Εθνικο Μετσοβιο Πολυτεχνειο
Σχολη Ηλεκτρολογων Μηχανικων και Μηχανικων Υπολογιστων
Τομεας Τεχνολογιας Πληροφορικης και Υπολογιστων

Κατασκευή Σύνοψης για Συναθροιστικές Ερωτήσεις

Διπλωματικη Εργασια

της

ΣΤΑΜΑΤΙΑΣ Ε. ΡΙΖΟΥ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εργαστηριο Συστηματων Βασεων Γνωσεων και Δεδομενων
Αθήνα, Σεπτέμβριος 2007

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Κατασκευή Σύνοψης για Συναθροιστικές Ερωτήσεις

Διπλωματικη Εργασια

της

ΣΤΑΜΑΤΙΑΣ Ε. ΡΙΖΟΥ

Επιβλέπων: Τιμολέων Σελλής
Καθηγητής Ε.Μ.Π.

Εγκρίθηκε από την τριμελή εξεταστική επιτροπή την 4η Σεπτεμβρίου 2007.

(Υπογραφή)	*(Υπογραφή)*	*(Υπογραφή)*
.....	.....	.....
Τιμολέων Σελλής	Ιωάννης Βασιλείου	Ανδρέας Γεώργιος Σταφυλοπάτης
Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.

Αθήνα, Σεπτέμβριος 2007

(Υπογραφή)

.....

Σταματια Πιζογ

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

Εθνικό Μετσόβιο Πολυτεχνείο
Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Τομέας Τεχνολογίας Πληροφορικής και Υπολογιστών
Εργαστήριο Συστημάτων Βάσεων Γνώσεων και Δεδομένων

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

Περίληψη

Ο μετασχηματισμός κυματιδίων χρησιμοποιείται ευρέως για την συμπίεση χρονικών σειρών και πολυδιάστατων δεδομένων. Στην βιβλιογραφία έχουν προταθεί διάφορες παραλλαγές του απλού μετασχηματισμού (restricted Haar) όπως το unrestricted Haar καθώς και ο μετασχηματισμός Haar$^+$, που στόχο έχουν να δημιουργήσουν μία πιο ακριβής περίληψη των αρχικών δεδομένων. Οι υπάρχοντες αλγόριθμοι στοχεύουν στην ελαχιστοποίηση του σημειακού σφάλματος των δεδομένων. Στόχος της παρούσας εργασίας είναι να μελετηθεί και να προταθεί η λύση του προβλήματος βέλτιστης περίληψης ώστε να ελαχιστοποιηθεί το σφάλμα για όλες τις δυνατές συναθροιστικές ερωτήσεις. Ο αλγόριθμος που προτείνεται είναι γενικός και ανεξάρτητος από το είδος του μετασχηματισμού Haar.

Λέξεις Κλειδιά

Μετασχηματισμός Κυματιδίων, Μετρικές Σφάλματος, Κατανεμημένες Μετρικές Σφάλματος, Δέντρο Σφάλματος, Περίληψη Κυματιδίων, Σημειακό Σφάλμα, Συναθροιστικό Σφάλμα, Στιγμιότυπο Περίληψης, Διάνυσμα Σφάλματος, Αλγόριθμος Μερικής Διάταξης

Abstract

The Wavelet Transformation has been used widely in data compression of time series and multidimensional data. In the bibliography many versions of the classical Wavelet Transformation (restricted Haar) have been proposed such as unrestricted Haar and Haar$^+$ Transformation, whose target is to construct a more precise synopsis of the data. The existing algorithms minimize error metrics that calculate the point errors of the data. In this diploma thesis we study and present a solution for the problem of best synopses that minimize error metrics for all possible range sum queries. The proposed algorithm is general and indepedent from the kind of Wavelet Transformation.

Keywords

Wavelet Transformation, Unrestricted Haar, Error Metrics, Distributed Error Metric, Error Tree, Wavelet Synopses, Point Error, Range Sum Error, Synopses Instance, Error Vector, Partial Progarmming Algorithm

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον καθηγητή μου, κ. Τίμο Σελλή, για την ευκαιρία που μου έδωσε να μελετήσω ένα τόσο ενδιαφέρον πρόβλημα και για την πολύτιμη βοήθεια που μου έχει προσφέρει. Επίσης ευχαριστώ θερμά τον διδακτορικό φοιτητή Δημήτρη Σαχαρίδη για τα εύστοχα και εποικοδομητικά σχόλιά του κατά τη διάρκεια εκπόνησης της εργασίας. Τέλος θέλω να ευχαριστήσω την οικογένεια και τους φίλους μου που με στηρίζουν σε κάθε μου βήμα.

