλλά άγνωστη.

Η απλούστερη παραλλαγή ονομάζεται Απλό Kriging (Simple Kriging - SK) και ορίζει πως η

παράμετρος $m(u)$ είναι σταθερή και γνωστή, δηλαδή $m(u)=m$. Η παραπάνω γενική σχέση λοιπόν ορίζει πλέον:

$$\hat{Z}(u)=m+\sum_{\alpha=1}^{n(u)} \lambda_\alpha [Z(u_\alpha)-m]$$

Αυτή η παραλλαγή απλοποιεί τους υπολογισμούς, ωστόσο πολλές φορές η υπόθεση πως η μέση τιμή m θα είναι σταθερή αλλά και γνωστή είναι μη ρεαλιστική.

Η αμέσως πιο πολύπλοκη υπολογιστικά τεχνική ονομάζεται Κανονικό Kriging (Ordinary Kriging - OK) και ορίζει πως η παράμετρος $m(u)$ είναι σταθερή τοπικά αλλά άγνωστη. Είναι δηλαδή:

$$\hat{Z}(u)=m(u)+\sum_{\alpha=1}^{n(u)} \lambda_\alpha(u)[Z(u_\alpha)-m(u)]$$
$$\hat{Z}(u)=\sum_{\alpha=1}^{n(u)} \lambda_\alpha(u) Z(u_\alpha)+[1-\sum_{\alpha=1}^{n(u)} \lambda_\alpha(u)]m(u)$$

Σε αυτή την περίπτωση, για την εξασφάλιση του περιορισμού $E\{\hat{Z}(u)-Z(u)\}=0$ προκύπτει πως τα προς υπολογισμό βάρη θα πρέπει να έχουν άθροισμα 1. Έτσι τελικά η σχέση απλοποιείται ως:

$$\hat{Z}(u)=\sum_{\alpha=1}^{n(u)} \lambda_\alpha(u) Z(u_\alpha) \quad \text{όπου} \quad \sum_{\alpha=1}^{n(u)} \lambda_\alpha(u)=1$$

Σε σχέση με το SK, το OK προσαρμόζεται καλύτερα σε τοπικά συσχετισμένες συμπεριφορές του υπο μελέτη φαινομένου, καθώς δεν περιορίζεται από την υπόθεση σταθερού και γνωστού μέσου.

Αυτή η παραλλαγή χρησιμοποείται αρκετά συχνά στην πράξη λόγω της απλότητάς της σε σχέση με άλλες τεχνικές. Επιπλέον, υπάρχει και μεθοδολογία συνδυασμού αυτής με άλλες τεχνικές, ένα ζήτημα το οποίο θα εξεταστεί σε επόμενο κεφάλαιο.

Άλλες παραλλαγές του Kriging είναι επιγραμματικά:

- Trend Kriging: Η περίπτωση οπού η παράμετρος $m(u)$ θεωρείται άγνωστη και μη σταθερή τοπικά.
- Cokriging: Εκτός απο την παραπάνω μεθοδολογία, λαμβάνονται υπ' όψην και μία ή περισσότερες δευτερεύουσες μεταβλητές για τις οποίες θεωρείται οτι υπάρχει συσχέτιση με το υπό μελέτη φαινόμενο.
- Indicator Kriging: Παραλλαγή του Kriging στην οποία οι προς αναζήτηση τιμές αντιπροσωπεύουν κατηγορίες και άρα λαμβάνουν διακριτές τιμές. Φυσικά, στην πράξη μπορεί να αφορά συνεχείς τιμές οι οποίες έχουν οργανωθεί σε διακριτές κατηγορίες.

Τέλος, κρίνεται σκόπιμο να αναφερθούν τόσο κάποια γενικά πλεονεκτήματα όσο και μειονεκτήματα της τεχνικής Kriging, ανεξάρτητα από την παραλλαγή που εφαρμόζεται. [2]

Μειονεκτήματα: Τα παρακάτω μειονεκτήματα της τεχνικής Kriging πηγάζουν από το γεγονός πως, ανεξάρτητα από την εμφάνιση της τυχαιότητας και των πολυπλοκότερων τεχνικών για την μελέτη χωρικής συσχέτισης, η τεχνική Kriging αποτελέι τεχνική χωρικής παρεμβολής. Έτσι:

- Αν τα γνωστά σημεία είναι αρκετά πυκνά και έχουν δειγματοληπτηθεί με ομοιόμορφο τρόπο, θα επιτευχθεί καλή ακρίβεια ανεξάρτητα της μεθόδου παρεμβολής που χρησιμοποιείται, δηλαδή ακόμα και με τις απλούστερες τεχνικές που παρουσιάστηκαν προηγουμένως.
- Αντίθετα, αν η δειγματοληψία δεν έχει γίνει με επαρκή τρόπο, το Kriging δεν θα καταφέρει αισθητά καλύτερη ακρίβεια απο τις άλλες μεθόδους παρεμβολής.
- Ως μέθοδος που βασίζεται στον σταθμισμένο μέσο όρο για τον πυρήνα των υπολογισμών της, παραμένει το πρόβλημα ανακρίβειας στις περιπτώσεις κορυφών και κοιλοτήτων, εφόσον ο μέσος όρος αδυνατεί να υπολογιστεί μεγαλύτερος από την μέγιστη και μικρότερος από την ελάχιστη τιμή αντίστοιχα.

Πλεονεκτήματα του Kriging:

- Η ιδιαίτερα πολύπλοκη επιλογή των βαρών για τον σταθμισμένο μέσο όρο, η οποία χρησιμοποιεί πληροφορίες σχετικές με την χωρική συσχέτιση μεταξύ των σημείων, βοηθάει στις περιπτώσεις οπού εμφανίζεται συσσώρευση γνωστών σημείων (data clustering).
- Ο υπολογισμός της διακύμανσης $\sigma_E^2(u)=Var\{\hat{Z}(u)-Z(u)\}$ (η ελαχιστοποίηση της οποίας αποτελεί τον βασικό στόχο για την επιλογή των βαρών λ_α) δίνει μια εικόνα του πιθανού σφάλματος της τελικά υπολογιζόμενης τιμής.

1.4 Χρησιμοποιώντας Μηχανική Μάθηση για Χωρική Παρεμβολή

Γιατί Μηχανική Μάθηση;

Οι τεχνικές Μηχανικής Μάθησης (Machine Learning) βασίζονται στην χρήση δεδομένων για την προσέγγιση της λύσης σε κάποιο πρόβλημα, ακολουθώντας μια τυπική διαδικασία μάθησης. Έτσι, η Μηχανική Μάθηση μπορεί να θεωρηθεί τομέας τόσο της στατιστικής όσο και της επιστήμης των υπολογιστών.

Έχοντας λοιπόν υπ' όψην την στατιστική θεώρηση των τεχνικών Μηχανικής Μάθησης, η χρήση τους για την αντιμετώπιση του προβλήματος της Χωρικής Παρεμβολής φαίνεται αρκετά λογική, δεδομένου του ότι όλες οι τεχνικές που έχουν αναλυθεί ως τώρα στην παρούσα εργασία βασίζονται στην στατιστική προσέγγιση. Πράγματι, τόσο οι τεχνικές Μηχανικής Μάθησης όσο και οι τεχνικές Γεωστατιστικής θεωρούνται προσεγγίσεις βασιζόμενες περισσότερο στα διαθέσιμα δεδομένα (“data-driven” approach) παρά σε κάποιο φυσικό μοντέλο.

Όσον αφορά τη σκοπιά της επιστήμης των υπολογιστών, οι τεχνικές μηχανικής μάθησης έχουν ως στόχο την επίλυση του προβλήματος μάθησης. Το πρόβλημα αυτό εκφράζει την επιθυμία εξόρυξης πληροφοριών απο ένα πεπερασμένο σύνολο δεδομένων, το οποίο αποκαλείται σύνολο εκπαίδευσης (training set). [3]

Διαφορές μεταξύ Μηχανικής Μάθησης και Γεωστατιστικής

Παρά την αρχική ομοιότητα που αφορά την στατιστική προσέγγιση τόσο στη Μηχανική Μάθηση όσο και την Γεωστατιστική, οι δύο τομείς έχουν ορισμένες βασικές διαφορές.

Ίσως η βασικότερη διαφορά, από την οποία πηγάζουν πολλές από τις υπόλοιπες, είναι το γεγονός πως η Μηχανική Μάθηση δεν περιορίζεται στην επίλυση ενός συγκεκριμένου προβλήματος, σε αντίθεση με την Γεωστατιστική, η οποία αποτελεί τομέα που έχει αναπτυχθεί συγκεκριμένα για την επίλυση του προβλήματος της χωρικής παρεμβολής. Αυτή η διαφορά ασκεί μεγάλη επιρροή, για παράδειγμα, στη διάσταση των δεδομένων εισόδου, τα οποία δεν είναι ασυνήθιστο να έχουν χιλιάδες διαστάσεις στην περίπτωση της Μηχανικής Μάθησης. [3]

Ως αποτέλεσμα της γενικής σκοπιάς των τεχνικών Μηχανικής Μάθησης, η τρόπος με τον οποίο αντιμετωπίζονται τα άγνωστα δεδομένα είναι αντίστοιχα πιο γενικός. Έτσι, ενώ όπως αναφέρθηκε στην Γεωστατιστική ακολουθείται μια θεώρηση πολλών ξεχωριστών τυχαίων μεταβλητών, οι οποίες όμως έχουν κάποια χωρική συσχέτιση μεταξύ τους, στην περίπτωση της Μηχανικής Μάθησης θεωρείται πως τα δεδομένα προέρχονται απο μια μοναδική τυχαία διαδικασία. Η διαδικασία μάθησης προσπαθεί να μοντελοποιήσει τελικά την κατανομή αυτή. [3]

Μια ακόμα βασική διαφορά είναι η επονομαζόμενη λογική “μαύρου κουτιού” (black box) που ακολουθείται στις τεχνικές Μηχανικές Μάθησης. Έτσι, στην πλειονότητα των τεχνικών Μηχανικής Μάθησης, αν και το πρόβλημα επιλύεται με ικανοποιητική ακρίβεια, το μοντέλο το οποίο δημιουργείται μέσω της διαδικασίας μάθησης και συνδέει την είσοδο με την έξοδο είναι δύσκολο ή αδύνατο να ερμηνευτεί.

Ως αποτέλεσμα, εμφανίζεται δυσκολία στην επιλογή των παραμέτρων (καθώς είναι δύσκολο να ερμηνευτούν πρακτικά). Πράγματι, στην πράξη οι παράμετροι επιλέγονται συνήθως μέσω πειραματισμού. Επιπλεόν, η διαδικασία μάθησης χρειάζεται εν γένει

περισσότερα δεδομένα απ' ο,τι στην περίπτωση της Γεωστατιστικής. Σημειώνεται ωστόσο οτι κάποια απο αυτά τα προβλήματα μπορούν να αντιμετωπιστούν εν μέρει εφόσον υπάρχει εκ των προτέρων γνώση για το πρόβλημα που αντιμετωπίζεται.[3]

Τέλος, στην σύγκριση μεταξύ Μηχανικής Μάθησης και Γεωστατιστικής κρίνεται σκόπιμο να αναφερθεί η διαφορά μεταξύ του όρου “μοντέλο” (model) μεταξύ αυτών.

Έτσι, στην Γεωστατιστική ο όρος αναφέρεται συνήθως σε φυσικό μοντέλο (physical model), το οποίο σκιαγραφεί την χωρική συμπεριφορά του υπό μελέτη φαινομένου και μπορεί να ερμηνευτεί από έναν παρατηρητή.

Αντίθετα, οι περισσότερες τεχνικές Μηχανικής Μάθησης θεωρούνται model-free υπό την έννοια ότι το μοντέλο Μηχανικής Μάθησης που δημιουργείται σκιαγραφέι τον υπολογισμό που συνδέει την είσοδο με την έξοδο, σύμφωνα πάντα με το σύνολο εκπαίδευσης που χρησιμοποιήθηκε. Το μοντέλο Μηχανικής Μάθησης ακολουθεί (όπως αναφέρθηκε) την λογική “μαύρου κουτιού”, δηλαδή δεν δίνει τη δυνατότητα εποπτικής κατανόησης του φαινομένου. [4]

Μηχανική Μάθηση στην Χωρική Παρεμβολή - Λεπτομέρειες

Όπως αναφέρθηκε, η Μηχανική Μάθηση αντιμετωπίζει έναν μεγάλο αριθμό προβλημάτων. Ως εκ τούτου, αρχικά αναλύεται ο τρόπος με τον οποίο οι γενικές αρχές της Μηχανικής Μάθησης τελικά αντικατοπτρίζονται στο πρόβλημα της χωρικής παρεμβολής.

Τρόπος Μάθησης

Ανάλογα με τον τρόπο μάθησης, οι αλγόριθμοι Μηχανικής Μάθησης χωρίζονται σε τρεις ευρείες κατηγορίες:

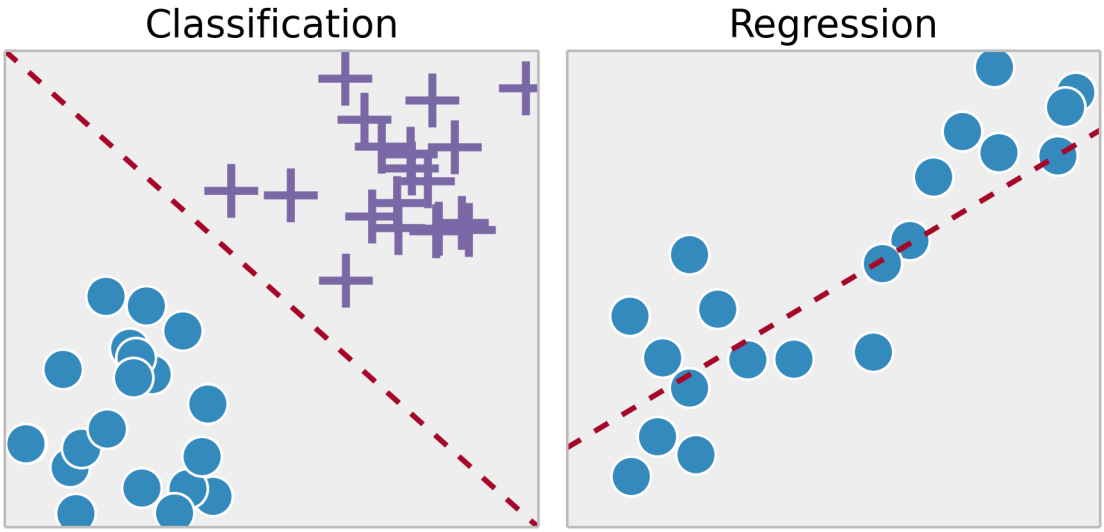
1. Επιβλεπόμενη Μάθηση (Supervised Learning): Σε αυτή την περίπτωση το μοντέλο δημιουργείται χρησιμοποιώντας ένα σύνολο εκπαίδευσης το οποίο περιέχει εισόδους και τις αντίστοιχες επιθυμητές εξόδους για το πρόβλημα που αντιμετωπίζεται. Συχνά περιγράφεται ως η περίπτωση ενός δασκάλου που εκπαιδεύει χρησιμοποιώντας παραδείγματα.
2. Μη-Επιβλεπόμενη Μάθηση (Unsupervised Learning): Σε αυτή την περίπτωση το σύνολο εκπαίδευσης δεν έχει επιθυμητές εξόδους για κάθε είσοδο. Στόχος είναι η εύρεση της διάταξης (pattern) που χαρακτηρίζει το σύνολο εκπαίδευσης.
3. Ενισχυτική Μάθηση (Reinforcement Learning): Σε αυτή την περίπτωση η εκπαίδευση γίνεται δυναμικά, με στόχο την κατά το δυνατόν ελαχιστοποίηση μιας μετρικής επίδοσης (performance metric). Κάθε απόφαση που λαμβάνεται επηρεάζει αυτή την μετρική αρνητικά η θετικά και με μελέτη αυτών των αλλαγών πραγματώνεται τελικά η διαδικασία της μάθησης.

Με βάση αυτές τις κατηγορίες, η Χωρική Παρεμβολή αποτελεί απο τη φύση της ένα πρόβλημα Επιβλεπόμενης Μάθησης. Έτσι, τα γνωστά σημεία που έχουν δειγματοληπτηθεί αποτελούν το σύνολο εκπαίδευσης, με την τιμή δειγματοληψίας στο καθένα απο αυτά να αποτελεί την επιθυμητή έξοδο για την συγκεκριμένη είσοδο.

Μορφή Εξόδου

Μια άλλη σκοπιά κατηγοριοποίησης των αλγορίθμων Μηχανικής Μάθησης που θεωρείται χρήσιμο να παρουσιαστεί αφορά την μορφή των εξόδων. Με βάση αυτές, οι αλγόριθμοι Μηχανικής μάθησης χωρίζονται σε δύο ευρείες κατηγορίες:

1. Αλγόριθμοι Ταξινόμησης (Classification): Σε αυτή την περίπτωση το πεδίο τιμών των εξόδων εκφράζει διακριτές κατηγορίες. Στόχος είναι λοιπόν η επιλογή της κατάλληλης κατηγορίας για δεδομένη είσοδο. Εναλλακτικά, στόχος είναι η εύρεση της γραμμής διαχωρισμού μεταξύ των διαφόρων κατηγοριών.
2. Αλγόριθμοι Παλινδρόμησης (Regression): Σε αυτή την περίπτωση το πεδίο τιμών των εξόδων είναι συνεχές. Στόχος είναι η κατά το δυνατόν βέλτιστη προσέγγιση της κατάλληλης τιμής για δεδομένη είσοδο. Εναλλακτικά, στόχος είναι η εύρεση της καμπύλης με το ελάχιστο σφάλμα όσον αφορά το σύνολο εκπαίδευσης.



Εποπτική παρουσίαση της διαφοράς μεταξύ προβλήματος ταξινόμησης και προβλήματος παλινδρόμησης

Τα προβλήματα Χωρικής Παρεμβολής μπορούν ταξινομηθούν και στις δύο κατηγορίες, ανάλογα με το ζητούμενο. Έτσι η αναζήτηση της τιμής του ύψους βροχόπτωσης σε άγνωστες θέσεις αποτελεί πρόβλημα παλινδρόμησης, ενώ ο χαρακτηρισμός ενός σημείου ως “πολύ”, “λίγο” ή “καθόλου” βροχερό με βάση τη θέση του αποτελεί πρόβλημα ταξινόμησης.

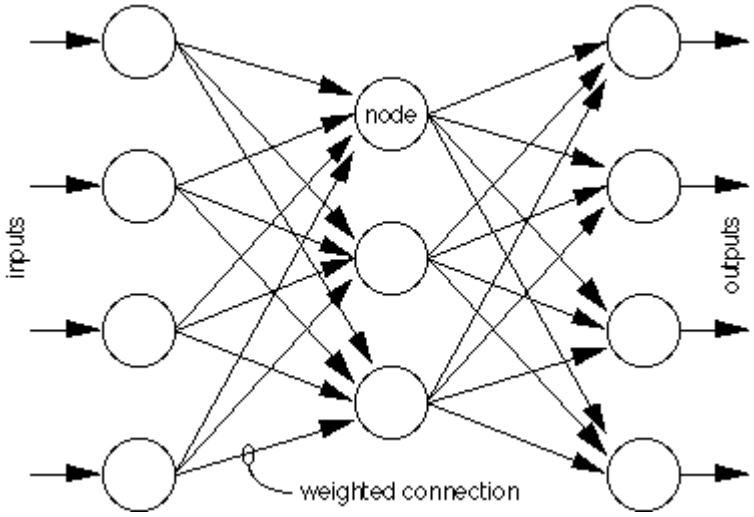
Στη συνέχεια παρουσιάζονται κάποιες βασικές τεχνικές Μηχανικής Μάθησης και η χρήση τους για την επίλυση του προβλήματος χωρικής παρεμβολής.

1.5 Βασικές Τεχνικές Μηχανικής Μάθησης

Σε αυτή την ενότητα παρουσιάζονται ορισμένες βασικές τεχνικές μηχανικής μάθησης, οι οποίες χρησιμοποιήθηκαν στη συγκεκριμένη μελέτη για την αντιμετώπιση του προβλήματος της παλινδρόμησης.

Πολυεπίπεδο Perceptron (Multi-Layer Perceptron – MLP)

Το πολυεπίπεδο Perceptron (Multi-Layer Perceptron, MLP) είναι μια βασική τεχνική επιβλεπόμενης μηχανικής μάθησης. Βασίζεται στην χρήση ενός πολυεπίπεδου δικτύου από συνδεδεμένους νευρώνες, καθένας απο τους οποίους (με εξαίρεση το επίπεδο εισόδου) χαρακτηρίζεται από μια μη γραμμική συνάρτηση ενεργοποίησης. Κάθε επίπεδο νευρώνων συνδέεται με τα γειτονικά του με συνάψεις, οι οποίες χαρακτηρίζονται από ένα βάρος. Συνεπώς, η εκπαίδευση του δικτύου επικεντρώνεται στην κατάλληλη προσαρμογή των βαρών αυτών, ωστέ για δεδομένη είσοδο να δίνεται η επιθυμητή έξοδος.



Τυπική μορφή ενός πολυεπίπεδου Perceptron

Το MLP μπορεί να χρησιμοποιηθεί τόσο σε προβλήματα Ταξινόμησης όσο και σε προβλήματα Παλινδρόμησης. Επιπροσθέτως, η δομή του είναι παραπλήσια και στις δύο περιπτώσεις, με αισθητή διαφορά μόνο την συνάρτηση ενεργοποίησης στο επίπεδο εξόδου, με βάση την οποία η έξοδος μπορεί να είναι είτε διακριτή (δηλαδή κατηγορία – class label) είτε συνεχής (πραγματική τιμή).