Κεφάλαιο 1

Εισαγωγή

Η ανάγκη για προσέγγιση των δεδομένων στις βάσεις προέκυψε ως αποτέλεσμα του μεγάλου χρόνου απόκρισης ερωτημάτων σε περιπτώσεις βάσεων μεγάλου μεγέθους (GB,TB) καθώς και της διαπίστωσης ότι σε πολλές εφαρμογές (OLAP/DSS συστήματα, βελτιστοποιητής ερωτημάτων) δεν είναι απαραίτητη μία ακριβής, παρά μία προσεγγιστική απάντηση. Προς αυτήν τη κατεύθυνση αναπτύχθηκαν διάφορες μέθοδοι με επικρατέστερες τα ιστογράμματα και τον μετασχηματισμό κυματιδίων (wavelet transform).

Τα ιστογράμματα διαμερίζουν το πεδίο ορισμού των δεδομένων σε διαστήματα (buckets), καθένα από τα οποία αντιπροσωπεύεται από μία και μόνο τιμή (συνήθως τον μέσο όρο των τιμών του διαστήματος). Η μέθοδος αυτή έχει ικανοποιητικά αποτελέσματα σε δεδομένα που έχουν κάποια ομοιομορφία αλλά αντίθετα υστερεί στην περίπτωση που τα δεδομένα παρουσιάζουν μεγάλες μεταβολές και ασυνέχειες.

Ο απλός μετασχηματισμός κυματιδίων (Haar wavelet) αντιπροτάθηκε αρχικά ως η εναλλακτική λύση για δεδομένα με απότομες αλλαγές τιμών. Η ιδέα προήλθε από την επιτυχημένη χρήση του μετασχηματισμού κυματιδίων στην επεξεργασία σήματος και εικόνας με στόχο την συμπίεση της πληροφορίας. Κατά αντιστοιχία με την εφαρμογή αυτή, στις βάσεις στόχος είναι η δημιουργία μίας όσο το δυνατόν πιο αντιπροσωπευτικής περίληψης των δεδομένων.

1.1 Αντικείμενο της διπλωματικής

Αντικείμενο της διπλωματικής είναι η μελέτη και επίλυση του προβλήματος βέλτιστης περίληψης για συναθροιστικές ερωτήσεις. Η εργασία αυτή είναι η πρώτη που μελετά ολοκληρωμένα τα συναθροιστικά σφάλματα αφού μέχρι στιγμής οι αλγόριθμοι που υπάρχουν δημιουργούν βέλτιστες περιλήψεις μόνο για τα σημειακά σφάλματα. Πιο συγκεκριμένα η διπλωματική αυτή περιλαμβάνει:

1. **Μελέτη των διαφόρων μετασχηματισμών Haar.** Γίνεται ανάλυση των κυριότερων μετασχηματισμών κυματιδίων και αναφορά των βασικών αλγορίθμων που λύνουν το πρόβλημα κατασκευής βέλτιστης σύνοψης για σημειακά ερωτήματα. Επιπλέον εισάγε-

ται και ένας ενιαίος συμβολισμός που παριστάνει αφαιρετικά όλα τα είδη μετασχηματισμών κυματιδίων.

2. **Θεωρητική μελέτη του προβλήματος.** Γίνεται η διατύπωση του γενικού προβλήματος της βελτιστοποίησης της περίληψης για συναθροιστικές ερωτήσεις ανεξάρτητα από το είδος του μετασχηματισμού κυματιδίων και προτείνεται αλγόριθμος για την επίλυση του γενικού προβλήματος. Επιπλέον παρέχεται το απαιτούμενο θεωρητικό υπόβαθρο, δηλαδή αποδεικνύονται κάποιες βασικές ιδιότητες που βοηθούν στην επίλυση του προβλήματος.

1.2 Διάρθρωση της διπλωματικής

Στο κεφάλαιο 2 αναλύουμε τα είδη του μετασχηματισμού Haar και αναφέρουμε τους βασικούς αλγορίθμους που επιλύουν το πρόβλημα βέλτιστης σύνοψης για σημειακά ερωτήματα. Επιπλέον εισάγουμε τον βασικό συμβολισμό που ενοποιεί σε ένα κοινό μοντέλο καθέναν από τους μετασχηματισμούς που αναφέρονται.