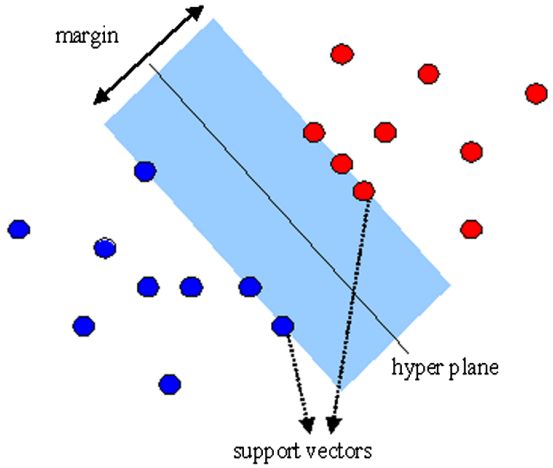
Ενδιαφέρον είναι το γεγονός πως τα MLP – ή αλλιώς Τεχνητά Νευρωνικά Δίκτυα (Artificial Neural Networks) – χρησιμοποιούνται στην πράξη σε πρακτικές εφαρμογές για επίλυση του προβλήματος χωρικής παρεμβολής. Σε αυτή την περίπτωση χρησιμοποιούνται σε συνδυασμό με τεχνικές γεωστατιστικής σε μια προσπάθεια εκμετάλλευσης των πλεοκτημάτων και των δύο προσεγγίσεων. [5] [6]

Ένα άμεσο πλεονέκτημα της χρήσης MLP έναντι των Γεωστατιστικών τεχνικών είναι η εύκολη προσαρμογή στην περίπτωση οπού η είσοδος έχει αυξημένη διάσταση. Με αυτό τον τρόπο μπορεί να αντικατοπτρίζεται για παράδειγμα η γνώση οτι το υπό μελέτη φαινόμενο εξαρτάται απο κάποιον επιπλεόν παράγοντα πέρα απο την γεωγραφική του θέση.

Παλινδρόμηση με Διανύσματα Υποστήριξης (Support Vector Regression – SVR)

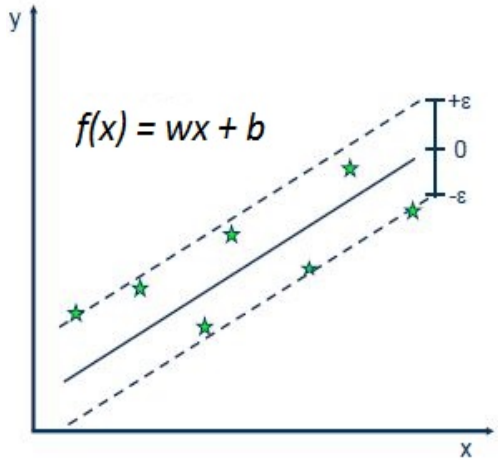
Η τεχνική SVR αποτελεί προσαρμογή των SVM (Support Vector Machines), τα οποία αντιμετωπίζουν το πρόβλημα της ταξινόμησης (classification), για την επίλυση του προβλήματος της παλινδρόμησης (regression).

Η γενική ίδεα πίσω από τα SVM βασίζεται στην αναπαράσταση των διαφόρων δεδομένων εισόδου ως σημεία στο χώρο. Με αυτό ως δεδομένο, και συγκεκριμένα για το πρόβλημα της ταξινόμησης, το SVM προσπαθεί να διαχωρίσει τις διάφορες κλάσεις ταξινόμησης με βέλτιστο τρόπο. Υπολογίζεται δηλαδή η γραμμή διαχωρισμού μεταξύ δύο κατηγοριών ώστε το μεταξύ αυτών κενό margin να είναι το μεγαλύτερο δυνατό.



Τυπική μορφή ταξινόμησης σε δύο κατηγορίες με χρήση διανυσμάτων υποστήριξης

Για την περίπτωση της παλινδρόμησης ειδικότερα, ο στόχος είναι η εύρεση της συνάρτησης $f(x)$, όπου x ο χώρος των μεταβλητών εισόδου, με την προϋπόθεση το σφάλμα να μην ξεπερνά την τιμή ε για κανένα απο τα στοιχεία του συνόλου εκπαίδευσης, ενώ ταυτόχρονα είναι επιθυμητό η $f(x)$ να έχει όσο το δυνατόν πιο επίπεδη μορφή.



Τυπική μορφή παλινδρόμησης με χρήση διανυσμάτων υποστήριξης

Αλλιώς, εφόσον η απόκλιση απο κάποια παρατήρηση θεωρείται πλήρως αποδεκτή εφόσον είναι μικρότερη απο ε , αλλά μη αποδεκτή αν ξεπερνά την τιμή ε . [7]

Ενδιαφέρον χαρακτηριστικό τόσο της τεχνικής SVM όσο και της τεχνικής SVR είναι η χρήση συναρτήσεων πυρήνα (Kernel Functions). Με αυτό τον τρόπο τα αρχικά δεδομένα μπορούν να απεικονιστούν σε χώρο μεγαλύτερης διαστατικότητας. Έτσι γίνεται εφικτός ο διαχωρισμός (ταξινόμηση) ή η προσέγγιση (παλινδρόμηση) αυτών με χρήση γραμμικής συνάρτησης $f(x)$.

K – Πλησιέστεροι γείτονες (K – Nearest Neighbors, KNN)

Η μέθοδος KNN είναι στην ουσία μέθοδος που δεν ανήκει στην μηχανική μάθηση, εφόσον δεν υπόκειται σε κάποια φάση εκπαίδευσης. Η λογική πίσω απο αυτή την μέθοδο είναι η εξής:

Για δεδομένη είσοδο, υπολογίζονται οι K – πλησιέστεροι γείτονες του συνόλου εκπαίδευσης.
Στη συνέχεια:

- στην περίπτωση της ταξινόμησης η έξοδος ορίζεται ως η κατηγορία στην οποία ανήκει η πλειοψηφία των K γειτόνων.
- στην περίπτωση της παλινδρόμησης (regression) υπολογίζεται ο μέσος όρος αυτών των K γειτόνων, ο οποίος αποτελεί την πρόβλεψη της μεθόδου για την συγεκριμένη είσοδο.

Το στοιχείο που οδηγεί στον χαρακτηρισμό της τεχνικής KNN ως μέθοδο μηχανικής μάθησης είναι ο τρόπος με τον οποίο γίνεται η επιλογή της παραμέτρου K, ώστε να ελαχιστοποιηθεί το συνολικό σφάλμα πρόβλεψης [3] . Η επιλογή αυτή γίνεται με την μέθοδο Διασταυρούμενης Επικύρωσης (Cross-Validation) που θα αναλυθεί στη συνέχεια.

Στην βασική της μορφή, η τεχνική KNN δεν ορίζει βάρη στον υπολογισμό της πλειοψηφίας ή του μέσου όρου των γειτόνων. Αυτή είναι και η βασική της διαφορά από τις τεχνικές IDW που αναφέρθηκαν προηγουμένως. Ωστόσο, η μέθοδος μπορεί να επεκταθεί με αντίστοιχη στάθμιση του υπολογισμού, οπότε και ταυτίζεται με τις τεχνικές IDW ως προς τον πυρήνα των υπολγισμών της.

Η τεχνική KNN αποτελεί μια ιδιαίτερα απλοϊκή τεχνική, ωστόσο βρίσκει εφαρμογή κυρίως λόγω της απλότητας αυτής και της ταχύτητας των υπολογισμών της. Πιο συγκεκριμένα, αποτελεί μια μορφή lazy learning, υπό την έννοια οτι η πρόβλεψη τιμών είναι εφικτή άμεσα, χωρίς να απαιτείται μια περίοδος εκπαίδευσης. Φυσικά, αυτό σημαίνει ότι το βάρος του υπολογισμού πέφτει κατά την πρόβλεψη τιμών, η οποία απαιτεί την έυρεση των πλησιέστερων γειτόνων.

2 Δέντρα Απόφασης

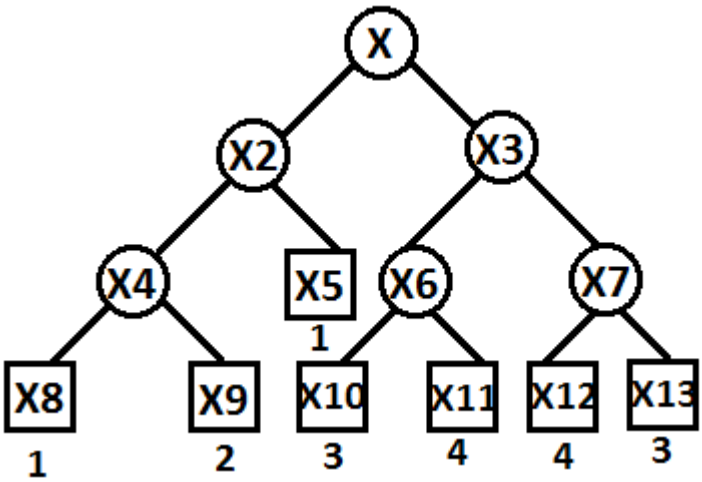
Σε αυτό το κεφάλαιο παρουσιάζεται η χρήση Δέντρων Απόφασης (Decision Trees) ως αλγορίθμων Μηχανικής Μάθησης. Ένα απο τα βασικότερα πλεονεκτήματα χρήσης των Decision Trees είναι η αντιμετώπιση σε μερικό βαθμό του προβλήματος της λογικής "μαύρου κουτιού" που αναφέρθηκε σε προηγούμενο κεφάλαιο.

Η σημασία αυτού του πλεονεκτήματος φαίνεται από το γεγονός οτι η επίλυση ενός προβλήματος με χρήση Μηχανικής Μάθησης έχει δύο βασικούς στόχους, των οποίων η σχετική σημασία διαφέρει ανάλογα με την περίσταση: Στόχος είναι έτσι αφ' ενός η επίτευξη ικανοποιητικής ακρίβειας πρόβλεψης εξόδου για δεδομένη είσοδο, αφ' ετέρου η κατανόηση της δομής που διέπει το πρόβλημα – φαινόμενο υπό μελέτη. [8]

2.1 Γενική Ιδέα των Δέντρων Απόφασης

Για λόγους απλότητας, αρχικά παρουσιάζεται η γενική λογική των Δέντρων Απόφασης (Decision Trees) όσον αφορά το πρόβλημα της Ταξινόμησης. Η περίπτωση της παλινδρόμησης θα εξεταστεί στη συνέχεια. Σε αυτή την παρουσίαση, ακολουθείται η μεθοδολογία και ο συμβολισμός που χρησιμοποιείται στο βιβλίο “Classification and Regression Trees”.[8]

Ένας ταξινομητής με δενδρική δομή κατασκευάζεται με διαδοχικούς διαχωρισμούς (splits) του πεδίου τιμών X του διανύσματος εισόδου σε υποσύνολα. Για την περίπτωση των δυαδικών δέντρων, που είναι και η πιο διαδεδομένη, κάθε διαχωρισμός χωρίζει το σύνολο X σε δύο υποσύνολα.



Ενδεικτική μορφή δέντρου απόφασης για ταξινόμηση

Στο παραπάνω παράδειγμα, υπάρχουν 4 κλάσεις ενώ τα σύνολα $X2, X3$ είναι συμπληρωματικά, με $X2 \cup X3 = X$. Αντίστοιχα για τα $X4, X5$ κ.ο.κ.

Τα φύλλα του δέντρου, τα οποία δεν διαχωρίζονται, δηλαδή τα $X5, X8, X9, X10, X11, X12, X13$ ονομάζονται τερματικά φύλλα και με βάση αυτά γίνεται η τελική ταξινόμηση για δεδομένη είσοδο. Τα τερματικά σύνολα αποτελούν μια διαμέριση του αρχικού συνόλου X και κάθε ένα απο αυτά χαρακτηρίζεται απο μια μόνο απο τις κλάσεις ταξινόμησης. Όπως και στο παράδειγμα, δύο ή περισσότερα τερματικά σύνολα μπορούν χαρακτηρίζονται από την ίδια κλάση ταξινόμησης.

Στο συγκεκριμένο παράδειγμα, οι κλάσεις ταξινομούνται ως εξής:

$$Class1 = X5 \cup X8$$
$$Class2 = X9$$
$$Class3 = X10 \cup X13$$
$$Class4 = X11 \cup X12$$

Κάθε διαχωρισμός γίνεται εφαρμόζοντας περιορισμούς στις συντεταγμένες του διανύσματος εισόδου. Θεωρητικά, οι περιορισμοί μπορούν να αφορούν πάνω απο μια συντεταγμένη για δεδομένο διαχωρισμό, ωστόσο στην πράξη συνήθως εφαρμόζονται περιορισμοί για μία μόνο συντεταγμένη ανά διαχωρισμό.
Η διαδικασία πρόβλεψης λοιπόν, ακολουθεί αυτό το “μονοπάτι περιορισμών” μέχρι να φτάσει σε κάποιο φύλλο – τερματικό υποσύνολο. Η πρόβλεψη είναι η κλάση που

χαρακτηρίζει το συγκεκριμένο τερματικό υποσύνολο.

Όπως φαίνεται, πρακτικά το πρόβλημα εκπαίδευσης ενός Δέντρου Απόφασης με βάση τα διαθέσιμα δεδομένα εκφράζεται σε τρία βασικά ζητήματα:

- Τον τρόπο με τον οποίο γίνεται κάθε διαχωρισμός (split).
- Το κριτήριο με το οποίο ορίζεται κάποιο σύνολο ως τερματικό. Αλλιώς, το κριτήριο τερματισμού της διαδικασίας διαχωρισμών.
- Τον τρόπο με τον οποίο χαρακτηρίζεται κάθε τερματικό σύνολο.

Στη συνέχεια, θα χρησιμοποιείται ορολογία γράφων για την περιγραφή των διαφόρων στοιχείων του δέντρου απόφασης. Για παράδειγμα, το αρχικό σύνολο X θα αποκαλείται ρίζα (root node) t_1 ενώ τα διάφορα υποσύνολα αποτελούν κόμβους (nodes). Με βάση αυτή την ορολογία, ακολουθούν κάποιοι ορισμοί:

- ως L ορίζεται το αρχικό σύνολο εκπαίδευσης (learning set) το οποίο εξετάζει J κλάσεις, ενω περιλαμβάνει συνολικά N παραδείγματα και N_j παραδείγματα της κλάσης j.
- ως $N(t)$ ορίζεται ο αριθμός συνολικών παραδειγμάτων στον κόμβο t. Έτσι, $N_j(t)$ είναι ο αριθμός παραδειγμάτων της κλάσης j που περιλαμβάνει ο κόμβος t.

Σε αυτό το σημείο πρέπει να σημειωθεί η διαφορά μεταξύ *παρατηρούμενης* και *πραγματικής* πιθανότητας εμφάνισης μιας κλάσης. Έτσι, μπορεί για παράδειγμα, με βάση το σύνολο L, η πρώτη κλάση να εμφανίζεται στο 50% των περιπτώσεων, ωστόσο να υπάρχει γνώση οτι στην πράξη εμφανίζεται σπανιότερα ή συχνότερα. Γι αυτό το λόγο ορίζεται:

- ως $\pi(j)$ η εξαρχής πιθανότητα (prior probability) εμφάνισης της κλάσης j στο δέντρο.

Με αυτό τον τρόπο μπορεί να ληφθεί υπ' όψην ήδη υπάρχουσα γνώση για το πρόβλημα. Σε περίπτωση που δεν υπάρχει τέτοια γνώση, ορίζεται $\pi(j)=\frac{N(j)}{N}$. Με βάση αυτό τον ορισμό μπορούμε να ορίσουμε:

- ως $p(j,t)$ την πιθανότητα εμφάνισης της κλάσης j στον κόμβο t.

Συνεπώς, με βάση τα παραπάνω, είναι:

$$p(j,t)=\pi(j)\frac{N_j(t)}{N_j}$$

Επομένως η πιθανότητα εμφάνισης οποιασδήποτε κλάσης στον κόμβο t είναι:

$$p(t)=\sum_j p(j,t)$$

Και με βάση αυτή προκύπτει η δεσμευμένη πιθανότητα οτι ένα πρότυπο ανήκει στην κλάση j δεδομένου του ότι βρίσκεται στον κόμβο t:

$$p(j\mid t)=\frac{p(j,t)}{p}(t)$$

2.2 Ο Τρόπος Διαχωρισμού

Ο βασικός στόχος διαχωρισμού του συνόλου X σε υποσύνολα είναι η απόκτηση δύο υποσυνόλων τα οποία παρουσιάζουν κατά το δυνατόν περισσότερη ομοιογένεια ως προς την ποικιλία κλάσεων ταξινόμησης. Ακολουθώντας αυτή τη λογική, τελικός στόχος είναι η δημιουργία τερματικών κόμβων που παρουσιάζουν πλήρη ομοιογένεια, δηλαδή χαρακτηρίζουν πλήρως μια από τις κλάσεις.

Φυσικά η πλήρης ομοιογένεια δεν είναι πάντα εφικτή, συνεπώς ορίζεται μια μετρική "ανομοιογένειας" (impurity) $i(t)$, ώστε το παραπάνω να μεταφράζεται σε ελαχιστοποίηση αυτής σε κάθε διαχωρισμό.

Για τον ορισμό της $i(t)$ ορίζεται αρχικά μια βοηθητική συνάρτηση ανομοιογένειας $\varphi(p(1),\ldots,p(J))$ με ορίσματα τις πιθανότητες $p(j)$, η οποία θέλουμε:

1. Να εμφανίζει μέγιστο μόνο στην περίπτωση που είναι εξίσου πιθανή η εμφάνιση όλων των κλάσεων, δηλαδή στη θέση $(\frac{1}{j},\frac{1}{j},\ldots,\frac{1}{j})$
2. Να εμφανίζει ελάχιστο μόνο στην περίπτωση πλήρους ομοιογένειας, δηλαδή ύπαρξης μοναδικής κλάσης $(1,0,\ldots,0),(0,1,0,\ldots,0),\ldots,(0,0,\ldots,0,1)$
3. Να είναι συμμετρική ως προς τα ορίσματα $p(1),\ldots,p(J)$

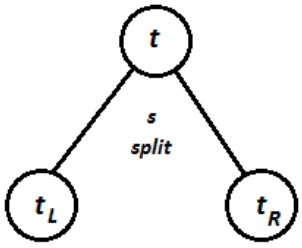
Δεδομένης αυτής της συνάρτησης φ, η μετρική ανομοιογένειας $i(t)$ μπορεί να υπολογιστεί χρησιμοποιώντας την δεσμευμένη πιθανότητα εμφάνισης κάθε κλάσης j στον κόμβο t:

$$i(t)=\varphi(p(1|t),p(2|t),\ldots,p(J|t))$$

Με βάση αυτή την μετρική μπορούμε να υπολογίσουμε την μείωση της ανομοιογένειας σε ένα διαχωρισμό s και άρα την αποτελεσματικότητα εφαρμογής αυτού ως εξής:

$$\Delta i(s,t)=i(t)-\frac{N_{t_L}}{N_t}i(t_L)-\frac{N_{t_R}}{N_t}i(t_R)$$

Όπου με t_L,t_R ορίζεται ο δεξιός και ο αριστερός αντίστοιχα κόμβος που προκύπτει με τον διαχωρισμό s.



Γενική μορφή διαχωρισμού δέντρου απόφασης

Αυτός ο ορισμός μπορεί να χρησιμοποιηθεί για να εκφράσει την συνολική ανομοιογένεια του δέντρου T που κατασκευάζεται. Ορίζοντας $\widetilde{T}$ το τρέχων σύνολο τερματικών κόμβων, η ανομοιογένεια $I(T)$ του δέντρου μπορεί να υπολογιστεί ως:

$$I(T)=\sum_{t\in\widetilde{T}} I(t)=\sum_{t\in\widetilde{T}} i(t)\,p(t)$$

Έτσι, η προσπάθεια επίτευξης της μέγιστης τιμής "μείωσης ανομειογένειας" $\Delta i(s,t)$, οπού υπενθυμίζεται πως η μεταβλητή που διερευνάται είναι ο διαχωρισμός s, ισοδυναμεί με την επιλογή του διαχωρισμού s που θα ελαχιστοποιήσει την συνολική ανομοιογένεια δέντρου $I(T)$.
Έτσι, το νέο δέντρο T' θα έχει ανομοιογένεια:

$$I(T')=I(T)+I(t_L)+I(t_R)$$
$$I(T')=\sum_{t\in\widetilde{T}} I(t)+I(t_L)+I(t_R)$$

Απ' όπου προκύπτει πως η μείωση ανομοιογένειας στο νέο δέντρο T' το οποίο δημιουργείται από τον διαχωρισμό s στον κόμβο t είναι:

$$\Delta I(s,t)=I(T)-I(T')=I(t)-I(t_L)-I(t_R)$$

Από αυτή την ανάλυση τελικά προκύπτει πως η επιλογή του καλύτερου διαχωρισμού σε δεδομένο κόμβο t εκφράζεται ως επιλογη του διαχωρισμού s που μεγιστοποιεί την παραπάνω ποσότητα $\Delta I(s,t)$.

Με αυτούς τους ορισμούς είναι λοιπόν δυνατό να επιλέξουμε τον βέλτιστο ανάμεσα σε διάφορους τρόπους διαχωρισμού.

Το επόμενο ζήτημα είναι με ποιον τρόπο ορίζονται οι τρόποι διαχωρισμού, ώστε να μπορεί μετά να γίνει επιλογή μεταξύ αυτών με τις παραπάνω τεχνικές. Τυπικά, στόχος είναι ο προσδιορισμός ενός συνόλου S που αποτελείται από διάφορους πιθανούς τρόπους διαχωρισμού, ώστε να επιλεγεί ο διαχωρισμός s που μεγιστοποιεί την παραπάνω ποσότητα.

Αρχικά, υποθέτουμε πως τα δεδομένα μας έχουν καθορισμένη δομή και άρα διαστατικότητα. Υποθέτουμε, δηλαδή, πως όλα τα διανύσματα μετρήσεων του συνόλου L έχουν σταθερή μορφή $x=(x_1, \ldots, x_M)$ οπού M η σταθερή διάσταση αυτών.

Σε αυτή την περίπτωση, το σύνολο διαχωρισμών S σε δεδομένο κόμβο t δημιουργείται ακολουθώντας τους εξής κανόνες:

1. Κάθε διαχωρισμός έχει ως συνθήκη την τιμή μίας μόνο εκ των μεταβλητών του διανύσματος x.
2. Για κάθε διατεταγμένη μεταβλητή x_m, οι πιθανοί διαχωρισμοί γίνονται με βάση τον περιορισμό $x_m \le c$, για όλες τις τιμές c εντός του διαστήματος $(-\infty,\infty)$.

3. Για κάθε κατηγορική μεταβλητή x_m , η οποία μπορεί να λάβει τιμές $\{b_1, b_2, \ldots, b_L\}$, οι πιθανοί διαχωρισμοί γίνονται με βάση την συνθήκη $x_m \in B$, οπού το B αντιπροσωπεύει όλα τα πιθανά υποσύνολα του συνόλου $\{b_1, b_2, \ldots, b_L\}$.

Ο δεύτερος κανόνας μπορεί να δώσει την εντύπωση οτι οδηγεί στην δημιουργία μη-πεπερασμένου αριθμού πιθανών διαχωρισμών. Ωστόσο αυτό δεν ισχύει καθώς κάθε διαχωρισμός στην ουσία εκφράζει δύο συμπληρωματικά υποσύνολα t_L, t_R του συνόλου που εμφανίζεται στον κόμβο t , το οποίο με τη σειρά του αποτελεί υποσύνολο του αρχικού συνόλου L .
Έτσι, οι πιθανές τιμές c που έχουν νόημα (δηλαδή θα δώσουν διαφορετικά σύνολα μετά τον διαχωρισμό) είναι πεπερασμένες και στην πράξη επιλέγονται ως οι ενδιάμεσες τιμές δύο διαδοχικών γνωστών τιμών της διατεταγμένης μεταβλητής x_m με βάση την οποία γίνεται ο εν λόγω διαχωρισμός.

Για τον τρίτο κανόνα φυσικά δεν εμφανίζεται τέτοια ανησυχία καθώς ο αριθμός υποσυνόλων του συνόλου πιθανών τιμών της κατηγορικής μεταβλητής $\{b_1, b_2, \ldots, b_L\}$ είναι πεπερασμένος.

Έχοντας λοιπόν ορίσει και τον τρόπο κατασκευής των διαφόρων πιθανών διαχωρισμών για έναν κόμβο, η επιλογή του διαχωρισμού για δεδομένο κόμβο t γίνεται με την εξής διαδικασία:

1. Υπολογίζεται το σύνολο πιθανών διαχωρισμών S .
2. Αναζητείται ο διαχωρισμός $s \in S$ ο οποίος μεγιστοποιεί την $\Delta I(s,t)$.

2.3 Το Κριτήριο Τερματισμού – Ορισμού Τερματικού Κόμβου

Ιδανικά, θα ορίζαμε κάθε κόμβο ως τερματικό αν ήταν πλήρως ομοιογενής, δηλαδή αν αντιπροσώπευε αποκλειστικά μία απο τις κλάσεις ταξινόμησης. Στην πράξη ωστόσο συνήθως αυτό δεν αποτελεί ρεαλιστική προσδοκία.

Επομένως ο τρόπος με τον οποίο ένας κόμβος ορίζεται τερματικός βασίζεται στο κριτήριο επιλογής διαχωρισμού που αναλύθηκε προηγουμένως. Συγκεκριμένα, δεδομένου ότι η τιμή ΔI εκφράζει έμμεσα το πόσο βελτιώνεται το δέντρο επιλογής (εφόσον γίνεται περισσότερο ομογενές), μπορούμε να ορίσουμε ένα κατώφλι επιθυμητής βελτίωσης $\beta > 0$, το οποίο αποτρέπει τον διαχωρισμό εφόσον δεν ξεπερνάται:
t τερματικός κόμβος εφόσον $max\, \Delta I(s,t) < \beta$, για $s \in S$.

Ορισμός της Κλάσης που αντιπροσωπεύει ο Τερματικός Κόμβος

Με βάση τους παραπάνω ορισμούς, είναι εύκολο να οριστεί η κλάση $\tilde{j}(t)$ που αντιπροσωπεύει ο κόμβος t χρησιμοποιώντας την δεσμευμένη πιθανότητα εμφάνισης της κλάσης j στον κόμβο t, $p(j|t)$. Επιπλέον σημειώνεται οτί κατά τον υπολογισμό της πιθανότητας αυτής λαμβάνεται υπ' όψην η πιθανή ύπαρξη προηγούμενης γνώσης σχετικά με την συχνότητα εμφάνισης κάθε κλάσης (υπενθυμίζεται οτι αυτό προκύπτει από τον τρόπο ορισμού των $\pi(j)$). Έτσι:

$$\tilde{j}(t) = j \quad \text{αν} \quad p(j|t) = \max_i p(i|t) \quad .$$

Άμεσο επακόλουθο αυτού του ορισμού είναι και δύο μετρικές της πιθανότητας λάθος ταξινόμησης:

- $r(t) = 1 - \max_j p(j|t)$ η πιθανότητα λάθος ταξινόμησης για τον κόμβο t.
- $R(T) = \sum_{t \in \tilde{T}} r(t)\, p(t) = \sum_{t \in \tilde{T}} R(t)$ η πιθανότητα λάθος ταξινόμησης για το δέντρο T.

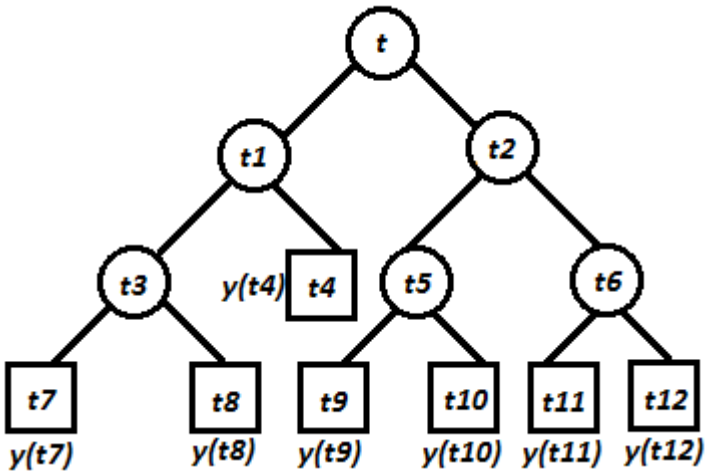
2.4 Δέντρα Παλινδρόμησης

Η παραπάνω μεθοδολογία αφορά την χρήση δέντρων απόφασης για επίλυση του προβλήματος ταξινόμησης. Η χρήση αυτών για επίλυση προβλημάτων παλινδρόμησης ακολουθεί την ίδια βασική λογική, ωστόσο υπάρχουν κάποιες βασικές διαφορές, οι οποίες κρίνεται σκόπιμο να αναλυθούν σε αυτό το κεφάλαιο.

Οι διαφορές μεταξύ των δύο προσεγγίσεων πηγάζουν από την διαφορά της μορφής επιθυμητών εξόδων σε κάθε περίπτωση. Έτσι, ενώ στην ταξινόμηση υπάρχει η απλότητα πεπερασμένου αριθμού διακριτών κατηγοριών στην έξοδο, η ιδιότητα αυτή δεν ισχύει στα δέντρα παλινδρόμησης.

Τυπικότερα, το σύνολο εκπαίδευσης L αποτελείται απο δεδομένα της μορφής (x, y) οπού το x είναι το διάνυσμα μετρήσεων και περιλαμβάνει μεταβλητές μετρήσεων, ενώ y είναι η μεταβλητή εξόδου, η οποία αποτελεί πραγματικό αριθμό. Τελικός στόχος είναι η κατασκευή ενός κανόνα πρόβλεψης (prediction rule), δηλαδή μια συνάρτηση $d(x)$ που συνδέει το διάνυσμα μετρήσεων με αντίστοιχες προβλέψεις – εξόδους.

Ενδεικτικά, η μορφή ενός δέντρου παλινδρόμησης είναι η εξής:



Ενδεικτική μορφή δέντρου παλινδρόμησης

Οπού, όπως φαίνεται κάθε τερματικός κόμβος t χαρακτηρίζεται πλέον απο μια μοναδική, σταθερή τιμή πρόβλεψης $y(t)$.

Όπως και στην περίπτωση της ταξινόμησης, η κατασκευή ενός Δέντρου Παλινδρόμησης (Regression Tree) με βάση το σύνολο γνωστών σημείων L εκφράζεται σε τρία βασικά ζητήματα:

- Τον τρόπο με τον οποίο γίνεται κάθε διαχωρισμός (split).
- Το κριτήριο με το οποίο ορίζεται κάποιο σύνολο ως τερματικό. Αλλιώς, το κριτήριο τερματισμού της διαδικασίας διαχωρισμών.
- Τον τρόπο με τον οποίο χαρακτηρίζεται κάθε τερματικό σύνολο με κάποια τιμή $y(t)$.

Επιλογή τιμής $y(t)$

Η επιλογή αυτή γίνεται με βασικό κριτήριο την κατά το δυνατόν ελάχιστη τιμή προβλεπόμενου σφάλματος για τον ταξινομητή d (δηλαδή το δέντρο παλινδρόμησης) που κατασκευάζεται.

Έτσι, το σφάλμα αυτό τυπικά ορίζεται ως:

$$R(d)=\frac{1}{N}\sum_{n}(y_n-d(x_n))^2$$

όπου τα N σε αριθμό στοιχεία (x_n, y_n) είναι τα γνωστά σημεία, δηλαδή τα στοιχεία του συνόλου L .

Προκύπτει λοιπόν πως οι κατάλληλες τιμές $y(t)$ για κάθε κόμβο t (μέσω των οποίων προκύπτουν οι διάφορες προβλέψεις $d(x_n)$ στην παραπάνω σχέση σφάλματος) είναι ο μέσος όρος των γνωστών τιμών y_n που περιλαμβάνει ο εκάστοτε κόμβος. Δηλαδή:

$$\bar{y}(t)=\frac{1}{N(t)}\sum_{x_n\in t} y_n$$

Ορίζοντας λοιπόν την τιμή $\bar{y}(t)$ μπορούμε να προχωρήσουμε στον προσδιορισμό της επιλογής διαχωρισμού, ο οποίος ακολουθεί λογική αντίστοιχη με αυτή που χρησιμοποιείται στην περίπτωση ταξινόμησης.

Τρόπος εφαρμογής διαχωρισμού:

Έτσι, όπως αντίστοιχα ορίστηκε η μετρική ανομοιογένειας, μπορούμε να ορίσουμε τώρα την μετρική προβλεπόμενου σφάλματος για έναν κόμβο t ως:

$$R(t)=\frac{1}{N}\sum_{x_n\in t}(y_n-\bar{y}(t))^2$$

Με βάση την οποία, εφαρμοζόμενη στο σύνολο τερματικών κόμβων $\tilde{T}$ ενός δέντρου T , προκύπτει το προβλεπόμενο σφάλμα για το συνολικό δέντρο:

$$R(T)=\sum_{t\in\tilde{T}} R(t)$$

Ακολουθώντας λοιπόν μεθοδολογία αντίστοιχη με αυτή των δέντρων ταξινόμησης, ορίζουμε την μείωση σφάλματος $\Delta R(s,t)$ για τον διαχωρισμό s του κόμβου t σε κόμβους t_L, t_R :

$$\Delta R(s,t)=R(t)-R(t_L)-R(t_R)$$

Οπότε κατά την διαδικασία διαχωρισμού επιλέγεται ο διαχωρισμός $\tilde{s}$ που επιτυγχάνει την μέγιστη μείωση σφάλματος:

$$\Delta R(\tilde{s},t)=max\,\Delta R(s,t) \quad \text{οπού} \quad s\in S$$

Σημειώνεται πως το σύνολο δυνατών διαχωρισμών S κατασκευάζεται με τον ίδιο τρόπο με τον οποίο κατασκευάζεται και στην περίπτωση ταξινόμησης.

Ένα ισχυρότερο κριτήριο επιλογής του βέλτιστου διαχωρισμού χρησιμοποιεί και την πιθανότητα $p(t)=\frac{N(t)}{N}$. Η πιθανότητα αυτή εκφράζει την πιθανότητα εμφάνισης μιας τυχαία επιλεγμένης περίπτωσης από την θεωρητική κατανομή (που χαρακτηρίζει το υπό μελέτη πρόβλημα) στον κόμβο t . Έτσι, ορίζουμε την προηγούμενη μετρική σφάλματος

$$s^2(t)=\frac{1}{N}\sum_{x_n\in t}(y_n-\bar{y}(t))^2$$

και η μετρική σφάλματος που χρησιμοποιείται τελικά – με τρόπο όμοιο με αυτόν που αναλύθηκε – είναι η:

$$R(T)=\sum s^2(t)p(t) \quad \text{, για} \quad t\in\widetilde{T}$$

Η συνθήκη τερματισμού, δηλαδή επιλογής ενός κόμβου ως τερματικού, προκύπτει με τον ίδιο τρόπο με την περίπτωση της ταξινόμησης, δηλαδή ορίζοντας ένα κατώφλι β για την παραπάνω μετρική σφάλματος.

2.5 Κλάδεμα (Pruning) του Δέντρου Απόφασης

Τα παραπάνω περιγράφουν την γενική λογική που εκφράζει τα δέντρα απόφασης. Στην πράξη όμως παρατηρήθηκε πως η συνθήκη επιλογής ενός κόμβου ως τερματικού δεν έδινε τα επιθυμητά αποτελέσματα: Τα δέντρα προκύπτουν ιδιαίτερα μεγάλα, ενώ η συγκεκριμένη συνθήκη απο τη φύση της τείνει να προσαρμόσει το δέντρο στα δεδομένα εκπαίδευσης, επομένως να οδηγήσει σε over-fitting.

Η λογική βέλτιστης επιλογής του κατωφλίου β , τόσο στην περίπτωση ταξινόμησης όσο και στη περίπτωση παλινδρόμησης, δεν αντιμετωπίζει το πρόβλημα. Πολύ χαμηλές τιμές του β οδηγούν σε υπερβολικά πολύ μεγάλο αριθμό διαχωρισμών και άρα σε υπερβολικά πολύ μεγάλα δέντρα απόφασης. Αντίθετα, με πολύ μεγάλες τιμές του β αυξάνεται η πιθανότητα πραγματοποίησης διαχωρισμών που τελικά δεν θα οδηγήσουν σε καλά αποτελέσματα.

Η λογική που φαίνεται να αντιμετωπίζει τα παραπάνω προβλήματα ξεφεύγει από τον κανόνα τερματισμού και συγκεντρώνεται στην εξής διαδικασία:

Εφαρμογή pruning (κλαδέματος) του δέντρου αντί για προσαρμογή του τρόπου τερματισμού. Αυτό σημαίνει πως δημιουργείται αρχικά ένα δέντρο το οποίο είναι σχετικά μεγάλο, εντούτοις έπειτα διαγράφονται κλάδοι αυτού με κατάλληλο τρόπο (δηλαδή με κατάλληλα κριτήρια) ώστε να επιτευχθεί η βέλτιστη δομή του.

Έτσι, η διαδικασία αρχικά περιλαμβάνει την δημιουργία ενός ιδιαίτερα μεγάλου δέντρου T_{max} του οποίου οι τερματικοί κλάδοι είναι είτε πολύ μικροί είτε πλήρως ομογενείς (δηλαδή περιλαμβάνουν πρότυπα μίας μόνο κλάσης στην περίπτωση ταξινόμησης). Η ιδανική μορφή αυτού του δέντρου, η οποία όμως έχει αυξημένο βάρος υπολογισμών, είναι η εφαρμογή διαχωρισμών μέχρις ότου κάθε κόμβος να περιλαμβάνει ένα μόνο πρότυπο εισόδου.

Στη συνέχεια πρέπει με κάποιο τρόπο να εφαρμοστεί pruning του T_{max} , ώστε να καταλήξουμε σε ένα δέντρο απόφασης που συνδυάζει καλή ακρίβεια και ικανότητα γενίκευσης. Αυτή η διαδικασία περιλαμβάνει την επιλογή μεταξύ των διαφόρων υποδέντρων δεδομένου μεγέθους σε κάθε βήμα. Η διαδικασία υπολογίζει την επίδοση όλων αυτών των πιθανών υποδέντρων, και συνεχίζεται μειώνοντας διαδοχικά το μέγεθος των υποδέντρων που υπολογίζονται, οπότε τελικά καταλήγουμε στον μοναδικό κόμβο-ρίζα του δέντρου.

Το κριτήριο που χρησιμοποιείται για την τελική επιλογή του βέλτισου υποδέντρου είναι η τιμή $R(T)$, υπολογισμένη για καθένα απο τα υποδέντρα ανάμεσα στα οποία θα γίνει η επιλογή. Έτσι, ενώ αυτή η μετρική δεν δίνει μια ικανοποιητική εικόνα του αντικειμενικού ποσοστού σφάλματος του δέντρου, αποτελεί χρήσιμη μετρική για την σύγκριση μεταξύ διαφόρων υποδέντρων που βασίζονται στο ίδιο σύνολο εκπαίδευσης.

3 Μάθηση Συνόλου

Η βασική ιδέα της μάθησης συνόλου είναι η δημιουργία περισσότερων του ενός ταξινομητών, ωστέ τελικά να συνδυαστούν τα αποτελέσματα τους για να επιτευχθεί απόδοση καλύτερη απο την απόδοση καθενός ξεχωριστά.
Εναλλακτικά, μπορούμε να θεωρήσουμε οτι η δημιουργία ενός ταξινομητή αποτελεί ουσιαστικά μια διαδικασία αναζήτησης στο χώρο υποθέσεων (που αφορά το εκάστοτε πρόβλημα) με σκοπό την εύρεση μιας υπόθεσης που δίνει τα καλύτερα αποτελέσματα. Υπο αυτή την σκοπιά, και εφόσον η εύρεση μιας υπόθεσης με πολύ καλά αποτελέσματα μπορεί να είναι ιδιαίτερα δύσκολη διαδικασία, η μάθηση συνόλου χρησιμοποιεί διάφορες υποθέσεις οι οποίες μπορεί να μην είναι τόσο ισχυρές, ωστέ τελικά ο συνδυασμός αυτών να δώσει τα καλύτερα αποτελέσματα.

Εξάλλου, δεδομένου του ότι κάποιοι αλγόριθμοι της μηχανικής μάθησης είναι εμπνευσμένοι από τον τρόπο με τον οποίο μαθαίνει ο άνθρωπος (π.χ. τεχνητά νευρωνικά δίκτυα), η μάθηση συνόλου μπορεί να αντιστοιχιστεί στην τάση των ανθρώπων να συμβουλευτούν περισσότερους του ενός "ειδικούς" προτού αποφασίσουν για κάποιο σημαντικό πρόβλημα.
[9]

3.1 Γενική Δομή μάθησης συνόλου – Ορισμοί

Όπως και προηγουμένως, αρχικά θα παρουσιαστούν οι μέθοδοι μάθησης συνόλου για την περίπτωση του προβλήματος της ταξινόμησης.

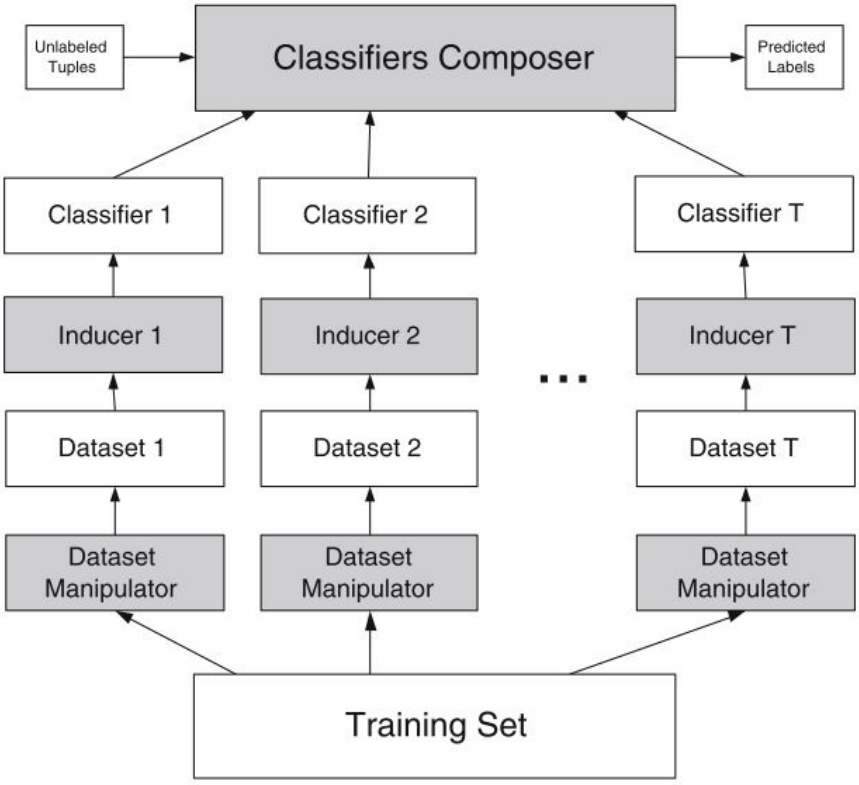
Αρχικά, η τυπική δομή ενος ταξινομητή βασιζόμενου στη μάθηση συνόλου περιλαμβάνει τα εξής:

- Σύνολο Εκπαίδευσης. Περιλαμβάνει τα γνωστά πρότυπα, τα οποία θα χρησιμοποιηθούν κατά την εκπαίδευση. Το σύνολο αυτό συνήθως χωρίζεται σε υποσύνολα ή μπορεί να εμπλουτιστεί με περεταίρω πληροφορίες (π.χ. βάρος σημασίας προτύπου), ανάλογα με την τεχνική που εφαρμόζεται.
- Βασικός Επαγωγέας (Base Inducer). Συνήθως οι τεχνικές μάθησης συνόλου βασίζονται στην δημιουργία πολλών ταξινομητών που ακολουθούν την ίδιο αλγόριθμο μηχανικής μάθησης (π.χ. πολλά δέντρα απόφασης). Ο αλγόριθμος αυτός που ορίζει την γενική σχέση μεταξύ εισόδου και εξόδο αποτελεί τον βασικό επαγωγέα.
- Γεννήτρια Ποικιλομορφίας (Diversity Generator). Εφόσον οι διάφοροι ταξινομητές παράγονται με βάση τον ίδιο βασικό επαγωγέα, είναι απαραίτητη η χρήση μιας τυπικής μεθόδου, ωστέ να εξασφαλίζεται οτι οι μεταξύ αυτών διαφορές έχουν νόημα.
- Συνδυαστής (Combiner). Εκφράζει την μέθοδο με την οποία συνδυάζονται τελικά οι προβλέψεις των διαφόρων ταξινομητών που δημιουργούνται.