Στο κεφάλαιο 3 διατυπώνουμε το πρόβλημα και παρέχουμε το απαιτούμενο μαθηματικό υπόβαθρο που χρησιμοποιείται στην επίλυση του προβλήματος. Επιπλέον περιγράφουμε τον προτεινόμενο γενικό αλγόριθμο για συναθροιστικές ερωτήσεις. Στόχος είναι να δοθεί μία γενική μέθοδος επίλυσης του προβλήματος που θα διαφοροποιείται σε ορισμένα σημεία ανάλογα με το είδος του μετασχηματισμού χωρίς να αλλάζει η βασική ιδέα του αλγόριθμου. Τέλος αναφέρουμε την επέκταση του αλγορίθμου με βάση τον μετασχηματισμό Haar^+.

Τέλος στο κεφάλαιο 4 παρουσιάζονται κάποια γενικά συμπεράσματα από τη μελέτη και επίλυση του προβλήματος και προτείνονται πιθανά προβλήματα για μελλοντική εργασία.

Κεφάλαιο 2

Περίληψη Κυματιδίων

2.1 Εισαγωγή

Η βασική ιδέα του απλού μετασχηματισμού κυματιδίων είναι η εύρεση των μέσων τιμών ανά δύο και των αποστάσεων των δεδομένων από αυτές και στη συνέχεια επανάληψη της διαδικασίας αναδρομικά θεωρώντας ως δεδομένα τις μέσες τιμές, μέχρι να υπολογιστεί ο ολικός μέσος όρος των αρχικών δεδομένων. Με αυτόν τον τρόπο δημιουργείται ένα δυαδικό δέντρο με τόσους συντελεστές όσους και το πλήθος των αρχικών δεδομένων. Η ρίζα του δέντρου περιλαμβάνει τον συντελεστή που αντιστοιχεί στην ολική μέση τιμή. Κάθε συντελεστής που βρίσκεται στους εσωτερικούς κόμβους συνεισφέρει θετικά στα δεδομένα που βρίσκονται στο αριστερό υπόδεντρο και αρνητικά στα δεδομένα του δεξιού υποδέντρου. Στόχος των αλγορίθμων συμπίεσης είναι να κρατήσουν $B \ll N$ από τους N συντελεστές του δέντρου έτσι ώστε να ελαχιστοποιείται κάποια μετρική σφάλματος.

Μέχρι στιγμής έχει μελετηθεί εκτενώς το πρόβλημα εύρεσης βέλτιστης περίληψης για σημειακά ερωτήματα. Συγκεκριμένα έχει προταθεί η χρήση ενός άπληστου αλγορίθμου για την ελαχιστοποίηση του συνολικού τετραγωνικού σφάλματος (μετρική $\mathcal{L}_2$). Δυστυχώς το πρόβλημα δεν μπορεί να λυθεί για οποιαδήποτε μετρική σφάλματος με άπληστο αλγόριθμο. Οι Garofalakis και Kumar ασχολήθηκαν με το πρόβλημα και ανέπτυξαν έναν αλγόριθμο δυναμικού προγραμματισμού που βελτιστοποιεί την περίληψη για κάθε κατανεμημένη μετρική σφάλματος. Αργότερα προτάθηκε η χρήση του unrestricted Haar, δηλαδή να μην ορίζονται οι τιμές των κόμβων μονοσήμαντα με βάση τις μέσες τιμές. Η ιδέα αυτή έδωσε ώθηση για περαιτέρω έρευνα στην περιοχή αυτή. Το Haar$^+$ γενικεύει ακόμη περισσότερο το unrestricted Haar προσθέτοντας σε κάθε συντελεστή του δέντρου (συντελεστής -κεφαλή) δύο επιπλέον συμπληρωματικούς κόμβους (αριστερά και δεξιά) που συνεισφέρουν πάντα θετικά την τιμή τους στο ανακτώμενο άθροισμα.

2.2 Μετρικές Σφάλματος

Στην ενότητα αυτή θα αναφέρουμε κάποιες βασικές έννοιες σχετικά με τις διάφορες μετρικές σφάλματος. Οι μετρικές σφάλματος σχετίζονται άμεσα με την δημιουργία περιλήψεων

κυματιδίων αφού στόχος κάθε αλγορίθμου συμπίεσης είναι να κρατήσει $B \ll N$ από τους N συντελεστές του δέντρου έτσι ώστε να ελαχιστοποιείται κάποια μετρική σφάλματος. Κάποιες συνήθεις μετρικές σφάλματος είναι το απόλυτο και το σχετικό σφάλμα που ορίζονται από τους παρακάτω τύπους αντίστοιχα:

$$\mathcal{L}_{abs}(i) = \left| D[i] - \hat{D}[i] \right|$$

$$\mathcal{L}_{rel}(i) = \frac{\left| D[i] - \hat{D}[i] \right|}{\max\{s, D[i]\}}$$