Οι τεχνικές μάθησης συνόλου χωρίζονται σε δύο ευρείες κατηγορίες, με βάση την σχέση μεταξύ των διαφόρων ταξινομητών που κατασκευάζονται. Έτσι, ως εξαρτώμενοι (dependent framework) ονομάζονται οι αλγόριθμοι κατά τους οποίους η έξοδος ενός ήδη κατασκευασμένου ταξινομητή χρησιμοποιείται για την κατασκευή του επόμενου, ωστέ αυτός να εκμεταλλευτεί πιθανή πληροφορία για την ως τώρα ακρίβεια των κατασκευαζόμενων ταξινομητών. Αντίθετα, αν κάθε ταξινομητής κατασκευάζεται ανεξάρτητα από τους προηγούμενους, ο αλγόριθμος μάθησης συνόλου ονομάζεται μη-εξαρτούμενος (independent framework). [10]

3.2 Μη-Εξαρτώμενοι Αλγόριθμοι μάθησης συνόλου

Σε αυτή την περίπτωση το αρχικό σύνολο εκπαίδευσης που είναι διαθέσιμο χωρίζεται σε υποσύνολα, καθένα των οποίων χρησιμοποιείται για την εκπαίδευση ενός μόνο ταξινομητή. Ανάλογα με τον συγκεκριμένο αλγόριθμο, τα υποσύνολα αυτά μπορεί να είναι αμοιβαία αποκλειόμενα ή να έχουν κοινά στοιχεία.



Γενική δομή ενός συστήματος μη-εξαρτώμενης μάθησης συνόλου

Σε αυτή την περίπτωση, εφόσον πρακτικά κάθε ταξινομητής που δημιουργείται είναι ανεξάρτητος από τους υπόλοιπους, είναι πιο εύκολη η χρήση διαφορετικών βασικών επαγωγέων. Έτσι, μπορεί κάποιοι ταξινομητές να βασίζονται για παράδειγμα στην μεθοδολογία τεχνητών νευρωνικών δικτύων ενώ κάποιοι άλλοι να χρησιμοποιούν μεθοδολογία δέντρων απόφασης. Οι έξοδοι αυτών των ταξινομητών τελικά συνδυάζονται για να παραχθεί η τελική πρόβλεψη του ολικού συστήματος για δεδομένη είσοδο.

Ένα ακόμα πλεονέκτημα που προκύπτει άμεσα απο την ανεξαρτησία που εκφράζει αυτή την προσέγγιση είναι η εύκολη παραλληλοποίηση της διαδικασίας και κατά συνέπεια η μείωση του συνολικού χρόνου εκτέλεσης. [10]

Παράδειγμα: Bagging

Ένας χαρακτηριστικός αλγόριθμος που ανήκει στην μη-εξαρτώμενη μεθοδολογία είναι ο αλγόριθμος Bagging. Ο αλγόριθμος αυτός εκφράζει την μη-εξαρτώμενη προσέγγιση στην πιο απλή της μορφή: Με δεδομένο το αρχικό σύνολο εκπαίδευσης S , κάθε ταξινομητής εκπαιδεύεται με βάση ένα υποσύνολο του S , μεγέθους μ .

Τα υποσύνολα αυτά δειγματοληπτούνται με αντικατάσταση απο το S . Αυτό έχει ως αποτέλεσμα την πιθανή εμφάνιση του ίδιου προτύπου περισσότερες απο μία φορές στο

ίδιο υποσύνολο καθώς και την πιθανή έλλεψη κάποιου προτύπου απο αυτό.

Με αυτό τον τρόπο εκπαιδεύεται ένα σύνολο ανεξάρτητων ταξινομητών, οι οποίοι έχουν κατασκευαστεί με βάση διαφορετικά – αλλά όχι ανεξάρτητα, εφόσον προέρχονται απο το ίδιο αρχικό υπερσύνολο – σύνολα εκπαίδευσης. Συνεπώς τελικά οι προβλέψεις καθενός συνδυάζονται με μια απλή λογική ψηφοφορίας, με αποτέλεσμα η πρόβλεψη κατηγορίας για δεδομένη είσοδο να είναι η κατηγορία την οποία προβλέπει – ψηφίζει η πλειοψηφία των ταξινομητών.

Η τεχνική Bagging, αν και απλοϊκή ως προς την βασική λογική της, μπορεί να δώσει καλύτερα αποτελέσματα απο την εκπαίδευση ενός μοναδικού ταξινομητή χρησιμοποιώντας τον αλγόριθμο του βασικού επαγωγέα και ολόκληρο το σύνολο εκπαίδευσης. Αυτή η βελτίωση αποτελεσμάτων είναι ιδιαίτερα πιθανή στην περίπτωση που οι χρησιμοποιούμενοι ταξινομητές θεωρούνται ασταθείς (unstable). [11]

Σημειώνεται οτι ασταθής θεωρείται ένας ταξινομητής του οποίου η τελική μορφή επηρεάζεται σε μεγάλο βαθμό απο μικρές συγκριτικά αλλαγές στο σύνολο εκπαίδευσης αυτού.

3.3 Εξαρτώμενοι Αλγόριθμοι μάθησης συνόλου

Σε αντίθεση με την προηγούμενη προσέγγιση, η δημιουργία και εκπαίδευση ενός ταξινομητή εξαρτάται απο τα αποτελέσματα των ταξινομητών που έχουν ήδη εκπαιδευτεί. Έτσι, γίνεται προσπάθεια εκμετάλλευσης της πληροφορίας που πιθανώς έχει ήδη εντοπιστεί για το συγκεκριμένο πρόβλημα.

Εκτός απο αυτή τη γενική ιδέα, οι εξαρτώμενοι αλγόριθμοι μάθησης συνόλου μπορούν να χωριστούν και σε δύο περεταίρω κατηγορίες, με βάση την προσέγγιση που ακολουθείται για την μετάδοση της γνώσης σε επόμενους ταξινομητές:

- Incremental Batch Learning: Οι προβλέψεις του προηγούμενου σε σειρά εκπαίδευσης ταξινομητή προσδίδονται στον επόμενο ως ήδη υπάρχουσα γνώση, ενώ ο επόμενος σε σειρά ταξινομητής χρησιμοποιεί το αρχικό σύνολο εκπαίδευσης σε συνδυασμό με τη γνώση αυτή για την εκπαίδευση του. Με αυτό τον τρόπο κατασκευάζεται αυξητικά ένας “όλο και καλύτερος” ταξινομητής. Συνεπώς, ο τελευταίος ταξινομητής που θα εκπαιδευτεί αποτελεί και τον ταξινομητή του οποίου οι έξοδοι θεωρούνται οι τελικές προβλέψεις του συστήματος για δεδομένη είσοδο.
- Model-guided Instance Selection: Σε αυτή την περίπτωση ή ήδη υπάρχουσα γνώση που θεωρούμε οτι έχει παραχθεί απο προηγούμενους ταξινομητές διαδίδεται στους επόμενους διαμέσου του συνόλου εκπαίδευσης. Έτσι, προηγούμενοι ταξινομητές τροποποιούν το σύνολο εκπαίδευσης κατάλληλα με σκοπόυ να καταστήσουν την εκπαίδευση επόμενων ταξινομητών πιο στοχευμένη. Μια συνήθης προσέγγιση είναι να αγνοούνται πρότυπα τα οποία έχουν ταξινομηθεί σωστά απο τον προηγούμενο ταξινομητή ωστέ να δωθεί περισσότερη έμφαση στα λάθος ταξινομημένα πρότυπα.

[10]

Παράδειγμα: Boosting

Χαρακτηριστική μεθοδολογία εξαρτώμενης μάθησης συνόλου είναι η τεχνική Boosting (ή αλλιώς arcing – Adaptive Resampling and Combining). Το Boosting αποτελεί Model-guided Instance Selection, δηλαδή βασίζεται στην κατασκευή πολλών ταξινομητών με κατάλληλη προσαρμογή του συνόλου εκπαίδευσης κατά τη διαδικασία εκπαίδευσης αυτών.
Η προσαρμογή αυτή γίνεται με κριτήριο την εκπαίδευση επόμενων ταξινομητών σε περιπτώσεις (δηλαδή πρότυπα εισόδου) στις οποίες οι ως τώρα εκπαιδευμένοι ταξινομητές δεν έχουν καλή απόδοση. Αυτό πρακτικά σημαίνει πως τα σύνολα εκπαίδευσης επόμενων ταξινομητών αποτελούνται κυρίως απο πρότυπα για τα οποία προηγούμενοι ταξινομητές είχαν εσφαλμένη πρόβλεψη. [12]

Ένας βασικός αλγόριθμος που βασίζεται στο Boosting αλλά εισάγει και κάποιες βελτιώσεις είναι το AdaBoost (Adaptive Boosting). Η βασική ιδέα του AdaBoost είναι προσθήκη βαρών σημασίας σε κάθε πρότυπο, με βάση τα οποία ορίζεται ο βαθμός και η σημασία συμμετοχής καθενός στην εκπαίδευση του κάθε ταξινομητή. Έτσι, ενώ αρχικά όλα τα πρότυπα εκπαίδευσης έχουν το ίδιο βάρος, τα βάρη αυτά αυξάνονται κάθε φορά που το αντίστοιχο πρότυπο ταξινομείται λάθος και μειώνονται όταν το αντίστοιχο πρότυπο ταξινομείται σωστά. Με αυτό τον τρόπο το AdaBoost εξασφαλίζει πως επόμενοι ταξινομητές θα στοχεύσουν περισσότερο στα “δύσκολα” πρότυπα που δεν ταξινομούνται σωστά από τους ως τώρα ταξινομητές.

Όσον αφορά τον συνδυασμό των διαφόρων αποτελεσμάτων, το AdaBoost αναθέτει ένα βάρος σε κάθε ταξινομητή ανάλογα με την ακρίβεια των προβλέψεων του, σε συνάρτηση με το βάρος των προτύπων που ταξινόμησε σωστά. Έτσι τελικά το βάρος κάθε ταξινομητή ορίζει την ισχύ που έχει η πρόβλεψη του στον τελικό υπολογισμό της εξόδου του συνολικού

συστήματος μάθησης συνόλου.

Το βασικό πρόβλημα που μπορεί να οδηγήσει σε μειώμενη απόδοση του αλγορίθμου AdaBoost είναι η υψηλή πιθανότητα εμφάνισης του φαινομένου Over-fitting. Αυτό προκύπτει ως άμεσο επακόλουθο της διαδικασίας “προσαρμογής” στα πιο δύσκολα πρότυπα εισόδου που περιγράφηκε παραπάνω. Ωστόσο, αυτό το πρόβλημα μπορεί να αντιμετωπιστεί σε περιορισμένο βαθμό μειώνοντας τον αριθμό επαναλήψεων (δηλαδή εκπαιδευόμενων ταξινομητών) του αλγορίθμου. [13] [14]

3.4 Συνδυασμός Προβλέψεων

Εφόσον έχουν κατασκευαστεί και εκπαιδευτεί οι διάφοροι ταξινομητές, τελικό βήμα στις μεθόδους μάθησης συνόλου είναι ο συνδυασμός των προβλέψεων καθενός. Υπάρχουν διάφοροι τρόποι για να γίνει αυτό, ωστόσο σε πρώτη φάση ορίζονται δύο γενικές κατηγορίες μεθόδων συνδυασμού:

- Μέθοδοι Στάθμισης (Weighting methods), οπού χρησιμοποιείται η λογική ανάθεσης βάρους σε κάθε ταξινομητή, με αποτέλεσμα κάθε ταξινομητής να έχει ισχύ ανάλογη του βάρους αυτού στην τελική πρόβλεψη.
- Μέθοδοι Meta-Learning, οπού ο στόχος είναι η περετάιρω εκπαίδευση μετά την ολοκλήρωση της δημιουργίας ταξινομητών, ωστέ τελικά κατά την φάση της πρόβλεψης να επιλέγεται με ευφυή τρόπο ποιος ή ποιοι από τους εκπαιδευμένους ταξινομητές θα χρησιμοποιηθούν.

Ορισμένες τεχνικές στάθμισης που χρησιμοποιούνται συνήθως είναι οι εξής:

- Πλειοψηφία (Majority Voting): για δεδομένη είσοδο επιλέγεται η κλάση την οποία προβλέπει η πλειοψηφία των ταξινομητών. Αποτελεί μια βασική μέθοδο, η οποία είναι ασφαλής εφόσον δεν υπάρχει επιπλεόν γνώση με βάση την οποία να μπορεί γίνει διαλογή μεταξύ των διαφόρων ταξινομητών.
- Στάθμιση Επίδοσης (Performance Weighting): σε κάθε ταξινομητή ανατίθεται ένα βάρος με βάση την επίδοση του σε ένα σύνολο επικύρωσης. Έτσι έχουν μεγαλύτερη ισχύ οι ταξινομητές που θεωρείται οτι έχουν καλύτερη επίδοση, με βάση αυτό το σύνολο επικύρωσης.

Οι μέθοδοι Meta-Learning έχουν αρκετά πιο περίπλοκη δομή καθώς αφορούν την ευφυή επιλογή μεταξύ διαφόρων classifiers. Μια αρκετά βασική τεχνική Meta-Learning είναι η εξής:

- Stacking: Σε αυτή τη μέθοδο αρχικά δημιουργείται ένα νέο σύνολο δεδομένων, με αριθμό στοιχείων όμοιο με αυτόν του αρχικού συνόλου εκπαίδευσης. Οι γνωστές κατηγορίες για κάθε πρότυπο διατηρούνται σε αυτό το νέο σύνολο, ωστόσο αντί για τις μεταβλητές εισόδου χρησιμοποιούνται οι προβλέψεις κάθε ταξινομητή για το εκάστοτε πρότυπο. Αυτό το νέο σύνολο δεδομένων χρησιμοποιείται για την εκπαίδευση ενός νέου “μέτα-ταξινομητή” με βάση των οποίο τελικά γίνεται η πρόβλεψη. Πρακτικά, το αρχικό σύνολο δεδομένων χωρίζεται σε δύο υποσύνολα στην περίπτωση του Stacking, με το πρώτο να χρησιμοποιείται για την εκπαίδευση των αρχικών ταξινομητών και το δεύτερο για την εκπαίδευση του μέτα-ταξινομητή.

[10]

3.5 Ποικιλομορφία Ταξινομητών (Classifier Diversity)

Όπως προαναφέρθηκε, για να έχει νόημα η μάθηση συνόλου, θα πρέπει οι διάφοροι ταξινομητές που δημιουργούνται να έχουν ουσιαστικές διαφορές, ωστέ ο συνδυασμός αυτών να προσεγγίζει καλύτερα τυχόν ιδιομορφίες του προβλήματος.

Μια πρώτη κατηγοριοποιήση αφορά τον τρόπο με τον οποίο επιχειρείται η εισαγωγή ποικιλομορφίας. Έτσι, σε τεχνικές όπως το bagging, οπού με τρόπο τυχαίο δημιουργούνται διαφορετικά σύνολα δεδομένων για την εκπαίδευση των διαφόρων ταξινομητών, μπορούμε να θεωρήσουμε οτι η ποικιλομορφία είναι *έμμεση* (implicit) αφού δεν εξασφαλίζεται με συστηματικό τρόπο. Αντίθετα σε τεχνικές όπως το boosting η ποικιλομορφία εισάγεται με τρόπο συστηματικό και προς επίτευξη συγκεκριμένου στόχου, άρα μπορούμε να την θεωρήσουμε *άμεση* ή *ρητή* (explicit).

Έτσι, αν θεωρήσουμε πως γενικός στόχος της μάθησης συνόλου είναι οι διάφοροι ταξινομητές που θα εκπαιδευτούν να κατέχουν διαφορετικά σημεία στον χώρο υποθέσεων του προβλήματος (και άρα προσεγγίζουν τις διάφορες πτυχές του καλύτερα), η έμμεση ποικιλομορφία βασίζεται περισσότερο στην τυχαιότητα για την επίτευξη αυτού του στόχου ενώ η άμεση ποικιλομορφία επιλέγει με ντετερμινιστικό τρόπο τα σημεία αυτά. [15]

Με βάση αυτή την προσέγγιση, μπορούμε να θεωρήσουμε τρεις διαφορετικούς τρόπους εισαγωγής ποικιλομορφίας σε έναν αλγόριθμο μάθησης συνόλου:

1. Μεταβολή του σημείου έναρξης στο Χώρο Υποθέσεων (Starting point in Hypothesis Space)
2. Μεταβολή του συνόλου Προσβάσιμων Υποθέσεων (Set of Accessible Hypotheses)
3. Μεταβολή του Τρόπου Διάσχισης του Χώρου Υποθέσεων (Traversal of Hypothesis Space)

Κάθε μια απο αυτές τις μεταβολές επηρεάζει τα σημεία του Χώρου Υποθέσεων στα οποία θα συγκλίνει κάθε ταξινομητής που εκπαιδεύεται. [15]
Με αυτό τον τρόπο, κάθε ταξινομητής ουσιαστικά προσεγγίζει το πρόβλημα απο διαφορετική οπτική γωνία.

Πρακτικά, οι παραπάνω θεωρητικές προσεγγίσεις μπορούν να εφαρμοστούν με τις παρακάτω τεχνικές:

- Μεταβολή του συνόλου εκπαίδευσης: Κάθε ταξινομητής εκπαιδεύεται με διαφορετικό σύνολο δεδομένων, συνήθως κάποιο υποσύνολο του αρχικού συνόλου εκπαίδευσης.
- Μεταβολή του επαγωγέα βάσης: Ο τρόπος με τον οποίο χρησιμοποιείται ο επαγωγέας βάσης μεταβάλλεται για κάθε ταξινομητή. Αυτή η μεταβολή μπορεί για παράδειγμα να αφορά παραμέτρους όπως τα αρχικά βάρη ενός τεχνητού νευρωνικού δικτύου (δηλαδή μεταβολή του σημείου έναρξης στο χώρο υποθέσεων).
- Μεταβολή του Τρόπου Αναπαράστασης της Μεταβλητής Εξόδου: Σε αυτή την περίπτωση αντί να χρησιμοποιηθεί ένας μοναδικός, ενδεχομένως πολύπλοκος ταξινομητής, χρησιμοποιούνται πολλοί ταξινομητές, καθένας απο τους οποίους αντιμετωπίζει το πρόβλημα θεωρώντας μια απλούστερη αναπαράσταση της μεταβλητής εξόδου. Με αυτό τον τρόπο μπορεί να μειωθεί η διαστατικότητα και η πολυπλοκότητα του κάθε υπο-προβλήματος που προκύπτει.
- Διαμέριση του Χώρου Αναζήτησης: Καθένας απο τους ταξινομητές εκπαιδεύεται ώστε να αντιμετωπίζει ένα μέρος του προβλήματος. Ο αρχικός χώρος αναζήτησης του προβλήματος χωρίζεται σε υπο-χώρους, οι οποίοι μπορεί να έχουν αυστηρά ή χαλαρά όρια, ενώ καθένας απο αυτούς τους υπο-χώρους ανατίθεται σε έναν

ταξινομητή. Για παράδειγμα, στην περίπτωση του προβλήματος χωρικής παρεμβολής, ένας υπο-χώρος μπορεί να είναι μια υπο-περιοχή της ολικής περιοχής που αφορά το υπό μελέτη πρόβλημα.

- Υβριδοποίηση: Αφορά την περίπτωση οπού ο βασικός επαγωγέας δεν είναι μοναδικός, οπότε εκπαιδεύονται ταξινομητές που βασίζονται σε διαφορετική μέθοδο μηχανικής μάθησης ή και μέθοδο μάθησης συνόλου.

3.6 Χρήση μάθησης συνόλου για το πρόβλημα της Παλινδρόμησης

Η προσαρμογή των παραπάνω τεχνικών για την αντιμετώπιση του προβλήματος της παλινδρόμησης είναι απόλυτα λογική. Έτσι:

- Οι αλγόριθμοι μάθησης που αποτελούν τον πυρήνα της μάθησης συνόλου χρησιμοποιούν την μορφή που αφορά το πρόβλημα της παλινδρόμησης. Για παράδειγμα, χρησιμοποιούνται δέντρα απόφασης για παλινδρόμηση στη θέση των δέντρων απόφασης για ταξινόμηση.
- Οι τεχνικές εισαγωγής ποικιλομορφίας μπορούν να εφαρμοστούν άμεσα και στην περίπτωση της παλινδρόμησης, καθώς αποτελούν μεθοδολογίες που περιγράφουν την προσέγγιση στο συγκεκριμένο πρόβλημα.
- Οι τεχνικές συνδυασμού των αποτελεσμάτων εφαρμόζονται με τρόπο αντίστοιχο με αυτό της ταξινόμησης, με τη διαφορά οτι η διαδικασία που χρησιμοποιείται αντί για την ψηφοφορία της ταξινόμησης είναι η λήψη μέσου όρου, με ή χωρίς κατάλληλα σταθμισμένα βάρη, ανάλογα με τη μεθοδολογία που ακολουθείται κάθε φορά.

4 Τυχαία Δάση

Σε αυτό το κεφάλαιο παρουσιάζεται ο αλγόριθμος μηχανικής μάθησης Τυχαίων Δασών (Random Forests), όπως αναλύθηκε απο τον L. Breiman. Εξετάζονται επίσης κάποια πρακτικά ζητήματα του αλγορίθμου, ενώ παρουσιάζεται η συμπεριφορά του σε σχέση με άλλους κλασικούς αλγορίθμους μάθησης συνόλου. Στο τέλος του κεφαλαίου παρουσιάζεται επίσης και ο τρόπος συνδυασμού αυτού του αλγορίθμου με κλασικές τεχνικές γεωστατιστικής για την επίτευξη καλύτερων αποτελεσμάτων στο πρόβλημα της χωρικής παρεμβολής, όπως πραγματοποιήθηκε στην μελέτη των Li et al. [16]

4.1 Η Βασική Ιδέα πίσω απο τα Τυχαία Δάση

Τα δέντρα απόφασης, όπως παρουσιάστηκαν σε προηγούμενο κεφάλαιο, αποτελούν δημοφιλείς ταξινομητές καθώς συνδυάζουν γρήγορη ταχύτητα εκτέλεσης με μια προσέγγιση που περιορίζει σε ένα βαθμό τη λογική "μαύρου κουτιού", κάτι το οποίο είναι επιθυμητό σε πολλές περιπτώσεις καθώς δεν αρκεί η απλή ακρίβεια προβλέψεων αλλά απαιτείται και κατανόηση της δομής που εκφράζει ένα πρόβλημα.

Ωστόσο, σύμφωνα με τον Ho [17] η χρήση των παραδοσιακών μεθόδων κατασκευής δέντρων απόφασης είναι προβληματική καθώς η δημιουργία δέντρων μεγάλου μεγέθους και πολυπλοκότητας συνήθως οδηγεί σε μείωση της ακρίβειας γενίκευσης. Στον αντίποδα, μείωση αυτής της πολυπλοκότητας οδηγεί σε μειωμένη ακρίβεια στα δεδομένα εκπαίδευσης.

Έχουν προταθεί διάφορες ευριστικές μέθοδοι που στοχεύουν στην αντιμετώπιση αυτού του προβλήματος μέσω επίτευξης της βέλτιστης δυνατής ακρίβειας με βάση τα δεδομένα ή περιορισμού του μεγέθους του δέντρου, ωστόσο η κατασκευή δέντρων απόφασης με το ίδιο σύνολο δεδομένων δεν φαίνεται να μπορεί να ξεφύγει από την τάση over-fitting σε αυτό το σύνολο εκπαίδευσης.

Τέτοιες μέθοδοι είναι ή μέθοδος pruning ενός μεγάλου σε μέγεθος δέντρου σε μικρότερο με σκοπό την μείωση του σφάλματος γενίκευσης, προκαλώντας όμως μείωση της απόδοσης στα δεδομένα εκπαίδευσης ή η χρήση πιθανοτικών μεθόδων επιλογής περισσότερων απο μίας διαδρομής διακλαδώσεων-αποφάσεων στο δέντρο απόφασης, με διαφορετικό βαθμό εμπιστοσύνης για κάθεμια.

Εντούτοις γενικώς εμφανίζεται το πρόβλημα ενός ανώτατου ορίου στην πολυπλοκότητα του δέντρου απόφασης που κατασκευάζεται: αν ένα δέντρο ξεπερνάει αυτό το όριο, εμφανίζεται over-fitting στο σύνολο εκπαίδευσης και η μετρική σφάλματος που υπολογίζεται κατά την εκπαίδευση δεν αντιπροσωπεύει την πραγματική ικανότητα γενίκευσης του ταξινομητή.

Σύμφωνα με τον Ho, για την αντιμετώπιση αυτού του προβλήματος είναι απαραίτητη η χρήση πολλών δέντρων απόφασης, καθένα απο τα οποία έχει την ικανότητα να γενικεύσει ανεξάρτητα απο τα υπόλοιπα, ενώ είναι επίσης απαραίτητη η χρήση μιας συνάρτησης που συνδυάζει τις προβλέψεις από κάθε δέντρο με τρόπο που διατηρεί την ακρίβεια καθενός. [17]
Η αντίστοιχη ανάλυση του Breiman [18] καταλήγει στο γεγονός πως το σφάλμα γενίκευσης ενός συστήματος μάθησης συνόλου αποτελούμενου από δέντρα απόφασης (ή αλλιώς, δάσος-forest) συνδέεται άμεσα με την ισχύ (strength) κάθε ταξινομητή-δέντρου που το απαρτίζει καθώς και στην μεταξύ τους συσχέτιση. Ως ισχύς (strength) ενός δέντρου απόφασης ορίζεται η διαφορά μεταξύ της πιθανότητας αυτού να προβλέψει την σωστή έξοδο από την πιθανότητα να προβλέψει οποιαδήποτε άλλη έξοδο για δεδομένη είσοδο. Έτσι, ο Breiman υποστηρίζει πως το ζητούμενο είναι η κατασκευή δασών που αποτελούνται από δέντρα απόφασης υψηλής ισχύος, τα οποία όμως έχουν μικρό βαθμό συσχέτισης.

Η αντιμετώπιση που προτείνεται απο τον Ho είναι η δημιουργία δέντρων καθένα απο τα οποία χρησιμοποιεί για την εκπαίδευση του ένα τυχαίο υποσύνολο των στοιχείων εισόδου. Έτσι, αν η είσοδος χαρακτηρίζεται απο μεταβλητές $(x_1, x_2, \ldots, x_n)$ ένα δέντρο μπορεί να εκπαιδευτεί χρησιμοποιώντας μόνο τις μεταβλητές $(x_2, x_5, \ldots x_{n-2})$ ενώ ένα άλλο δέντρο τις μεταβλητές $(x_1, x_4, \ldots, x_n)$. Συγκεκριμένα, το χαρακτηριστικό που θεωρείται χρήσιμο είναι η ύπαρξη 2^m υποσυνόλων για διάνυσμα εισόδου διάστασης m . Η χρήση τυχαιότητας αποτελεί έναν βολικό τρόπο διάσχισης αυτού του χώρου.

4.2 Η κατά Breiman προσέγγιση των Τυχαίων Δασών

Οι βασικές ιδέες που έχουν προταθεί και χρησιμοποιηθεί για την βελτίωση της απόδοσης συστημάτων μάθησης συνόλου με δέντρα απόφασης είναι το Bagging, το Boosting, αλλά και η επιλογή Τυχαίων Μεταβλητών εισόδου (Random Feature Selection), όπως αναφέρθηκε παραπάνω.

Τα Τυχαία Δάση (Random Forests) που προτείνονται απο τον Breiman [18] στην ουσία συνδυάζουν την λογική του Bagging με το Random Feature Selection. Έτσι στην πράξη, κάθε νέο σύνολο εκπαίδευσης (που θα χρησιμοποιηθεί για την δημιουργία του επόμενου δέντρου απόφασης στο δάσος) επιλέγεται ως υποσύνολο του αρχικού συνόλου εκπαίδευσης, με αντικατάσταση (δηλαδή η μεθοδολογία του Bagging), ωστόσο στη συνέχεια η εκπαίδευση του δέντρου γίνεται με Random Feature Selection απο αυτό το υποσύνολο. Σημειώνεται επίσης πως δεν εφαρμόζεται pruning των δέντρων απόφασης που δημιουργούνται.

Η χρήση του Bagging συνήθως οδηγεί σε αύξηση της ακρίβειας στην περίπτωση επιλογής τυχαίων μεταβλητών εισόδου. Επιπλέον, χρησιμοποιώντας την τεχνική του Bagging γίνεται δυνατή η χρήση μιας μετρικής του σφάλματος γενίκευσης του συνολικού συστήματος μάθησης συνόλου καθώς και των δύο παραμέτρων που θεωρούνται σημαντικές για την μείωση του σφάλματος αυτού, δηλαδή της ισχύος ταξινομητή (classifier strength) και της μεταξύ τους συσχέτισης (correlation). Οι μετρικές αυτές ορίζονται ακολουθώντας λογική out-of-bag, ως εξής:

Έστω το αρχικό σύνολο εκπαίδευσης T . Με βάση αυτό φτιάχνουμε τα υποσύνολα εκπαίδευσης T_k , τα οποία χρησιμοποιούνται για την εκπαίδευση κάποιων ταξινομητών $h(x, T_k)$. Σε αυτό τον συμβολισμό οι ταξινομητές ξεχωρίζουν μεταξύ τους λόγω του διαφορετικού συνόλου T_k που χρησιμοποιήθηκε για την εκπαίδευση καθενός, ενώ το x συμβολίζει το διάνυσμα εισόδου.

Για τον υπολογισμό της τελικής εξόδου – πρόβλεψης y για δεδομένη είσοδο x , χρησιμοποιείται η πλειοψηφία των προβλέψεων όλων των ταξινομητών $h(x, T_k)$. Για όλους τους συνδυασμούς γνωστής εισόδου – εξόδου που αποτελούν το σύνολο T , συγκεντρώνουμε την απόκριση του συνολικού συστήματος χρησιμοποιώντας όμως τις προβλέψεις μόνο των ταξινομητών που δεν εκπαιδεύτηκαν χρησιμοποιώντας το αντίστοιχο ζευγάρι x, y , η αλλιώς για τους οποίους ισχύει $\{x, y\} \notin T_k$. Έτσι σχηματίζεται ο ταξινομητής out-of-bag (out-of-bag classifier), ο οποίος αποτελεί προσπάθεια προσέγγισης της επίδοσης του συστήματος σε πρότυπα με τα οποία δεν έχει εκπαιδευτεί. Αντίστοιχα, το ποσοστό σφάλματος του ταξινομητή out-of-bag αποτελεί το out-of-bag εκτιμώμενο ποσοστό σφάλματος γενίκευσης του συστήματος μάθησης συνόλου.

Η χρησιμότητα της εκτίμησης out-of-bag είναι οτι επιτρέπει την εκτίμηση του σφάλματος γενίκευσης (και άρα μια χρήσιμη μετρική για την εκπαίδευση του συστήματος) χωρίς την ανάγκη χρήσης ενός εξειδικευμένου συνόλου επικύρωσης (validation set) όπως για παράδειγμα στην περίπτωση των τεχνητών νευρωνικών δικτύων. Επιπλέον, αν και δεν έχουν την βέλτιστη απόδοση ως εκτιμητές απόδοσης, η οποία αποδίδεται στην τεχνική Cross-Validation, η απόδοση τους την πλησιάζει σε ικανοποιητικό βάθμο, ενώ έχει το παραπάνω πλεονέκτημα οτι μπορούν να υπολογιστούν παράλληλα με τη διαδικασία εκπαίδευσης με ελάχιστη άυξηση του υπολογιστικού κόστους. [19]

4.3 Πρακτικά Ζητήματα Εφαρμογής των Τυχαίων Δασών

Όσον αφορά την πρακτική εφαρμογή της τεχνικής Random Feature Selection (ή αλλιώς Feature-Bagging) του αλγορίθμου των Random Forests, ο Breiman [18] ορίζει δύο γενικές τεχνικές, τα Forest-RI και τα Forest-RC:

Forest-RI: Αποτελεί τον απλούστερο τρόπο δημιουργίας Random Forests με Random Feature Selection, οπού για κάθε δέντρο επιλέγεται σε κάθε διαχωρισμό (split), με τυχαίο τρόπο, ένα μικρό σύνολο μεταβλητών εισόδου (features) στα οποία θα εφαρμοστεί ο διαχωρισμός. Το δέντρο απόφασης δημιουργείται με την μεθοδολογία που έχει αναλυθεί παραπάνω, στο μέγιστο δυνατό μέγεθος, και δεν εφαρμόζεται pruning. Ορίζεται επίσης η παράμετρος F η οποία εκφράζει το μέγεθος του μικρού συνόλου μεταβλητών εισόδου στα οποία εφαρμόζεται ο διαχωρισμός. Η παράμετρος F θεωρείται σταθερή.

Σημειώνεται πως, με βάση τα αποτελέσματα των πειραμάτων του Breiman, η διαδικασία δεν φαίνεται να επηρεάζεται σε μεγάλο βαθμό απο αλλαγές στην τιμή F . Συγκεκριμένα, η μέση απόλυτη διαφορά μεταξύ του ποσοστού σφάλματος για τιμή $F=1$ και για μεγαλύτερες τιμές της παραμέτρου F είναι μικρότερη απο 1% . Επιπλέον, η διαφορά αυτή γίνεται αισθητή κυριώς σε μεγαλύτερα σύνολα δεδομένων.

Forest-RC: Αποτελεί έναν τρόπο αντιμετώπισης της περίπτωσης οπού ο αριθμός μεταβλητών εισόδου είναι σχετικά μικρός. Αυτό αποτελεί πρόβλημα διότι, ορίζοντας τιμή F η οποία αποτελεί υπολογίσιμο κλάσμα του αριθμού μεταβλητών εισόδου M μπορούμε απο τη μία να έχουμε άυξηση της ισχύος κάθε δέντρου, ωστόσο αυξάνεται και η συσχέτιση μεταξύ αυτών.Ο τρόπος με τον οποίο αντιμετωπίζεται αυτό είναι η παραγωγή περισσότερων μεταβλητών εισόδου, χρησιμοποιώντας συνδυασμούς των αρχικών μεταβλητών.

Αυτό γίνεται με χρήση τυχαίων γραμμικών συνδυασμών των αρχικών μεταβλητών: Ορίζεται η τιμή L , δηλαδή ο αριθμός μεταβλητών προς συνδυασμό. Σε δεδομένο κόμβο του δέντρου απόφασης, επιλέγονται τυχάια L μεταβλητές εισόδου, οι οποίες προστίθενται με συντελεστές οι οποίοι είναι ομοιόμορφα κατανεμημένοι τυχάιοι αριθμοί οι οποίοι ανήκουν στο πεδίο $[-1, 1]$. Με αυτό τον τρόπο παράγονται F γραμμικοί συνδυασμοί μεταβλητών, και η τελική αναζήτηση καλύτερου διαχωρισμού γίνεται με βάση τους F αυτούς συνδυασμούς.

Μια περίπτωση που εμφανίζει πρόβλημα τόσο για τα Forest-RI αλλά κυρίως για τα Forest-RC είναι η περίπτωση των κατηγορικών μεταβλητών, δηλαδή των μεταβλητών των οποίων οι τιμές εκφράζουν κατηγορίες, όπως για παράδειγμα *colour = { blue , red , green }.* Το ζητούμενο είναι η εύρεση κάποιας τεχνικής που θα επιτρέπει τον γραμμικό συνδυασμό αυτών των μεταβλητών με τις αριθμητικές μεταβλητές.

Η προσέγγιση του Breiman είναι η εξής: Εφόσον μια κατηγορική μεταβλητή επιλέγεται για να συμμετέχει σε κάποιο διαχωρισμό, επιλέγεται ένα τυχαίο υποσύνολο των πιθανών κατηγοριών αυτής της μεταβλητής και ορίζεται μια νέα βοηθητική μεταβλητή η οποία έχει την τιμή “1” αν το αντίστοιχο πρότυπο έχει τιμή κατηγορίας που ανήκει στο τυχαίο υποσύνολο που επιλέχτηκε και “0” αν όχι.
Άμεσο αποτέλεσμα της πρακτικής αυτής είναι η άυξηση της πιθανότητας επιλογής μιας κατηγορικής μεταβλητής κατά το διαχωρισμό, καθώς μια κατηγορική μεταβλητή “ξεδιπλώνεται” σε πολλές περισσότερες, ενώ οι αριθμητικές μεταβλητές δεν προσαρμόζονται κάπως σε αυτή την αλλαγή. Επομένως, είναι σημαντικό να χρησιμοποιηθεί μεγαλύτερη τιμή F σε αυτές τις περιπτώσεις, για να αποφευχθούν πιθανές πτώσεις στη ισχύ του κάθε ταξινομητή.

4.4 Τυχαία Δάση και Παλινδρόμηση (Regression)

Η προσέγγιση του προβλήματος της παλινδρόμησης με χρήση Random Forests είναι ακριβώς αντίστοιχη με την προσέγγιση του προβλήματος ταξινόμησης. Έτσι, εφαρμόζεται πάλι τεχνική Bagging σε συνδυασμό με Random Feature Selection (Feature Bagging). Οι διαφορές εμφανίζονται στις τεχνικές που εφαρμόζονται στον πυρήνα του αλγορίθμου.

Στην παλινδρόμηση θεωρείται οτι η δημιουργία των δέντρων απόφασης βασίζεται σε ένα τυχαίο διάνυσμα Θ (αντί του T_k προηγουμένως) ωστέ να δημιουργείται ένας tree predictor $h(x, \Theta)$ ο οποίος σε αυτή την περίπτωση προβλέπει (δηλαδή έχει έξοδο) αριθμητικές τιμές αντί για κατηγορίες.

Η έξοδος του συστήματος μάθησης συνόλου για δεδομένη είσοδο προκύπτει ως ο μέσος όρος των k συνολικά ταξινομητών της μορφής $h(x, \Theta_k)$ που δημιουργούνται.

Οι εκτιμήσεις out-of-bag προκύπτουν με τρόπο αντίστοιχο με αυτό της ταξινόμησης. Η διαφορά έγκειται πάλι στο γεγονός οτι πλέον αντιμετωπίζονται συνεχείς μεταβλητές και όχι διακριτές κατηγορίες.

Σε αντίθεση με την περίπτωση της ταξινόμησης, η παλινδρόμηση απαιτεί σχετικά μεγάλη τιμή F για την επίτευξη αποδεκτής ακρίβειας. Αυτό αποδίδεται στο γεγονός πως η συσχέτιση μεταξύ δέντρων αυξάνει με αργό ρυθμό καθώς αυξάνει η τιμή F , σε αντίθεση με την περίπτωση της ταξινόμησης.

4.5 Τυχαία Δάση και Προβληματικές Περιπτώσεις

Στη συγκεκριμένη ενότητα παρουσιάζεται συνοπτικά η απόδοση του αλγορίθμου Random Forests σε περιπτώσεις που παραδοσιακά θεωρούνται δύσκολες για τους αλγορίθμους μηχανικής μάθησης. Η πειραματική ανάλυση αυτών των περιπτώσεων προέρχεται από τον Breiman. [18]

Θόρυβος δεδομένων

Η ύπαρξη θορύβου στο σύνολο εκπαίδευσης είναι μια απο τις πιο συνηθισμένες προβληματικές περιπτώσεις σχετικά με την μηχανική μάθηση. Ο θόρυβος αποτελεί ένα τυχαίο σφάλμα διακύμανσης της μετρούμενης μεταβλητής, και τα σύνολα δεδομένων που αφορούν περιβαλλοντικές εφαρμογές είναι συνήθως θορυβώδη. [20]
Συνεπώς, η ικανότητα ανέχειας στο θόρυβο (noise robustness) αποτελεί ένα επιθυμητό χαρακτηριστικό για το μελετούμενο πρόβλημα της χωρικής παρεμβολής.

Σύμφωνα με την μελέτη του Kalapanidas [20] , στην περίπτωση της παλινδρόμησης φαίνεται πως η γραμμική παλινδρόμηση καταφέρνει να αντιμετωπίσει την εισαγωγή θορύβου καλύτερα απο τους άλλους αλγορίθμους που μελετούνται. Ωστόσο όπως παρουσιάστηκε σε προηγούμενο κεφάλαιο, η γραμμική παλινδρόμηση αποτελεί μια αρκετά απλοϊκή προσέγγιση στο πρόβλημα της χωρικής παρεμβολής, επομένως είναι επιθυμητή η χρήση κάποιου ευφυέστερου αλγόριθμου, ο οποίος να εμφανίζει και ικανοποιητική ανέχεια στο θόρυβο.

Η μελέτη του Breiman επικεντρώνεται στη σύγκριση μεταξύ Random Forests και AdaBoost σε ο,τι αφορά την ανέχεια στο θόρυβο. Σε αυτή τη σύγκριση τα Random Forests, τόσο στην προσέγγιση Forest-RI όσο και στην προσέγγιση Forest-RC εμφανίζουν αισθητά καλύτερη ανέχεια στο θόρυβο.

Αν και αυτή η μελέτη αυτή δεν μπορεί να συγκριθεί άμεσα με την προαναφερθείσα, είναι ενδεικτική της καλύτερης συμπεριφοράς των Random Forests στην περίπτωση ύπαρξης θορύβου.

Δεδομένα με Πολλές Αδύναμες Εισόδους

Αποτελεί μια περίπτωση συνόλων δεδομένων που εμφανίζεται σε όλο και περισσότερα προβλήματα. Σε αυτές τις περιπτώσεις το διάνυσμα εισόδου αποτελείται απο έναν μεγάλο αριθμό μεταβλητών, κάθεμια από τις οποίες θεωρείται “αδύναμη” υπό την έννοια οτι δεν προσδίδει αρκετή πληροφορία για την αντιμετώπιση του προβλήματος απο μόνη της. Τέτοια σύνολα δεδομένων εμφανίζονται συχνά στους τομείς της ιατρικής διάγνωσης ή της ανάκτησης εγγράφων. Αυτή η μορφή δεδομένων αποτελεί μια ιδιαίτερα δύσκολη περίπτωση για τους συνήθεις ταξινομητές μηχανικής μάθησης όπως τα τεχνητά νευρωνικά δίκτυα ή τα δέντρα απόφασης.

Σε αυτή τη μελέτη, τα πειράματα του Breiman έδωσαν ποσοστά σφάλματος τα οποία δεν απείχαν ιδιαίτερα απο τα ποσοστά σφάλματος Bayes (τα οποία αποτελούν τα θεωρητικά βέλτιστα που μπορούν να επιτευχθούν). Έτσι εκφράζεται η ισχύς των Random Forests ως αλγόριθμος μάθησης συνόλου: καταφέρνει ικανοποιητικά ποσοστά σφάλματος, χρησιμοποιώντας ένα μεγάλο πλήθος πιθανώς αδύναμων ταξινομητών.

Ενδεικτικά, η τεχνική AdaBoost δεν ήταν δυνατό να εκτελεστεί στο σύνολο δεδομένων που παρήχθηκε για τη συγκεκριμένη μελέτη απόδοσης. [18]

4.6 Συνδυασμός Μεθόδων μέσω των Υπολοίπων (Residuals)

Στα πλαίσια της σύγκρισης μεταξύ διαφόρων μεθόδων αντιμετώπισης του προβλήματος της χωρικής παρεμβολής, χρησιμοποιήθηκε και ο συνδυασμός της τεχνικής των Random Forests με τεχνικές γεωστατιστικής. Πιο συγκεκριμένα, χρησιμοποιήθηκαν οι συνδυασμοί Random Forests – Inverse Distance Squared (RFIDS) και Random Forests – Ordinary Kriging (RFOK).

Τυπικά, ο συνδυασμός δύο μεθόδων που αφορούν το πρόβλημα της παλινδρόμησης γίνεται χρησιμοποιώντας τις τιμές Residuals. Οι τιμές αυτές εκφράζουν την διαφορά της προβλεπόμενης από την πραγματική τιμή για δεδομένο σημείο, δηλαδή:

$$e = y - \hat{y}$$

Χρησιμοποιώντας αυτές τις τιμές, είναι δυνατό να ερευνήσουμε αν το μοντέλο με βάση το οποίο υπολογίστηκαν επιδέχεται βελτίωση. Εναλλακτικά, οι residual τιμές παρέχουν εκτίμηση της πληροφορίας που δεν έχει εκμεταλλευτεί το μοντέλο.

Ο μαθηματικός τρόπος υπολογισμού και εκτίμησης αυτής της περίπτωσης γίνεται τυπικά μέσω του συντελεστή συσχετισμού (Correlation Coefficient) μεταξύ των αρχικών τιμών εξόδου και των residuals που προκύπτουν. Έτσι, αν προκύψει υψηλός βαθμός συσχέτισης μεταξύ των δύο, μπορούμε να θεωρήσουμε πως το μοντέλο σφάλλει κατά τρόπο προβλέψιμο, και άρα επιδέχεται βελτίωση.