όπου $\hat{D}$ είναι το ανακατασκευασμένο με βάση την περίληψη διάνυσμα των δεδομένων και s ένας αριθμός που λειτουργεί σαν κατώφλι ώστε να μην κυριαρχούν πολύ μικρές τιμές στο σχετικό σφάλμα. Με την βοήθεια των παραπάνω τύπων μπορούμε να βρούμε το σφάλμα κάθε στοιχείου του πίνακα. Στη συνέχεια για να υπολογίσουμε το σφάλμα όλου του διανύσματος αθροίζουμε τα επιμέρους σφάλματα ή βρίσκουμε το μέγιστο από αυτά. Σε κάθε περίπτωση χρειάζεται να υπολογίσουμε το σφάλμα ξεχωριστά για κάθε σημείο και στη συνέχεια να το συνδυάσουμε με τα υπόλοιπα για την εύρεση του ολικού σφάλματος. Γενικά στους τύπους των μετρικών σφάλματος που περιγράψαμε παραπάνω μπορεί να εμφανίζονται και βάρη που να διαφοροποιούν την επίδραση του κάθε σημείου για τον υπολογισμό του ολικού σφάλματος. Παρακάτω παραθέτουμε μία πιο γενική μορφή των κανονικοποιημένων μετρικών σφάλματος:

$$\mathcal{L}_p^w(D, \hat{D}) = \sum_i \sqrt[p]{\frac{\left(w(i) \left| D[i] - \hat{D}[i] \right|\right)^p}{n}}$$

Οι παραπάνω μετρικές σφάλματος ανήκουν σε μία ευρύτερη κατηγορία που χαρακτηρίζονται ως κατανεμημένες μετρικές σφάλματος.Σε αυτό το σημείο είναι σημαντικό να εισάγουμε την έννοια της κατανεμημένης μετρικής σφάλματος.

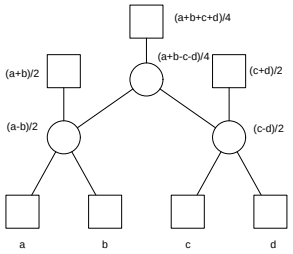
Ορισμός. Έστω $\hat{D}$ το ανακατασκευασμένο με βάση την περίληψη διάνυσμα των δεδομένων. Με $f(R)$ συμβολίζουμε το σφάλμα που περιέχει η περίληψη για όλο το εύρος τιμών R του D. Λέμε ότι η μετρική σφάλματος f είναι κατανεμημένη αν και μόνο αν, για κάθε διαμέριση ξένων διαστημάτων $R_1, R_2, \ldots, R_k$ υπάρχει συνδυαστική συνάρτηση g τέτοια ώστε το σφάλμα της ένωσης των επιμέρους διαστημάτων $\bigcup_{i=1}^{k} R_i$ α μπορεί να εκφραστεί ως:

$$f(\bigcup_{i=1}^{k} R_i) = g(f(R_1), f(R_2), \ldots, f(R_k))$$

2.3 Μετασχηματισμός Haar

2.3.1 Το Δέντρο Haar

Ο μετασχηματισμός Haar όπως περιγράφτηκε και στην εισαγωγή δημιουργεί ένα δυαδικό δέντρο και τοποθετεί στα φύλλα τα δεδομένα, στους εσωτερικούς κόμβους τις αποκλίσεις από την μέση τιμή και στην ρίζα του δέντρου τον ολικό μέσο όρο. Έστω, λοιπόν, ότι έχουμε το διάνυσμα τιμών A, μεγέθους N, με το N να είναι μια δύναμη του 2. Στο πρώτο βήμα



Σχήμα 2.1: Παράδειγμα Μ/Σ Haar για το διανύσμα (a, b, c, d).