Στην συγκεκριμένη περίπτωση, η βελτίωση αυτή εφαρμόζεται μέσω της εισαγωγής δεύτερης μεθόδου για την κατασκευή του τελικού μοντέλου επίλυσης της χωρικής παρεμβολής. Έτσι:

1. Εκπαιδεύεται το μοντέλο με με την μεθοδολογία Random Forests, με τρόπο ακριβώς αντίστοιχο με την περίπτωση εφαρμογής μόνο Random Forests.
2. Υπολογίζονται οι προβλέψεις του μοντέλου αυτού για τα σημεία που απαρτίζουν το σύνολο εκπαίδευσης. Δεδομένου οτι τα Random Forests δεν ειναι ακριβής μέθοδος, οι προβλέψεις αυτές είναι εν γένει διαφορετικές απο τις αρχικές τιμές.
3. Υπολογίζονται τα Residuals ως η διαφορά των αρχικών απο τις προβλεπόμενες τιμές
4. Κατασκευάζεται το μοντέλο με την δέυτερη μέθοδο (Inverse Distance Squared ή Ordinary Kriging) χρησιμοποιώντας ως σύνολο γνωστών σημείων τις θέσεις του συνόλου εκπαίδευσης σε συνδυασμό με τις αντίστοιχες τιμές Residuals που υπολογίστηκαν.
5. Για δεδομένη άγνωστη είσοδο, η πρόβλεψη του μοντέλου υπολογίζεται με βάση τόσο το μοντέλο Random Forests όσο και το δεύτερο μοντέλο, οι τιμές των οποίων συνδυάζονται.

5 Πρωτόκολλο Πειραμάτων

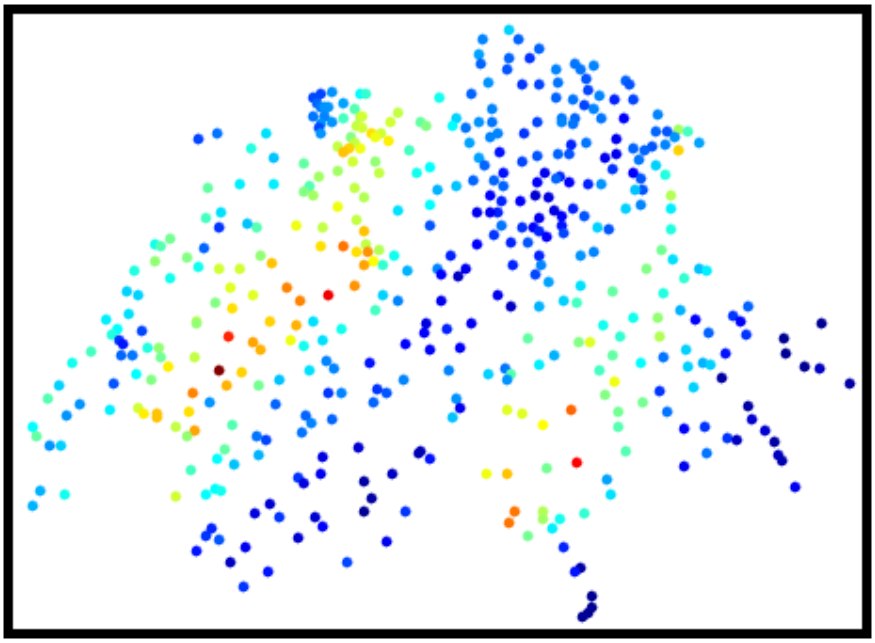
Σε αυτό το κεφάλαιο παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για τα πειράματα. Αναλύονται επίσης οι μετρικές με βάση τις οποίες συγκρίθηκε η απόδοση των διαφόρων τεχνικών χωρικής παρεμβολής, καθώς και η τεχνική Διασταυρούμενης Επικύρωσης (Cross-Validation), με βάση την οποία έγινε η επιλογή των υπερ-παραμέτρων που απαιτούνται για τις τεχνικές μηχανικής μάθησης.

5.1 Παρουσίαση Συνόλων Δεδομένων SIC (Spatial Interpolation Comparison)

Οι αρχικές μελέτες πραγματοποιήθηκαν με χρήση των συνόλων δεδομένων του επιστημονικού διαγωνισμού χωρικής παρεμβολής (Spatial Interpolation Comparison – SIC) για τις χρονίες 1997 και 2004 (SIC97 και SIC2004 αντίστοιχα). Πρόκειται για τα σύνολα δεδομένων που χρησιμοποιήθηκαν από τους Gilardi & Bengio. [4][3]

Ο διαγωνισμός αυτός είναι προσανατολισμένος προς την χρήση τεχνικών χωρικής παρεμβολής με τελικό σκοπό την χρήση των πληροφοριών που θα προκύψουν στην διαδικασία λήψης αποφάσεων σχετικά με μια φυσική καταστροφή. Αυτό προκύπτει απο το γεγονός πως μετρήσεις που αφορούν μια φυσική καταστροφή (όπως για παράδειγμα, την απελευθέρωση ραδιενεργούς ρύπου στην ατμόσφαιρα) λαμβάνονται απο τα ήδη υπάρχονται δίκτυα μετρήσεων, και άρα προκύπτουν σε άμεση αναφορά με την γεωγραφική θέση. Έτσι, η διαδικασία αυτή εκφράζει άμεσα το πρόβλημα χωρικής παρεμβολής: χρήση ενός περιορισμού αριθμού γνωστών σημείων για τη κατά το δυνατόν βέλτιστη πρόβλεψη της συμπεριφοράς μιας χωρικής συνάρτησης. Για το συγκεκριμένη πρόβλημα, ιδιαίτερη σημασία έχει η πρόβλεψη σημείων οπού εμφανίζονται ακραίες τιμές (για παράδειγμα, υψηλή τιμή ραδιενέργειας) καθώς και το μέτρο αβεβαιότητας που συνδέεται με τις προβλέψεις αυτές. [21]

Πιο λεπτομερώς, το σύνολο δεδομένων SIC97 αποτελείται από 467 μετρήσεις που αφορούν την ημερήσια βροχόπτωση (σε 1/10 του mm) στην Ελβετία για την 8η Μαρτίου 1986. Απο αυτές τις 467 σε πλήθος μετρήσεις οι 100 χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης, ενώ οι υπόλοιπες 367 θεωρήθηκαν άγνωστες και χρησιμοποιήθηκαν για τον έλεγχο της απόδοσης των διαφόρων μεθόδων. Τα στοιχεία που δόθηκαν για τον διαγωνισμό SIC97 περιελάμβαναν και το υψόμετρο για κάθε σημείο, ωστόσο σύμφωνα με τους Gilardi & Bengio η χρήση αυτών δεν επιφέρει ιδιαίτερο πλεονέκτημα για κάποιες απο τις τεχνικές μηχανικής μάθησης. Έτσι, σε αντιστοιχία με αυτή την ανάλυση, δεν χρησιμοποιήθηκε για να υπάρχει πιο ξεκάθαρη σύγκριση μεταξύ των μεθόδων. [4]



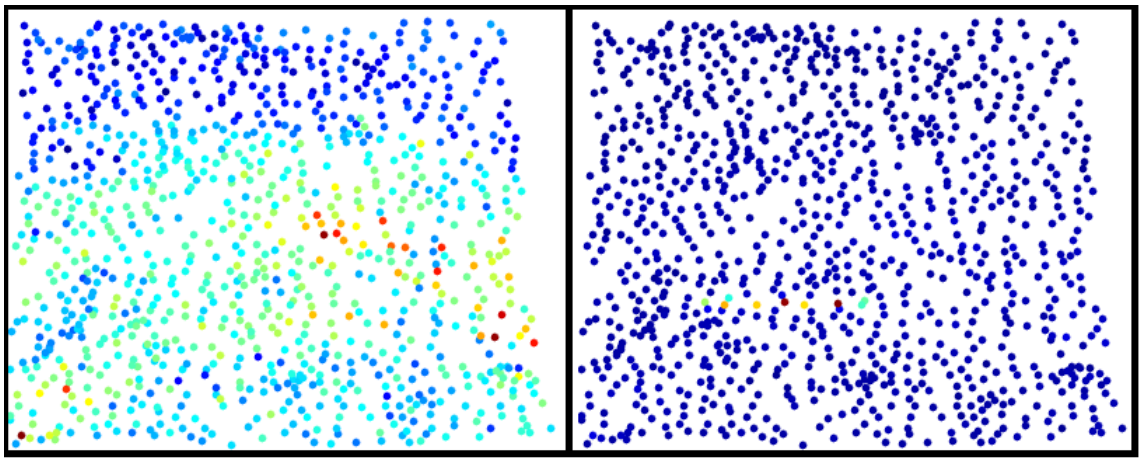
Το σύνολο δεδομένων SIC97

Το σύνολο δεδομένων SIC2004 αποτελείται από δύο περιπτώσεις μελέτης: την περίπτωση “Natural” και την περίπτωση “Joker”. Και στις δύο περιπτώσεις οι μετρήσεις αποτελούν την ημερήσια μέση τιμή της ακτινοβολίας γάμμα, υπολογισμένες σε διάφορα σημεία της περιοχής της Γερμανίας. Σημειώνεται πως οι μετρήσεις αυτές στην πράξη δεν έχουν ληφθεί την ίδια μέρα: πρόκειται για ένα τεχνητό σύνολο δεδομένων που δημιουργήθηκε για τις ανάγκες του διαγωνισμού. [22]

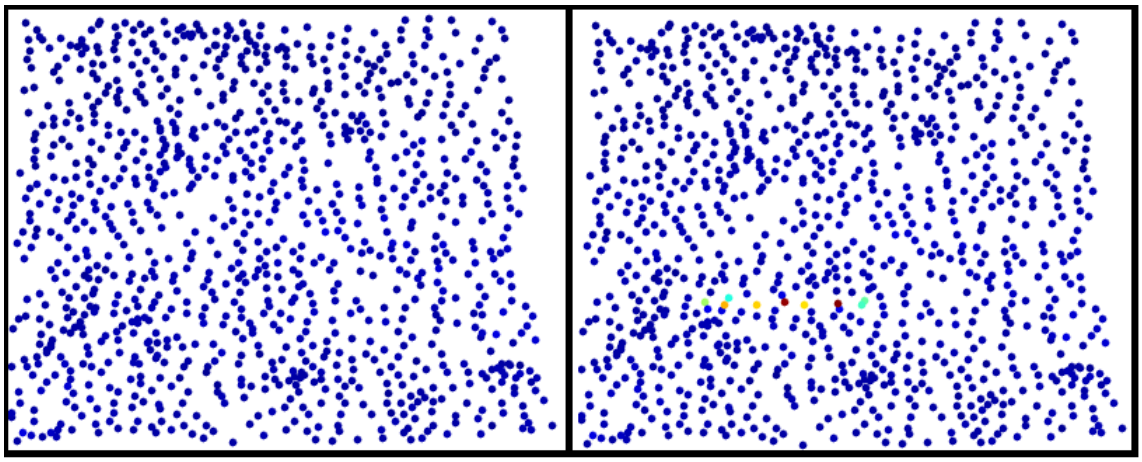
Στην περίπτωση “Natural” χρησιμοποιούνται οι πραγματικές μετρήσεις, όπως προέκυψαν απο τους σταθμούς μέτρησης. Στην περίπτωση “Joker” οι τιμές αυτές έχουν αλλοιωθεί, ωστέ να προσομοιωθέι ένα ραδιενεργό ατύχημα στη θέση “Joker” (Joker Spot).

Σημειώνεται επίσης πως, σε αντίθεση με τις ανάγκες του διαγωνισμού, τα σύνολα εκπάιδευσης και ελέγχου του διαγωνισμού αντιστρέφονται στην συγκεκριμένη μελέτη, κατ' αντιστοιχία με την μεθοδολογία των Gilardi & Bengio. [3]

Έτσι, ενώ στο διαγωνισμό χρησιμοποιήθηκε σύνολο εκπαίδευσης 200 σημείων και σύνολο ελέγχου 808 σημείων (τοσο για την περίπτωση “Natural” οσο και για την περίπτωση “Joker”), στην συγκεκριμένη μελέτη το σύνολο 808 σημείων θεωρείται σύνολο εκπαίδευσης και το σύνολο 200 σημείων σύνολο ελέγχου.



Τα σύνολα δεδομένων SIC04, Natural (αριστερά) και Joker (δεξιά). Σημειώνεται πως η κάθε χρωματική κλίμακα είναι προσαρμοσμένη στις τιμές του αντίστοιχου συνόλου.



Η παραπάνω απεικόνιση, με χρήση κοινής κλίμακας χρωμάτων και για τα δύο σύνολα δεδομένων. Χαρακτηριστική είναι η διαφορά των ακραίων τιμών του συνόλου Joker από όλες τις υπόλοιπες τιμές.

5.2 Μέθοδοι Εκτίμησης των Διαφόρων Μοντέλων

Για την εκτίμηση των διαφόρων μοντέλων και τη σύγκριση μεταξύ αυτών χρησιμοποιήθηκαν δύο διαφορετικές μέθοδοι εκτίμησης: το μέσο απόλυτο σφάλμα (Mean Absolute Error – MAE) και το μέσο τετραγωνισμένο σφάλμα (Mean Squared Error – MSE).

Τα σφάλματα αυτά ορίζονται τυπικά ως εξής:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Στην περίπτωση του *MSE* χρησιμοποιήθηκε στην πράξη η ρίζα της τιμής, δηλαδή η τιμή *RMSE* (Root Mean Squared Error). Το πλεονέκτημα αυτής της πρακτικής είναι ότι η τιμή *RMSE* εκφράζεται στην ίδια μονάδα με τις τιμές $y_i, \hat{y}_i$ επομένως δίνει μια αμεσότερη εικόνα για την επίδοση του μοντέλου.

Οι απόψεις σχετικά με την χρήση της μιας μεθόδου υπέρ της άλλης είναι διάφορες, ωστόσο η βασική διαφορά μεταξύ των δύο μεθόδων είναι η εξής: το *MAE* προσδίδει ίσο βάρος σε όλα τα σφάλματα κατά τον υπολογισμό της συνολικής επίδοσης, ενω το *RMSE* από την φύση της ύψωσης στο τετράγωνο δίνει μεγαλύτερο βάρος σε σφάλματα με μεγαλύτερη απόλυτη τιμή. Έτσι, εκ φύσεως η τιμή *RMSE* δεν προκύπτει ποτέ μικρότερη της τιμής *MAE* για το ίδιο σύνολο δεδομένων. [23]

Έτσι, στην περίπτωση που η διαφορά μεταξύ δύο μοντέλων δεν είναι ξεκάθαρη (δηλαδή δεν υπερτερεί κάποιο ως προς τις δύο μετρικές ταυτόχρονα), προκύπτει πως το μοντέλο με μικρότερη τιμή *MAE* προσεγγίζει το πρόβλημα με πιο ομαλό τρόπο συνολικά, ωστόσο το μοντέλο με μικρότερη τιμή *RMSE* προσεγγίζει τις ακραίες τιμές με μεγαλύτερη ακρίβεια.

Στη συγκεκριμένη μελέτη χρησιμοποιούνται και οι δύο μέθοδοι, ενώ δεν δημιουργείται ιδιαίτερος προβληματισμός για τις παραπάνω διαφορές καθώς βασικός στόχος είναι η εποπτική σύγκριση μεταξύ διαφόρων μεθόδων επίλυσης του προβλήματος χωρικής παρεμβολής.

5.3 Το πρωτόκολλο Διασταυρούμενης Επικύρωσης

Για να επιτευχθεί η καλύτερη απόδοση των διαφόρων τεχνικών για δεδομένο σύνολο δεδομένων είναι απαραίτητο να γίνει κατάλληλη επιλογή των υπερ–παραμέτρων (hyper-parameters) που σχετίζονται με κάθε μέθοδο.

Η επιλογή αυτή γίνεται χρησιμοποιώντας το πρωτόκολλο Διασταυρούμενης Επικύρωσης (Cross-Validation). Η βασική ιδέα πίσω απο το Cross-Validation είναι η χρήση ενός μόνο μέρους του συνόλου εκπαίδευσης για την εκπαίδευση του μοντέλου, χρησιμοποιώντας τα σημεία που δεν έλαβαν μέρος στην εκπαίδευση για τον υπολογισμό της απόδοσης του μοντέλου.

Με αυτό τον τρόπο το πρωτόκολλο δίνει μια εικόνα σχετικά με την ικανότητα γενίκευσης του μοντέλου που εξετάζεται. Έτσι, στην προκειμένη περίπτωση (δηλ. στην επιλογή υπερ-παραμέτρων) δίνει μια εικόνα σχετικά με την βέλτιστη επιλογή για την τιμή της κάθε παραμέτρου.
Επιπροσθέτως, χρησιμοποιώντας το πρωτόκολλο Cross-Validation μειώνεται το πρόβλημα του overfitting στο σύνολο εκπαίδευσης.

Οι διάφοροι τρόποι υλοποίησης του πρωτοκόλλου Cross-Validation μπορούν να ταξινομηθούν σε δύο κατηγορίες: α) Εξαντλητικό (Exhaustive) Cross-Validation και β) Μη-Εξαντλητικό (Non-Exhaustive) Cross-Validation. Σύμφωνα με το Εξαντλητικό Cross-Validation εξετάζονται όλοι οι πιθανοί τρόπου διαχωρισμού το αρχικού συνόλου εκπαίδευσης σε ύπο-σύνολα εκπαίδευσης και επικύρωσης.

Στη συγκεκριμένη μελέτη χρησιμοποιείται μη-εξαντλητικό Cross-Validation, υπό την μορφή του
K-Fold Cross-Validation. Στο K-Fold Cross-Validation το αρχικό σύνολο εκπαίδευσης χωρίζεται σε K υποσύνολα ίσου μεγέθους. Στη συνέχεια, απο αυτά τα K υποσύνολα, ένα χρησιμοποιείται ως σύνολο επικύρωσης ενώ τα υπόλοιπα K-1 απαρτίζουν το σύνολο εκπαίδευσης.
Η διαδικασία αυτή επαναλαμβάνεται K φορές, ωστέ τελικά καθένα απο τα K υποσύνολα χρησιμοποιείται μία φορά ως σύνολο επικύρωσης. Τελικώς, τα αποτελέσματα απόδοσης του μοντέλου για καθεμία απο τις K επαναλήψεις συνδυάζονται.

Αυτή η διαδικασία επαναλαμβάνεται για όλους τους πιθανούς συνδυασμούς των υπερ-παραμέτρων, ωστέ τελικά λαμβάνεται ένα σύνολο που περιγράφει την απόδοση του μοντέλου για καθένα απο αυτούς τους συνδυασμούς, σύμφωνα με το Cross-Validation.

6 Αποτελέσματα πειραμάτων SIC

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα των πειραμάτων στα σύνολα δεδομένων SIC, καθώς και σύντομος σχολιασμός επι αυτών. Αναφέρονται επίσης οι διάφορες υπερ-παράμετροι που χρειάστηκε να επιλεχθούν για την επίτευξη της βέλτιστης απόδοσης κάθε τεχνικής. Παρουσιάζονται επίσης οι τιμές που προέκυψαν για τις παραμέτρους αυτές.

6.1 Αποτελέσματα SIC97

Στο σύνολο δεδομένων SIC9 χρησιμοποιήθηκαν οι εξης μέθοδοι:

- Δέντρο Παλινδρόμησης (Regression Tree, R-Tree)
- K – Πλησιέστεροι Γείτονες (K – Nearest Neighbors, KNN)
- Πολυεπίπεδο Perceptron (Multi-Layer Perceptron - MLP)
- Παλινδρόμηση με Διανύσματα Υποστήριξης (Support Vector Regression - SVR)
- Ordinary Kriging (OK)
- Random Forests (RF)
- Random Forests σε Συνδυασμό με Inverse Distance Squared (RFIDS)
- Random Forests σε Συνδυασμό με Ordinary Kriging (RFOK)

Αρχικά παρουσιάζονται οι υπερ-παράμετροι για κάθε μια απο αυτές τις μεθόδους, όπως προέκυψαν μέσω Cross-Validation.

- KNN: Η μοναδική παράμετρος προς προσδιορισμό ήταν ο αριθμός K των γειτόνων που λαμβάνονται υπ' όψην για τον υπολογισμό της τιμής οποιασδήποτε άγνωστης θέσης. Η τιμή αυτή προέκυψε 1 .
- MLP: Οι παράμετροι προς προσδιορισμό ήταν ο αριθμός νευρώνων στο κρύφο επίπεδο και ο αριθμός εποχών εκπαίδευσης. Οι τιμές αυτές προέκυψαν 25 και 900 αντίστοιχα.
- SVR: Οι παράμετροι προς προσδιορισμό ήταν η παράμετρος γ της Gaussian συνάρτησης πυρήνα, η τιμή επιτρεπόμενου σφάλματος ε , και η παράμετρος εμπιστοσύνης στις μετρήσεις C . Οι τιμές αυτές προέκυψαν $3.4896 \cdot 10^{-9}$, 5 και 1000 αντίστοιχα.
- OK: Οι παράμετροι προς προσδιορισμό ήταν τα *nugget, sill* και *range*. Σε αυτή την περίπτωση δεν χρησιμοποιήθηκε Cross-Validation, αλλά υπολογίστηκε το variogram των τιμών εκπαίδευσης, στο οποίο προσδιορίζονται οι τιμές αυτές. Οι τιμές προέκυψαν 0 , $1.5 \cdot 10^4$ και $6 \cdot 10^4$ αντίστοιχα.
- RF: Οι παράμετροι προς προσδιορισμό ήταν ο αριθμός των κατασκευαζόμενων δέντρων και ο αριθμός features που επιλέγονται σε κάθε διαχωρισμό. Οι τιμές αυτές προέκυψαν 3000 και 1 αντίστοιχα.
- RFIDS: Οι παράμετροι του RF είναι ίδιες με παραπάνω, καθώς αντιμετωπίζει το ίδιο πρόβλημα. Το IDS δεν απαιτεί τον προσδιορισμό κάποιας υπερ-παραμέτρου.
- RFOK: Οι παράμετροι του RF είναι ίδιες με παραπάνω, καθώς αντιμετωπίζει το ίδιο πρόβλημα. Για τις παραμέτρους του OK υπολογίστηκε το variogram με βάση τα residuals που προκύπτουν, οπότε οι νέες τιμές *nugget, sill* και *range* προέκυψαν 0 , 450 και $5 \cdot 10^4$ αντίστοιχα.
-

Με βάση τις παραμέτρους αυτές, τα αποτελέσματα για το σύνολο δεδομένων SIC97 είναι τα εξής:

	R-Tree	KNN	MLP	SVR	OK	RF	RFIDS	RFOK
MAE	70.17	46.08	49.91	44.74	43.62	47.65	43.8	43.27
RMSE	93.93	74.61	69.29	66.67	62.82	69.14	66.53	65.89

Όπως φαίνεται, σε σχετικά μικρό σύνολο δεδομένων (υπεμθυμίζεται οτι το σύνολο εκπαίδευσης περιελάμβανε 100 μόνο σημεία, ενώ 367 σημεία είναι προς πρόβλεψη), οι τεχνικές μηχανικής μάθησης δεν υπερτερούν του Ordinary Kriging. Ο συνδυασμός τεχνικών μηχανικής μάθησης με άλλες τεχνικές αδυναμεί επίσης να δώσει αισθητά καλύτερα αποτελέσματα. Συνεπώς σε τέτοιες περιπτώσεις επιβεβαιώνεται η υπόθεση των Gilardi & Bengio πως οι κλασσικές τεχνικές γεωστατιστικής είναι πιο αποτελεσματικές. [4]

6.2 Αποτελέσματα SIC2004 – Natural

Στην περίπτωση του συνόλου δεδομένων SIC2004 χρησιμοποιήθηκαν όλες οι παραπάνω μέθοδοι.