σχηματίζουμε ένα νέο διάνυσμα $S_1[0 \ldots (N/2-1)]$ μεγέθους $N/2$ επιλέγοντας διαδοχικά ζεύγη τιμών από το A και υπολογίζοντας το ημιάθροισμά τους. Με τον ίδιο τρόπο σχηματίζουμε το διάνυσμα ημιδιαφορών $D_1[0 \ldots (N/2-1)]$, δηλαδή για διαδοχικά ζεύγη τιμών του A υπολογίζουμε την ημιδιαφορά τους και την τοποθετούμε στο D. Παρατηρούμε ότι ως εδώ δεν έχουμε απώλεια πληροφορίας, αφού κάθε στοιχείο του αρχικού διανύσματος ανακτάται εύκολα από τα στοιχεία των νέων διανυσμάτων. Για παράδειγμα, $A[0] = S[0] + D[0] = (A[0]+A[1])/2+(A[0]-A[1])/2$, $A[1] = S[0]-D[0] = (A[0]+A[1])/2-(A[0]-A[1])/2$, κοκ. Η διαδικασία που περιγράψαμε επαναλαμβάνεται αναδρομικά για το διάνυσμα ημιαθροισμάτων, μέχρι να φτάσουμε σε διανύσματα μεγέθους 1 . Το συνολικό πλήθος ημιδιαφορών που υπολογίζουμε είναι $1 + 2 + \ldots + N/2 = N - 1$. Από αυτή τη διαδικασία σχηματίζεται το διάνυσμα W μεγέθους N, το οποίο αποτελείται από τό μέσο όρο των τιμών του αρχικού σήματος (το τελευταίο ημιάθροισμα που υπολογίζουμε) και τις $N-1$ ημιδιαφορές. Το διάνυσμα αυτό είναι ο μη-κανονικοποιημένος μετασχηματισμός Haar του A. Μπορούμε διαισθητικά να καταλάβουμε ότι για να έχουμε κανονικοποιημένο αποτέλεσμα πρέπει να δώσουμε περισσότερο βάρος στα τελευταία από τα $\log N$ βήματα, καθώς αυτά αφορούν περισσότερα στοιχεία του αρχικού σήματος. Έτσι, πολλαπλασιάζουμε κάθε συντελεστή που υπολογίστηκε στο k-οστό βήμα με $\sqrt{N/2^{\log N-k}}$ και παράγουμε τον κανονικοποιημένο μετασχηματισμό C (συντελεστές wavelet) του A.

Θεωρούμε ένα σήμα A μεγέθους $N = 4$, με $A = (a, b, c, d)$. Με βάση τα όσα προηγήθηκαν για να βρούμε τον Μ/Σ Haar, υπολογίζουμε διαδοχικά τα διανύσματα ημιαθροισμάτων και ημιδιαφορών (βλ. Σχήμα 2.1).

$$S_1 = [\frac{a+b}{2}, \frac{c+d}{2}] \quad D_1 = [\frac{a-b}{2}, \frac{c-d}{2}]$$

$$S_2 = [\frac{a+b+c+d}{4}] \quad D_2 = [\frac{a+b-c-d}{4}]$$

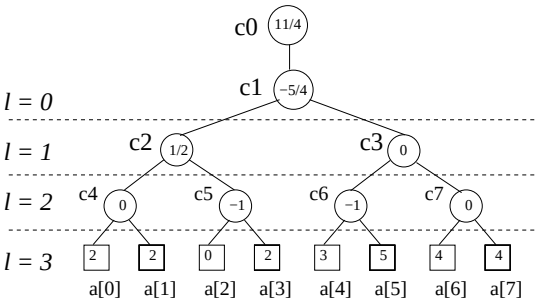
Έτσι, παίρνουμε το διάνυσμα $W = [\frac{a+b+c+d}{4}, \frac{a+b-c-d}{4}, \frac{a-b}{2}, \frac{c-d}{2}]$. Το αρχικό σήμα A μπορεί να ανακατασκευαστεί από τους μη-κανονικοποιημένους συντελεστές $(W[N])$.

$$A[0] = W[0] + W[1] + W[2] = \frac{a+b+c+d}{4} + \frac{a+b-c-d}{4} + \frac{a-b}{2} = a$$

$$A[1] = W[0] + W[1] - W[2] = \frac{a+b+c+d}{4} + \frac{a+b-c-d}{4} - \frac{a-b}{2} = b$$

Resolution	Averages	Detail Coefficients
3	[2, 2, 0, 2, 3, 5, 4, 4]	-----
2	[2, 1, 4, 4]	[0, −1, −1, 0]
1	[3/2, 4]	[1/2, 0]
0	[11/4]	[−5/4]

(a)



(b)

Σχήμα 2.2: Παράδειγμα δένδρου σφάλματος για το διανύσμα a.

$$A[0] = W[0] - W[1] + W[3] = \frac{a+b+c+d}{4} - \frac{a+b-c-d}{4} + \frac{c-d}{2} = c$$

$$A[0] = W[0] - W[1] - W[3] = \frac{a+b+c+d}{4} - \frac{a+b-c-d}{4} - \frac{c-d}{2} = c$$

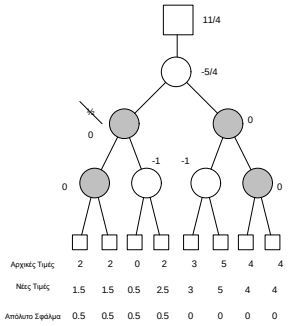
Θα προχωρήσουμε με ένα αριθμητικό παράδειγμα. Έστω, λοιπόν, ότι έχουμε το διάνυσμα τιμών $a = [2, 2, 0, 2, 3, 5, 4, 4]$, μεγέθους $N = 8$ (βλ. Σχήμα 2.2). Ο μη-κανονικοποιημένος μετασχηματισμός του a είναι το διάνυσμα $w_a = [11/4, -5/4, 1/2, 0, 0, -1, -1, 0]$ μεγέθους $N = 8$. Αποτελείται από το μέσο όρο των τιμών του a και από τις τιμές των διανυσμάτων ημιδιαφορών. Οι τιμές αυτές αποτελούν τους συντελεστές του μετασχηματισμού wavelet ενώ οι ημιδιαφορές ονομάζονται και *λεπτομέρειες* (detail coefficients). Στο παράδειγμά μας, ο μέσος όρος των τιμών του a είναι $11/4$, η λεπτομέρεια επιπέδου $l = 0$ είναι $-5/4$, οι λεπτομέρειες επιπέδου $l = 1$ είναι $1/2$ και 0 και οι λεπτομέρειες επιπέδου $l = 2$ είναι $0, -1, -1$ και 0.

Ένας καλός τρόπος να παραστήσουμε και να κατανοήσουμε την ιεραρχική φύση του μετασχηματισμού Haar είναι το δένδρο σφαλμάτων, όπως αυτό φαίνεται στο Σχήμα 2.2(b). Η ρίζα του δένδρου, c_0, είναι ο μέσος όρος των τιμών, οι εσωτερικοί κόμβοι αντιστοιχούν στους υπόλοιπους συντελεστές (λεπτομέρειες) και τα φύλλα αντιστοιχούν στα αρχικά δεδομένα. Παρατηρήστε ότι η τιμή ενός φύλλου μπορεί να ανασκευαστεί από τις τιμές των $\log N + 1$ εσωτερικών κόμβων που βρίσκονται στο μονοπάτι από τη ρίζα προς το φύλλο. Για παράδειγμα, $a[5] = c_0 - c_1 + c_3 - c_6 \Leftrightarrow 5 = \frac{11}{4} - \left(-\frac{5}{4}\right) + 0 - (-1)$. Το πρόσημο του όρου στο άθροισμα είναι $+$ ή $-$ όταν το φύλλο βρίσκεται στο αριστερό ή το δεξί φύλλο του όρου, αντίστοιχα.

2.3.2 Κατασκευή Περιλήψεων

Όπως μπορούμε να παρατηρήσουμε από την περιγραφή του M/Σ Haar, όταν γειτονικά δεδομένα έχουν παρόμοιες τιμές, παράγονται συντελεστές - λεπτομέρειες με μικρό μέτρο (κοντά στο 0). Αν θέσουμε την τιμή αυτών των συντελεστών ίση με 0 περιμένουμε ότι το σφάλμα που θα παρουσιάζεται στα νέα ανακατασκευασμένα δεδομένα, θα είναι 'μικρό'. Αυτή η ιδιότητα του M/Σ Haar τον κάνει κατάλληλο στη χρήση του για κατασκευή περιλήψεων δεδομένων.

Μια περίληψη B όρων $\hat{C}$ ορίζεται επιλέγοντας ένα σύνολο $\Lambda \subset C$ συντελεστών, με $B = |\Lambda| \ll N$, ενώ οι υπόλοιποι $N - B$ όροι θεωρούνται ίσοι με μηδέν.Το πόσο 'μικρό'



Σχήμα 2.3: Παράδειγμα περίληψης για τη μετρική L_∞.