Όπως προηγουμένως, αρχικά παρουσιάζονται οι υπερ-παράμετροι:

- KNN: Η μοναδική παράμετρος προς προσδιορισμό ήταν ο αριθμός K των γειτόνων που λαμβάνονται υπ' όψην για τον υπολογισμό της τιμής οποιασδήποτε άγνωστης θέσης. Η τιμή αυτή προέκυψε 2 .
- MLP: Οι παράμετροι προς προσδιορισμό ήταν ο αριθμός νευρώνων στο κρύφο επίπεδο και ο αριθμός εποχών εκπαίδευσης. Οι τιμές αυτές προέκυψαν 30 και 800 αντίστοιχα.
- SVR: Οι παράμετροι προς προσδιορισμό ήταν η παράμετρος γ της Gaussian συνάρτησης πυρήνα, η τιμή επιτρεπόμενου σφάλματος ε , και η παράμετρος εμπιστοσύνης στις μετρήσεις C . Οι τιμές αυτές προέκυψαν $3.5 \cdot 10^{-9}$, 0.5 και 10 αντίστοιχα.
- OK: Οι παράμετροι προς προσδιορισμό ήταν τα *nugget, sill* και *range*. Όπως και πριν, οι τιμές προέκυψαν 100 , 550 και $3.5 \cdot 10^{5}$ αντίστοιχα.
- RF: Οι παράμετροι προς προσδιορισμό ήταν ο αριθμός των κατασκευαζόμενων δέντρων και ο αριθμός features που επιλέγονται σε κάθε διαχωρισμό. Οι τιμές αυτές προέκυψαν 75 και 11 αντίστοιχα.
- RFIDS: Οι παράμετροι του RF είναι ίδιες με παραπάνω, καθώς αντιμετωπίζει το ίδιο πρόβλημα. Το IDS δεν απαιτεί τον προσδιορισμό κάποιας υπερ-παραμέτρου.
- RFOK: Οι παράμετροι του RF είναι ίδιες με παραπάνω, καθώς αντιμετωπίζει το ίδιο πρόβλημα. Για τις παραμέτρους του OK οι νέες τιμές *nugget, sill* και *range* προέκυψαν 3 , 60 και $2 \cdot 10^{4}$ αντίστοιχα.

Με βάση τις παραμέτρους αυτές, τα αποτελέσματα για το σύνολο δεδομένων SIC2004 - Natural είναι τα εξής:

	R-Tree	***KNN***	***MLP***	***SVR***	***OK***	***RF***	***RFIDS***	***RFOK***
MAE	8.81	9.39	8.97	8.88	7.81	7.69	7.61	7.49
RMSE	11.96	12.72	11.77	11.6	10.3	10.3	10.11	10.05

Σε αυτή την περίπτωση, ενώ οι κλασικοί αλγόριθμοι μηχανικής μάθησης πάλι δεν έχουν καλύτερα αποτελέσματα απο την μέθοδο Ordinary Kriging, φαίνεται πως η μέθοδος μάθησης συνόλου Random Forests εμφανίζει οριακά καλύτερη επίδοση. Η διαφορά αυτή γίνεται πιο αισθητή στην περίπτωση του συνδυασμού μεθόδων.

6.3 Αποτελέσματα SIC2004 – Joker

Όπως προηγουμένως, αρχικά παρουσιάζονται οι υπερ-παράμετροι:

- KNN: Η τιμή K προέκυψε 3 .
- MLP: Οι τιμές νευρώνων και εποχών προέκυψαν 30 και 900 αντίστοιχα.
- SVR: Οι τιμές γ , ε και C προέκυψαν $6.173 \cdot 10^{-9}$, 0.1 και 1000 αντίστοιχα.
- OK: Οι παράμετροι *nugget, sill* και *range* προέκυψαν 0 , 100000 και 10^5 αντίστοιχα.
- RF: Οι παράμετροι προς προσδιορισμό ήταν ο αριθμός των κατασκευαζόμενων δέντρων και ο αριθμός features που επιλέγονται σε κάθε διαχωρισμό. Οι τιμές αυτές προέκυψαν 7000 και 1 αντίστοιχα.
- RFIDS: Οι παράμετροι του RF είναι ίδιες με παραπάνω, καθώς αντιμετωπίζει το ίδιο πρόβλημα. Το IDS δεν απαιτεί τον προσδιορισμό κάποιας υπερ-παραμέτρου.
- RFOK: Οι παράμετροι του RF είναι ίδιες με παραπάνω, καθώς αντιμετωπίζει το ίδιο πρόβλημα. Για τις παραμέτρους του OK οι νέες τιμές *nugget, sill* και *range* προέκυψαν 7 , 250 και $1.2 \cdot 10^5$ αντίστοιχα.

Με βάση τις παραμέτρους αυτές, τα αποτελέσματα για το σύνολο δεδομένων SIC2004 - Joker είναι τα εξής:

	R-Tree	*KNN*	*MLP*	*SVR*	*OK*	*RF*	*RFIDS*	*RFOK*
MAE	13.49	21.48	28.28	23.33	24.74	15.94	16.28	16.71
RMSE	44.29	116.79	114.97	90.66	94.18	62.61	62.62	60.38

Τα αποτελέσματα αυτού του πειράματος παρουσιάζουν ιδιαίτερο ενδιαφέρον. Υπενθυμίζεται οτι η βασική ιδέα πίσω απο την δημιουργία του συνόλου δεδομένων Joker ήταν η διερεύνηση της ικανότητας εντοπισμού ενός σημείου ραδιενεργού ατυχήματος, του λεγόμενου “hot-spot”. Στο συγκεκριμένο σημείο εμφανίζονται ιδιαίτερα ακραίες τιμές, επομένως η μετρική *RMSE* καθορίζεται κυρίως απο την ακρίβεια κάθε μοντέλου στο να εντοπίσει και να υπολογίσει σωστά τα σημεία που βρίσκονται κοντά στο “hot-spot”.

Έτσι, το μοντέλο SVR έχει καλύτερα αποτελέσματα απο το Ordinary Kriging σε αυτή την περίπτωση, γεγονός που προέρχεται απο το γεγονός οτι εντοπίζει το “hot-spot”, κάτι που οι άλλες κλασικές μέθοδοι μηχανικής μάθησης δεν φαίνεται να έχουν καταφέρει.

Ενδιαφέρον είναι επίσης το γεγονός πως η τεχνική Random Forests επιτυγχάνει ιδιαίτερα καλύτερη απόδοση, τόσο ως προς το *MAE* όσο και ως προς το *RMSE*. Αυτή η βελτίωση μπορεί να αποδοθεί στην φύση των αλγορίθμων μάθησης συνόλου: Μέσω του Bagging και του Feature Bagging το μοντέλο έχει τη δυνατότητα να εστιάσει σε υπο-περιπτώσεις του προβλήματος, συνεπώς καταφέρνει να προσεγγίσει το “hot-spot” χωρίς να θυσιάζει την ακρίβεια στα υπόλοιπα, πιο ομάλα σημεία.

Αυτή η λογική μπορεί να εξηγήσει και την συγκριτική χειροτέρευση του *MAE* στην περίπτωση του RFIDS: Δεδομένου οτι το IDS δεν χρησιμοποιεί κάποιο ευφυή τρόπο επιλογής των γειτόνων, είναι πιθανό επιλεγμένοι γείτονες που ανήκουν στο “hot-spot” να παρασύρουν την πρόβλεψη ενός μη-ακραίου σημείου, αυξάνοντας έτσι το *MAE.*

Ενδιαφέρον παρουσιάζει και η περίπτωση του RFOK: φαίνεται πως δεδομένου οτι το OK έχει την τάση “ομαλοποίησης” και άρα οχι τόσο καλής συμπεριφοράς σε απότομες μεταβολές, η προσπάθεια καλύτερης προσέγγισης του “hot-spot” (μειωμένο *RMSE*), οδηγεί στην εμφάνιση μεγαλύτερων σφαλμάτων στις ομαλές περιοχές (αυξημένο *MAE*), σε σχέση

με την εφαρμογή απλών RF.

Εντυπωσιακά είναι επίσης τα αποτελέσματα του R–Tree στο συγκεκριμένο σύνολο δεδομένων. Όπως φαίνεται, σε αντίθεση με τις προηγούμενες περιπτώσεις οπού το R–Tree είχε αισθητά χειρότερη απόδοση απο τις άλλες μεθόδους, σε αυτή την περίπτωση έχει τα καλύτερα αποτελέσματα. Φαίνεται λοιπόν πως η ομαλοποίηση που προκύπτει μέσω των γεωστατιστικών μεθόδων IDS και OK οδηγεί σε μεγάλυ απώλεια ακρίβειας για το συγκεκριμένο, ακραίο σύνολο δεδομένων, όπως αναφέρθηκε και παραπάνω.

Επιλέον, το R–Tree έχει καλύτερη απόδοση και απο την εφαρμογή RF, γεγονός που οδηγεί στο συμπέρασμα πως στη συγκεκριμένη περίπτωση ο ταξινομητής που κατασκευάζεται (δηλαδή το δέντρο απόφασης) είναι ιδιαίτερα πιο ισχύρος από τους πολλούς, ασθενείς ταξινομητές που δημιουργούνται με βάση το RF, σε σημείο οπού ο συνδυασμός των προβλέψεων αυτών να μην οδηγεί σε καλύτερη ακρίβεια.

7 Περαιτέρω Βελτιστοποίηση απο την πλευρά της Γεωστατιστικής

Σε αυτό το κεφάλαιο γίνεται μια προσπάθεια βελτίωσης της απόδοσης της τεχνικής RFOK, η οποία κρίθηκε ως η πιο αποτελεσματική με βάση τα προηγούμενα πειράματα. Προς αυτή την κατεύθυνση, εξετάζονται τεχνικές που χρησιμοποιούνται στην πράξη στον τομέα της γεωστατιστικής για την αντιμετώπιση του προβλήματος της χωρικής παρεμβολής. Παρουσιάζεται επίσης και το πρωτότυπο σύνολο δεδομένων, το οποίο σχηματίστηκε με δεδομένα απο το επιστημονικό πρόγραμμα earthscope και χρησιμοποιήθηκε για τον έλεγχο των παραπάνω βελτιώσεων.

7.1 Προσομοίωση υπο Συνθήκη

Ανάλυση των συντελεστών συσχετισμού μεταξύ των αρχικών εισόδων και των Residuals που προκύπτουν απο τη συνολική μέθοδο RFOK έδειξε ικανοποιητικό βαθμό συσχετισμού μεταξύ των δύο, για όλα τα σύνολα δεδομένων για τα οποία υπολογίστηκε. Έτσι, μπορούμε να υποθέσουμε οτι υπάρχει πληροφορία την οποία το μοντέλο δεν έχει εκμεταλλευτεί και άρα περιθώριο βελτίωσης.

Συγκεκριμένα, στις πρακτικές εφαρμογές γεωστατιστικής, το βήμα που προστίθεται στην θέση της εφαρμογής του OK, ονομάζεται Προσομοίωση υπο Συνθήκη (Conditional Simulation) και προσπαθεί να αντιμετωπίσει το πρόβλημα της ομαλότητας που εκ φύσεως προκύπτει μέσω του OK (όπως παρατηρήθηκε και στα πρώτα πειράματα).

Με την χρήση conditional simulation παράγονται πολλές πιθανές υλοποιήσεις (realizations) του υπό μελέτη χώρου, οι οποίες είναι πιστές στα δεδομένα εισόδου, αλλά έχουν τιμές που διαφέρουν στα ενδιάμεσα διαστήματα. Κάθε μια απο αυτές τις υλοποιήσεις θεωρείται το ίδιο πιθανή με τις υπόλοιπες, και οι τιμές αυτών συνδυάζονται για την παραγωγή του τελικού αποτελέσματος. Έτσι, με τον υπολογισμό ενός μεγάλου αριθμού υλοποιήσεων γίνεται πιο πιθανός και αποτελεσματικός ο εντοπισμός πιθανών ακραίων τιμών. [24]

7.2 Εξερευνητική Ανάλυση Δεδομένων

Όπως αναφέρθηκε και σε προηγούμενο κεφάλαιο όλες οι μέθοδοι χωρικής παρεμβολής επηρεάζονται άμεσα απο το σύνολο δεδομένων εκπαίδευσης, άλλες περισσότερο και άλλες λιγότερο. Επομένως, ένας άλλος τομέας μελέτης για το ενδεχόμενο βελτίωσης της επίδοσης της χωρικής παρεμβολής είναι το σύνολο δεδομένων εκπαίδευσης.

Πράγματι, στις πρακτικές εφαρμογές γεωστατιστικής το πρώτο στάδιο είναι η μελέτη των δεδομένων, η οποία αποκαλείται Εξερευνητική Ανάλυση Δεδομένων (Exploratory Data Analysis, EDA). Η ανάλυση αυτή περιλαμβάνει τον έλεγχο της εγκυρότητας τον δεδομένων καθώς και την αφαίρεση δεδομένων που θεωρούνται οτι έχουν μετρηθέι εσφαλμένα (outliers). Επιπλέον, σε αυτό το στάδιο μπορεί να συμπεριληφθεί και η μετατροπή των δεδομένων σε μορφές πιο βολικές για τις μεθόδους που θα εφαρμοστούν (π.χ. κανονικοποίηση των δεδομένων). Η μελέτη αυτή γενικώς είναι μη-τετριμμένη και μπορεί να καταλαμβάνει μέχρι και το 75% του συνολικού χρόνου μιας γεωστατιστικής μελέτης χωρικής παρεμβολής. [24]

7.3 Ζητήματα υλοποίησης: Προσομοίωση υπο Συνθήκη

Υπάρχουν διάφορες τεχνικές conditional simulation οι οποίες μπορούν να εφαρμοστούν, ωστόσο η πιο δημοφιλής, η οποία εφαρμοστηκε και στη συγκεκριμένη μελέτη, ονομάζεται Sequential Gaussian Simulation (SGSIM).

Η βασική ιδέα του SGSIM είναι ο "αυξητικός" υπολογισμός των τιμών σε κάθε σημείο του χώρου, με χρήση κάποιας απο τις μεθόδους Kriging ως πυρήνα των υπολογισμών και τη διαδοχική εισαγωγή των ήδη υπολογισμένων σημείων στο σύνολο γνωστών σημείων που χρησιμοποιείται για τον υπολογισμό.
Έτσι:

1. Αρχικά ο συνολικός χώρος μελέτης οργανώνεται ως πλέγμα (grid), οπού κάθε υπο-χώρος που ορίζεται με βάση το πλέγμα μπορεί να περιλαμβάνει ένα ή περισσότερα σημεία του πραγματικού χώρου υπο μελέτη.
2. Με τυχαίο τρόπο, επιλέγεται μια "διαδρομή" υπολογισμών σύμφωνα με την οποία θα υπολογιστούν διαδοχικά οι διάφοροι υπο-χώροι του πλέγματος.
3. Ακολουθώντας την σειρά που ορίζει η τυχαία διαδρομή υπολογισμών, υπολογίζονται διαδοχικά οι υπο-χώροι που ορίζει το πλέγμα, θεωρώντας τους ήδη υπολογισμένους χώρους ως γνωστά σημεία κατά τον υπολογισμό της τεχνικής Kriging που χρησιμοποιείται.
4. Για τον επιθυμητό αριθμό υλοποιήσεων του conditional simulation, υπολογίζεται κάθε φορά νεά διαδρομή υπολογισμών, και επαναλαμβάνεται ο συνολικός υπολογισμός όπως παραπάνω.

Η χρήση τυχαίας διαδρομής υπολογισμών εξασφαλίζει πως οι υλοποιήσεις θα εμφανίζουν διαφορές μεταξύ τους, οπότε προκύπτει το σύνολο ισοπίθανων υλοποιήσεων που τελικά χρησιμοποιείται για την βελτίωση των αποτελεσμάτων.
Στην πράξη, υπολογίζεται ένας μεγάλος αριθμός υλοποιήσεων. Ένας τυπικός αριθμός υλοποιήσεων είναι 1000. [24] Ο όγκος υπολογισμών που συνεπάγεται αυτός ο αριθμός είναι ιδιαίτερα μεγάλος, συνεπώς στην συγκεκριμένη εφαρμογή υπολογίστηκε ένας περιορισμένος αριθμός υλοποιήσεων, γεγονός που περιορίζει την βελτίωση της απόδοσης απο αυτή την τεχνική.

7.4 Ζητήματα υλοποίησης: Εξερευνητική Ανάλυση Δεδομένων

Για την καλύτερη παρουσίαση της Εξερευνητικής Ανάλυσης Δεδομένων, κρίνεται σκόπιμο να παρουσιαστούν πρώτα κάποιες πληροφορίες για την φύση του πρωτότυπου συνόλου δεδομένων που χρησιμοποιήθηκε.

Το πρωτότυπο σύνολο δεδομένων κατασκευάστηκε χρησιμοποιώντας δεδομένα απο το επιστημονικό πρόγραμμα earthscope, αντικείμενο του οποίου είναι η γεωγραφική μορφή και εξέλιξη της Βορειοαμερικάνικης ηπείρου.

Πιο λεπτομερώς, μέσω του επιστημονικού προγράμματος earthscope διατίθεται ένα ιδιαίτερα μεγάλο σύνολο μετρήσεων, εκτεταμένο τόσο χωρικά όσο και χρονικά. Ωστόσο για τις ανάγκες της μελέτης ενδιαφέρει μόνο η χωρική έκταση των δεδομένων. Έτσι, για την δημιουργία του “βέλτιστου” συνόλου δεδομένων απο απόψη χωρικής έκτασης εφαρμόστηκε μια αναζήτηση στο συνολικό πλήθος των σταθμών προκειμένου να βρεθεί η ημέρα την οποία λειτουργεί ταυτόχρονα ο μεγαλύτερος αριθμός σταθμών. Η μέρα αυτή ευρέθηκε να είναι η 27-02-2015 για την οποία υπάρχουν 1879 μετρήσεις.

Κάθε μια απο αυτές τις μετρήσεις αφορά την γεωγραφική θέση ενός σταθμού τη συγκεκριμένη μέρα. Αυτή η τιμή μπορεί να συγκριθεί με την θέση αναφοράς (reference position) του ίδιου σταθμού για να υπολογιστεί η μετατόπισή του. Οι θέσεις αυτές εκφράζονται σε δύο διαφορετικά συστήματα συντεταγμένων: XYZ, ή NEU. Στους υπολογισμούς τελικά χρησιμοποιείται το δεύτερο. Έτσι, το σύνολο δεδομένων περιλαμβάνει τρεις μεταβλητές εισόδου (συντεταγμένες αναφοράς του σταθμού στο σύστημα NEU) και ζητούμενο είναι ο προσδιορισμός τριών μεταβλητών εξόδου, η οποίες εκφράζουν τις συντεταγμένες του σταθμού τη συγκεκριμένη χρονική στιγμή (επίσης εκφρασμένες στο σύστημα NEU).

Στη διαδικασία του EDA επίσης σημαντικός είναι ο εντοπισμός outliers, δηλαδή τιμών που απέχουν ιδιαίτερα απο το μέσο όρο των μετρήσεων για δεδομένο σταθμό. Αυτές οι τιμές δεν συμπεριλαμβάνονται στο τελικό σύνολο δεδομένων καθώς είναι ιδιαίτερα πιθανό να αποτελούν εσφαλμένες μετρήσεις, ενώ η παρουσία τους μπορεί να προκαλέσει σημαντικές αλλαγές στα μοντέλα που δημιουργούνται, όπως φάνηκε και στα πειράματα που αφορούσαν το Joker σύνολο δεδομένων.

Ο εντοπισμός των outliers αποτελεί ένα πρόβλημα το οποίο δεν έχει μοναδική λύση. Υπάρχουν διάφορες προσεγγίσεις, όπως η χρήση του απλού μέσου όρου ή της τιμής mean absolute deviaton που εκφράζει την μέση απόκλιση όλων των τιμών από το μέσο όρο.

Στη συγκεκριμένη περίπτωση, χρησιμοποιήθηκε η τεχνική του Γενικευμένου ESD (extreme studentized deviate) Test.

Η τεχνική αυτή απαιτεί τον ορισμό της παραμέτρου r , το οποίο αποτελεί ένα άνω φράγμα στον αριθμό αναμενόμενων outliers. Με βάση αυτή την τιμή, το Γενικευμένο ESD Test πραγματοποιεί r διαφορετικούς ελέγχους: έναν έλεγχο για την περίπτωση ενός outlier, έναν έλεγχο για την περίπτωση δύο outliers, μέχρι και την περίπτωση r outliers.

Για κάθε μία απο αυτές τις περιπτώσεις υπολογίζονται δύο τιμές:

- Την τιμή R_i που εκφράζει την μέγιστη απόκλιση από τον μέσο όρο, έχοντας πρώτα αφαιρέσει τις i τιμές μέγιστη απόκλιση απο τον μέσο όρο.
- Την τιμή λ_i που εκφράζει την κρίσιμη τιμή που θα χρησιμοποιηθεί για τον έλεγχο, στην περίπτωση i outliers.

Έχοντας υπολογίσει τις δύο παραπάνω τιμές, ο έλεγχος εντοπίζει την μέγιστη τιμή i για την οποία προκύπτει $R_i > \lambda_i$. [25]

Ο παραπάνω έλεγχος εφαρμόστηκε σε όλες τις μετρήσεις όλων των σταθμών, και με βάση τα αποτελέσματα αυτού εξετάστηκε αν εμφανίζεται outlier την ημερομηνία 27-02-2015, απο την οποία προκύπτουν τα δεδομένα που τελικά χρησιμοποιήθηκαν πειραματικά, όπως αναφέρθηκε προηγουμένως. Με αυτό τον παραπάνω έλεγχο ο τελικός αριθμός δεδομένων είναι 1876 μετρήσεις.

Απο αυτές, οι 1300 χρησιμοποιήθηκαν ως σύνολο εκπαίδευσης ενώ οι υπόλοιπες 576 χρησιμοποιήθηκαν ως σύνολο ελέγχου της εγκυρότητας των διαφόρων μοντέλων.

8 Αποτελέσματα Earthscope

Σε αυτό το κεφάλαιο παρουσιάζονται τα αποτελέσματα των πειραμάτων στο πρωτότυπο σύνολο δεδομένων earthscope, καθώς και σύντομος σχολιασμός επι αυτών. Αναφέρονται επίσης οι διάφορες υπερ-παράμετροι που χρειάστηκε να επιλεχθούν για την επίτευξη της βέλτιστης απόδοσης κάθε τεχνικής. Τέλος, παρουσιάζονται οι τιμές που προέκυψαν για τις παραμέτρους αυτές.

8.1 Παρουσίαση Παραμέτρων

Για το σύνολο δεδομένων earthscope χρησιμοποιήθηκαν οι εξης μέθοδοι, με χρήση διαφορετικών υπερ-παραμέτρων για την βέλτιστη επίδοση ανάλογα με την κάθε μεταβλητή εξόδου N, E, ή U:

- Ordinary Kriging (OK)
- Random Forests (RF)
- Random Forests σε Συνδυασμό με Ordinary Kriging (RFOK)

Επιπλέον χρησιμοποιήθηκε η τεχνική RF σε συνδυασμό με Sequential Gaussian Simulation (SGSIM), όπως εξετάστηκε προηγουμένως, μόνο για την περίπωση της N μεταβλητής εξόδου. Η μελέτη μίας μόνο μεταβλητής προέκυψε λόγω του υπολογιστικού κόστους που προκύπτει μέσω του SGSIM, ενώ ο αριθμός υλοποιήσεων που πραγματοποιήθηκαν μέσω του SGSIM θεωρείται περιορισμένος (100 υλοποιήσεις έναντι του τυπικού 1000).

Στη συνέχεια παρουσιάζονται οι υπερ-παράμετροι, όπως προέκυψαν απο την χρήση Cross-Validation για την περίπτωση των Random Forests και την μελέτη του variogram για την περίπτωση του Ordinary Kriging:

Ordinary Kriging Parameters:

	N	***E***	***U***
Nugget	50	100	0
Sill	170	800	$1.8\ 10^6$
Range	100	100	2000

Random Forest Parameters:

	N	***E***	***U***
Number of Trees	3000	7000	5000
Number of Features	1	1	2

Ordinary Kriging on Random Forest Residuals Parameters:

	N	***E***	***U***
Nugget	0.001	0.003	0
Sill	0.030	0.275	27
Range	180	200	1900

8.2 Παρουσίαση αποτελεσμάτων

Με αυτές τις παραμέτρους, τα αποτελέσματα για τις διάφορες μεθόδους ήταν τα εξής:

North Latitude Results:

N	*OK*	*RF*	*RFOK*	*RF-SGSIM*
MAE	2.0423	0.0531	0.0523	0.0496
RMSE	3.7639	0.2634	0.2548	0.2487

East Longtitude Results:

E	*OK*	*RF*	*RFOK*	*RF-SGSIM*
MAE	1.8797	0.1189	0.1180	0.1139
RMSE	4.9278	1.0287	0.9748	0.9627

Height Results:

U	*OK*	*RF*	*RFOK*	*RF-SGSIM*
MAE	1.1564	4.2262	4.0316	4.0121
RMSE	18.4219	24.3498	18.1798	18.0527

Στις δύο πρώτες περιπτώσεις, η εφαρμογή των Random Forests δίνει ιδιαίτερα καλύτερα αποτελέσματα απ' ο,τι η εφαρμογή του Ordinary Kriging, ενω ο συνδυασμός Random Forests – Ordinary Kriging βελτιώνει την απόδοση ελάχιστα. Φυσικά, στην περίπτωση της συντεταγμένης *N* τα αποτελέσματα βρίσκονται ήδη σε εντυπωσιακά καλές τιμές: ο μέσος όρος των δεδομένων είναι *39.3091,* επομένως τόσο το *MAE* οσο και το *RMSE* εκφράζουν σφάλματα μικρότερα του 1% του μέσου όρου των εξεταζόμενων τιμών. Το ίδιο φαίνεται να ισχύει και για την μεταβλητή *E,* οπού ο μέσος όρος των δεδομένων είναι *247.9005,* δικαιολογώντας τις αυξημένες τιμές σφαλμάτων σε σχέση με την συντεταγμένη *N.*

Όσον αφορά την τεχνική RF-SGSIM, παρατηρείται μια μικρή βελτίωση της επίδοσης. Αυτή η ιδιαίτερα μικρή βελτίωση μπορεί να δικαιολογηθεί απο το γεγονός οτι οι υλοποιήσεις που δημιουργήθηκαν χρησιμοποιώντας τον αλγόριθμο ήταν λίγες σε σχέση με τον τυπικό αριθμό που χρησιμοποιείται (100 έναντι 1000 υλοποιήσεις), ενώ δεν εφαρμόζεται κάποιας μορφής ανάλυση αβεβαιότητας (uncertainty analysis) ώστε να αγνοηθεί η περιοριστεί η επιρροή λιγότερο πιθανών υλοποιήσεων στο τελικό αποτέλεσμα.

9 Συμπεράσματα – Μελλοντικές Επεκτάσεις

Σε αυτό το κεφάλαιο παρουσιάζονται ορισμένα γενικά συμπεράσματα που προκύπτουν απο τα πειράματα που πραγματοποιήθηκαν. Επιπλέον, αναφέρονται ορισμένες πιθανές επεκτάσεις και βελτιστοποιήσεις της μεθοδολογίας που τελικά έδωσε τα καλύτερα αποτελέσματα στην αντιμετώπιση του προβλήματος της χωρικής παρεμβολής.

9.1 Συμπεράσματα με Βάση τα Αποτελέσματα των Πειραμάτων

Όπως σχολιάστηκε και στα αντίστοιχα κεφάλαια αποτελεσμάτων, το γενικό συμπέρασμα που προκύπτει απο τα αποτελέσματα των πειραμάτων είναι πως δεν υπάρχει ξεκάθαρη υπεροχή κάποιας απο της τεχνικές για όλες τις μορφές με τις οποίες μπορεί να εμφανιστεί το πρόβλημα της χωρικής παρεμβολής.

Έτσι, για τα σχετικά μικρά σύνολα δεδομένων SIC η σύγκριση μεταξύ τεχνικών μηχανικής μάθησης και τεχνικών γεωστατιστικής (οι οποίες αντιπροσωπεύονται κυριώς απο το Ordinary Kriging στη συγκεκριμένη μελέτη) περιορίζεται σε σχετικά μικρές διαφορές. Το OK φαίνεται να υπερτερεί των κλασικών τεχνικών μηχανικής μάθησης, ενώ εμφανίζει επίδοση αρκετά παραπλήσια με τα Random Forests.

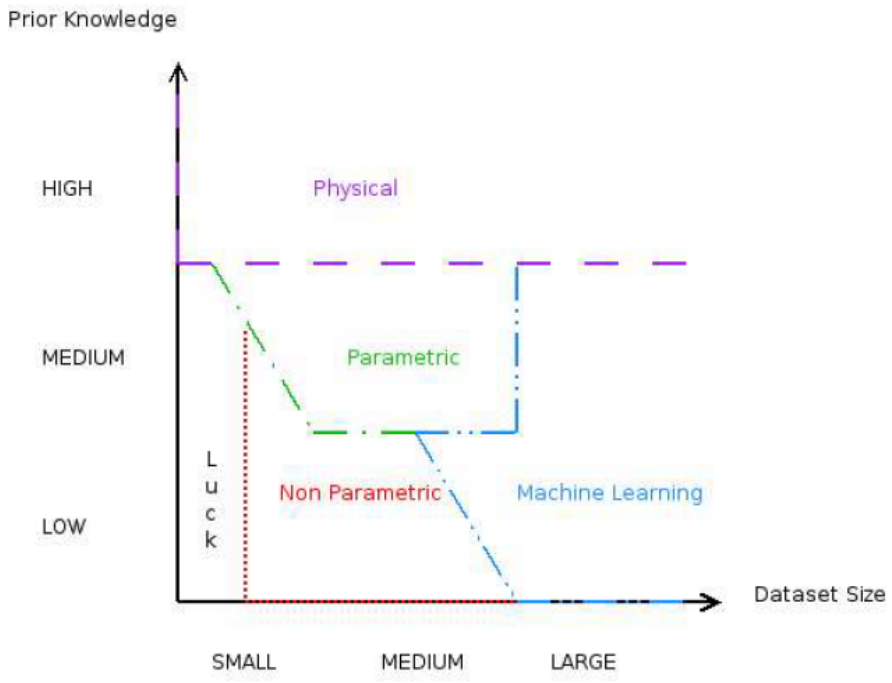
Σημαντικό είναι επίσης να σημειωθεί πως για σύνολα δεδομένα με ιδιαίτερα απότομες κορυφές στις τιμές των δεδομένων (όπως το SIC2004 Joker) τα Random Forests καταφέρνουν να ξεχωρίσουν απο το OK, και να προσεγγίσουν τις απότομες αυτές αλλαγές καλύτερα. Αντίθετα, για σχετικά ομαλά σύνολα δεδομένων τα RF εμφανίσαν σχεδόν αμελητέα καλύτερα επίδοση από το OK.

Για μεγαλύτερα σύνολα δεδομένων, που στην συγκεκριμένη μελέτη αντιπροσωπεύονται απο το πρωτότυπο σύνολο δεδομένων earthscope, η διαφορά μεταξύ OK και RF γίνεται πιο αισθητή. Τα RF υπερτερούν σε 2 απο τις 3 περιπτώσεις που μελετήθηκαν, παρουσιάζοντας καλύτερα αποτελέσματα ως προς τις δύο μετρικές επίδοσης που λήφθηκαν υπ' όψην.

Ο συνδυασμός RF με OK φαίνεται οτι, ενώ εν γένει μπορεί να θεωρηθεί βελτιστοποίηση της απλής εφαρμογής RF, εντούτοις ακολουθεί την τάση που εξηγήθηκε παραπάνω. Έτσι, έχοντας ως παράδειγμα την U – συντεταγμένη του συνόλου δεδομένων earthscope, απο τη στιγμή που η εφαρμογή RF έχει αποτελέσματα αισθητά χειρότερα από το απλό OK, ο συνδυασμός αυτών οριακά καταφέρνει να επιτύχει αποτελέσματα αντίστοιχα με την απλή εφαρμογή OK. Αντίθετα, στις άλλες δύο συντεταγμένες του ίδιου συνόλου δεδομένων ο συνδυασμός μεθόδων καταφέρνει να βελτιώσει τα αποτελέσματα σε σχέση με την ήδη καλή επίδοση του RF.

Στην παραπάνω σύγκριση δεν λαμβάνεται υπ' όψην το ενδεχόμενο ύπαρξης εκ των προτέρων γνώσης σχετικά με το πρόβλημα που αντιμετωπίζεται. Πιο συγκεκριμένα, τα RF αποτελούν έναν αλγόριθμο ο οποίος εκ φύσεως δεν απαιτεί – εκμεταλλέυεται την πιθανή ύπαρξη εκ των προτέρων γνώσης. Αντίθετα, το OK παρουσιάζει τόσο την φαινομενική απαίτηση αλλά ταυτόχρονα και την δυνατότητα χρήσης πληροφορίας αυτής της μορφής μέσω του μοντέλου variogram. Αυτό αποτελεί ένα χαρακτηριστικό το οποίο πρέπει να λαμβάνεται υπ' όψην, καθώς δεν αποτελεί ξεκάθαρο πλεονέκτημα ή μειονέκτημα κάποιας απο αυτές τις μεθόδους.

Τέλος, παρουσιάζεται μια ενδιαφέρουσα γραφική παράσταση προερχόμενη από τους N. Gilardi και S. Bengio [3] . Η οποία συνοψίζει ποιοτικά την σχέση μεταξύ μεγέθους συνόλου δεδομένων, διαθέσιμης εκ των προτέρων γνώσης και μεθόδου χωρικής παρεμβολής που ενδείκνυται να εφαρμοστεί:



Η σχέση μεταξύ μεγέθους συνόλου δεδομένων, διαθέσιμης εκ των προτέρων γνώσης και ενδεικνύμενης προσέγγισης χωρικής παρεμβολής

9.2 Μελλοντικές Επεκτάσεις

Από την πλευρά της γεωστατιστικής, σημειώνεται η εξής επέκταση της μεθόδολογίας που τελικά οδήγησε στα καλύτερα αποτελέσματα:

Uncertainty Analysis

Όπως αναφέρθηκε και προηγουμένως, η τεχνική Conditional Simulation εφαρμόστικε ενδεικτικά και επομένως δεν φαίνεται να βελτίωσε την απόδοση στο μέγιστο δυνατό. Για να επιτευχθεί αυτό, είναι απαραίτητη η δημιουργία ενός αρκετά μεγαλύτερου αριθμού υλοποιήσεων (ενδεικτικό νούμερο είναι 1000 υλοποιήσεις [24]), ενώ δεν έχει εφαρμοστεί το βήμα της ανάλυσης αβεβαιότητας που συνοδεύει αυτή τη μέθοδο. Το βήμα αυτό μπορεί να οδηγήσει σε βελτίωση της απόδοσης καθώς στοχεύει στην μείωση της επίδρασης που έχουν στο τελικό αποτέλεσμα υλοποιήσεις που τελικά κρίνονται λιγότερο πιθανές.

Από την πλευρά της μηχανικής μάθησης και δεδομένου οτι ο αλγόριθμος μηχανικής μάθησης που τελικά έδωσε τα καλύτερα αποτελέσματα ήταν τα RF, η επέκταση της μεθοδολογίας που μελετήθηκε αφορά πιθανές επεκτάσεις και βελτιστοποιήσεις του αλγορίθμου RF. Συγκεκριμένα:

Extremely Randomized Trees (ExtraTrees)

Τα ExtraTrees αποτελούν έναν αλγόριθμο μάθησης συνόλου ο οποίος έχει κοινά σημεία με τα RF όπως παρουσιάστηκαν, ωστόσο έχει ως βασικό στόχο την εισαγωγή ενός μεγαλύτερου βαθμού τυχαιότητας. Αυτή η προσέγγιση ακολουθείται καθώς η εισαγωγή τυχαιότητας φαίνεται να οδηγεί σε καλύτερη απόδοση, όπως φάνηκε για παράδειγμα από την χρήση RF στα προηγούμενα πειράματα. Τα αποτελέσματα εφαρμογής του αλγορίθμου αυτού φαίνονται να είναι καλύτερα από τα RF. [26]

Οι βασικές διαφορές των ExtraTrees σε σχέση με τα RF είναι *α)* η χρήση όλων των στοιχείων του συνόλου εκπαίδευσης για την δημιουργία κάθε δέντρου (σε αντίθεση με την χρήση υποσυνόλων – bagging) και *β)* η επιλογή διαχωρισμού με τρόπο τελείως τυχαίο σε κάθε κόμβο του δέντρου που δημιουργείται (σε αντιθέση με την προσπάθεια ελαχιστοποίησης κάποιας μετρικής απόδοσης)

Βελτιώσεις στον αλγόριθμο των Random Forests

Οι βελτιώσεις στον αλγόριθμο RF που παρουσιάζονται σε αυτό το σημείο στοχέυουν κυρίως στην προσπάθεια βελτίωσης της ισχύος πρόβλεψης καθενός απο τα δέντρα απόφασης που αποτελούν το RF, χωρίς ωστόσο να μειωθεί η ποικιλία μεταξύ αυτών. Σύμφωνα με τον M. Robnik [27] ένας τρόπος να επιτευχθεί αυτό είναι η χρήση πολλών τεχνικών αξιολόγησης διαχωρισμών, σε αντίθεση με την χρήση μιας μόνο, όπως συμβάινει στον βασικό αλγόριθμο RF. Η λογική πίσω απο αυτή την προσέγγιση είναι η προσπάθεια απόκτησης μιας γενικότερης εικόνας κατά την εφαρμογή διαχωρισμών, ωστόσο η προσέγγιση δεν φαίνεται να βελτιώνει τα αποτελέσματα [27].

Η δεύτερη προσέγγιση του Robnik είναι η χρήση σταθμισμένου μέσου όρου κατά τον υπολογισμό της τελικής πρόβλεψης του συνόλου για δεδομένη είσοδο. Το επιθυμητό είναι μεμονομένοι ταξινομητές (δηλ. δέντρα απόφασης) που παρουσιάζουν καλύτερη απόδοση σε πρότυπα εισόδου παρόμοια με το τρέχων πρότυπο εισόδου να εμφανίζουν μεγαλύτερη ισχύ κατά την πρόβλεψη της τελικής εισόδου. Η τεχνική αυτή φαίνεται να οδηγεί σε ικανοποιητική βελτίωση της απόδοσης. [27]

Χρησιμοποιώντας Random Forests ως Συνάρτηση Πυρήνα (Kernel Function)

Οι Davies και Ghahramani παρουσιάζουν μια προσέγγιση σύμφωνα με την οποία είναι δυνατή η χρήση τυχαίων διαμερίσεων ως συναρτήσεις πυρήνα. [28] Πιο συγκεκριμένα, εξετάζεται η χρήση των Random Forests ως συνάρτηση πυρήνα και τα αποτελέσματα που προκύπτουν εμφανίζουν αισθητά καλύτερη απόδοση σε σχέση με τις τυπικές συναρτήσεις πυρήνα όπως η γραμμική η ή Radial Basis Function. [28]

Στη συγκεκριμένη μελέτη, η μέθοδος που χρησιμοποιεί Kernel Function έιναι το SVR. Δεδομένου λοιπόν οτι η χρήση SVR με RBF χρειάστηκε βελτιστοποίηση 3 παραμέτρων, και δεδομένων των αποτελεσμάτων του [28] , η χρήση Random Forest Kernel αναμένεται να οδηγήσει σε βελτίωση της απόδοσης του SVR, ενώ αποφεύγεται η ανάγκη βελτιστοποίησης υπερ-παραμέτρων καθώς το Random Forest Kernel δεν απαιτεί προσδιορισμό αυτών.

Βιβλιογραφία

[1] : C. Deutsch, A. Journel. GSLIB: Geostatistical Software Library and User's Guide,
[2] : G. Bohling. Kriging, 2005
[3] : N. Gilardi, S. Bengio. Machine Learning for Automatic Environmental Mapping: When and How, 2005
[4] : N. Gilardi, S. Bengio. Comparison of Four Machine Learning Algorithms for Spatial Data Analysis, 1997
[5] : V. Demyanov, M. Kanevsky, E. Savelieva, V. Timonin, S. Chernov, V. Polishchuk. Neural Network Residual Stochastic Cosimulation for Environmental Data Analysis,
[6] : V. Demyanov, M Kanevsky, S. Chernov, E. Savelieva, V. Timonin. Neural Network Residual Kriging Application for Cilmatic Data, 1998
[7] : A. Smola, B. Scholkopf. A Tutorial on Support Vector Regression, 1998
[8] : L. Breiman, J. Friedman, R. Olshen, C. Stone. Classification and Regression Trees, 1998
[9] : R. Polikar. Ensemble Based Systems in Decision Making, 2006
[10] : L. Rokach. Ensemble-Based Classifier, 2009
[11] : L. Breiman. Bagging Predictors, 1994
[12] : D. Opitz, R. Maclin. Popular Ensemble Methods: An Empirical Study, 1999
[13] : J. R. Quinlan. Bagging, Boosting, and C4.5, 2006
[14] : Y. Freund, R. Schapire. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting, 1996
[15] : G. Brown, J. Wyatt, R. Harris, X. Yao. Diversity Creation Methods: A Survey and Categorisation, 2005
[16] : J. Li, A. Heap, A. Potter, J. Daniell. Application of Machine Learning Methods to Spatial Interpolation of Environmental Variables, 2011
[17] : T. K. Ho. Random Decision Forests, 1995
[18] : L. Breiman. Random Forests, 2001
[19] : L. Breiman. Out-Of-Bag Estimation, 1996
[20] : E. Kalapanidas, N. Avouris, M. Craciun, D. Neagu. Machine Learning Algorithms: A Study on Noise Sensitivity, 2003
[21] : G. Dubois, J. Malczewski, M. De Cort. Mapping Radioactivity in the Environment - Spatial Interpolation Comparison 97, 2003
[22] : G. Dubois, S. Galmarini. Spatial Interpolation Comparison (SIC) 2004: Introduction to the Exercise and Overview of Results, 2005
[23] : T. Chai, R. R. Draxler. Root Mean Square Error (RMSE) or Mean Absolute Error (MAE)? - Arguements against avoiding RMSE in the Literature, 2014
[24] : J. Yarus, R. Chambers. Practical Geostatistics - An Armchair Overview for Petroleum Reservoir Engineers, 2006
[25] : C. Croarkin, P. Tobias, J. Filiben, B. Hembree, W. Guthrie, L. Trutna, J. Prins, C. Zey. NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/, 2003
[26] : P. Geurts, D. Ernst, L. Wehenkel. Extremely Randomized Trees, 2006
[27] : M. Robnik-Sikonja. Improving Random Forests, 2004
[28] : A. Davies, Z. Ghahramani. The Random Forest Kernel and Creating other Kernels for Big Data from Random Partitions, 2014