είναι το σφάλμα που εισάγεται στα δεδομένα μας όταν επιλέξουμε να θεωρήσουμε κάποιους συντελεστές ίσους με 0 και να κρατήσουμε την τιμή των υπολοίπων, μπορεί να μετρηθεί με διάφορους τρόπους. Το σφάλμα αυτό υπολογίζεται από συναρτήσεις που ονομάζονται μετρικές σφάλματος. Αν A είναι το αρχικό σήμα και $\hat{A}$ το ανακατασκευασμένο από την περίληψη σήμα, η μετρική σφάλματος στην ουσία μας παρέχει ένα μέτρο του διανύσματος $A - \hat{A}$.

$$error = f_{metric}(A - \hat{A}) = ||A - \hat{A}||_{f_{metric}}$$

Το πρόβλημα που καλούμαστε να λύσουμε δεδομένων του διαθέσιμου χώρου περίληψης B και μιας μετρικής $f_{metric}()$ είναι να επιλέξουμε τους B όρους από το μετασχηματισμένο σήμα ώστε το σφάλμα που δίνει η μετρική να είναι ελάχιστο.

Μία μετρική είναι η L_∞ ή μετρική μέγιστου απόλυτου σφάλματος.

$$L_\infty = maxAbsErr(A - \hat{A}) = \max_{0 \le i < N} |\hat{A}[i] - A[i]|$$

Αν υποθέσουμε ότι έχουμε το παράδειγμα του σχήματος 2.3 και θέλουμε να φτάξουμε μια περίληψη με $B = 4$ συντελεστές ώστε να ελαχιστοποιείται η μετρική L_∞, επιλέγουμε να αγνοήσουμε τους συντελεστές c_2, c_3, c_4 και c_7 (οι τρεις τελευταίοι είναι ήδη μηδενικοί) με μέγιστο απόλυτο σφάλμα 1/2.

2.3.3 Κατασκευή Σύνοψης για σημειάκα σφάλματα πάνω από Μ/Σ Haar

Εισαγωγή

Στην ενότητα αυτή εξετάζουμε αλγορίθμους που υπολογίζουν το ελάχιστο σφάλμα που μπορεί να περιέχει μια περίληψη B όρων, αφού πρώτα εφαρμοστεί ο κλασσικός μετασχηματισμός Haar στο αρχικό σήμα. Οι αλγόριθμοι αυτοί, δηλαδή, δέχονται ως είσοδο το αρχικό σήμα μετασχηματισμένο με τον κλασσικό ΜΣ Haar και το μέγιστο επιτρεπόμενο μέγεθος περίληψης B και δίνουν ως έξοδο το ελάχιστο δυνατό σφάλμα, υπολογισμένο με τη χρήση μιας μετρικής. Η μετρική που χρησιμοποιείται ανήκει στις weighted-Lp μετρικές για point errors. Η επιλογή των συντελεστών με τους οποίους επιτυγχάνεται η βέλτιστη περίληψη, μπορεί να γίνεται προγραμματιστικά είτε σε ένα δεύτερο πέρασμα, αφού πρώτα υπολογιστεί το ελάχιστο σφάλμα, είτε παράλληλα με τον υπολογισμό του ελάχιστου σφάλματος. Ενδεικτικά θα αναφερθούμε στους δύο πιο βασικούς αλγορίθμους που λύνουν το πρόβλημα εύρεσης βέλτιστης σύνοψης για σημειακά δεδομένα. Οι δύο αυτοί αλγόριθμοι αντιπροσωπεύουν και δύο διαφορετικές τεχνικές προγραμματισμού. Αρχικά παρουσιάζουμε έναν άπληστο αλγόριθμο που λύνει το πρόβλημα μόνο για το τετραγωνικό σφάλμα και στην συνέχεια έναν αλγόριθμο δυναμικού προγραμματισμού που βρίσκει τη βέλτιστη σύνοψη για κάθε κατανεμημένη μετρική σφάλματος. Για λόγους απλότητας, η επιλογή των συντελεστών δεν περιγράφεται στην περιγραφή των περισσότερων αλγορίθμων που ακολουθούν.

Ο Άπληστος αλγόριθμος επιλογής συντελεστών για τη μετρική L_2

Στην ενότητα αυτή παρουσιάζεται ο απλούστερος αλγόριθμος περίληψης. Ο αλγόριθμος αυτός κρατάει στην περίληψη τους B μεγαλύτερους κατ' απόλυτη τιμή συντελεστές, πρόκειται δηλαδή για έναν άπληστο αλγόριθμο επιλογής συντελεστών. Όταν εφαρμόζεται σε σύνολα συντελεστών που έχουν προκύψει από τον κλασσικό Μ/Σ Haar ελαχιστοποιεί τη μετρική σφάλματος L_2. Η μετρική σφάλματος L_2 ή SSE - Sum of Squared Errors όπως ονομάζεται ορίζεται ως εξής:

$$SSE(e) = \sum_{i=0}^{N-1} e_i^2$$

με A το αρχικό σήμα-διάνυσμα, C το μετασχηματισμένο και κανονικοποιημένο διάνυσμα, $\hat{C}$ η περίληψη Λ, $\hat{A} = W^{-1}\{\hat{C}_A\}$ τα δεδομένα όπως επανακτώνται μετά την περίληψη και $e = A - \hat{A}$ το διάνυσμα σφάλματος. Η μετρική L_2 ταυτίζεται με το τετράγωνο της ευκλείδειας νόρμας του διανύσματος σφάλματος, δηλαδή ισούται με το άθροισμα των τετραγώνων των (κανονικοποιημένων) συντελεστών που μένουν εκτός της περίληψης. Έτσι, το SSE γίνεται ελάχιστο επιλέγοντας τους B μεγαλύτερους κατ' απόλυτη τιμή συντελεστές.

$$SSE = L_2(e) = ||e||^2 = \sum_{i \notin \Lambda} c_i^2$$

Ο αλγόριθμος δουλεύει ως εξής: αρχικά, σε χρόνο $O(N)$, φτιάχνουμε ένα σωρό (max-

heap, δηλαδή δυαδικό δένδρο στο οποίο κάθε κόμβος έχει μεγαλύτερη τιμή από τα παιδιά του) των N κανονικοποιημένων συντελεστών του μετασχηματισμού, υπολογίζουμε και αποθηκεύουμε το άθροισμα των τετραγώνων όλων των συντελεστών (έστω Sum). Στη συνέχεια, επιλέγουμε και αφαιρούμε το συντελεστή που βρίσκεται στην κορυφή του σωρού και σε χρόνο $O(\log N)$ ξαναφτιάχνουμε το σωρό. Το βήμα αυτό επαναλαμβάνεται B φορές. Το σφάλμα υπολογίζεται αφαιρώντας κατά την επιλογή των συντελεστών τα τετράγωνά τους από το Sum. Η πολυπλοκότητα του αλγορίθμου είναι $O(N)$ σε χώρο και $O(N + B \log N)$.

Ο Αλγόριθμος των Garofalakis και Kumar για τη μετρική L_∞

Η μετρική σφάλματος L_∞ μας δίνει το μέγιστο από τα σφάλματα που περιέχει μια περίληψη για τα στοιχεία του αρχικού σήματος (μέγιστο σημειακό σφάλμα). Το σφάλμα αυτό μπορεί να είναι απόλυτο ή σχετικό. Έχουμε, λοιπόν δύο δυνατούς ορισμούς για αυτή τη μετρική

$$L_\infty = absErr(a - \hat{a}) = \max_{1 \le i \le N} |\hat{a}[i] - a[i]|$$

και

$$L_\infty = relErr(a - \hat{a}) = \max_{1 \le i \le N} \frac{|\hat{a}[i] - a[i]|}{max\{|a[i]|, s\}}$$

όπου s μια σταθερά εκλογίκευσης του σχετικού σφάλματος που δεν επιτρέπει σε πολύ μικρές τιμές του σήματος να κυριαρχούν στη μέτρηση του σφάλματος. Το πρόβλημά που αντιμετωπίζουμε, λοιπόν, είναι η επιλογή Β συντελεστών της περίληψης ώστε να ελαχιστοποιείται το μέγιστο σημειακό σφάλμα. Η επιλογή του ενός ή του άλλου ορισμού δεν επηρεάζει σημαντικά τη λύση του προβλήματος.

Η αρχική μορφή του αλγορίθμου

Οι Garofalakis και Kumar στο [2] προτείνουν έναν αλγόριθμο δυναμικού προγραμματισμού (MinMaxErr), που λύνει το πρόβλημα για τη μετρική L_∞. Η βασική ιδέα του αλγορίθμου είναι ότι λύνει το (υπο)πρόβλημα της επιλογής συντελεστών για το υποδέντρο με ρίζα τον κόμβο c_i λαμβάνοντας υπόψιν την επιλογή συντελεστών που έχει γίνει για το μονοπάτι από τη ρίζα του δένδρου (κόμβος c_0).

Προτού προχωρήσουμε στην περιγραφή του αλγορίθμου δίνουμε κάποιους ορισμούς και παραδοχές που θα χρησιμοποιήσουμε. Μας δίνονται το αρχικό σήμα μαζί με το μετασχηματισμένο και κανονικοποιημένο σήμα (μέσω του δένδρου σφάλματος) και το πλήθος Β των συντελεστών που θα αποτελέσουν την περίληψή μας. Με T_j συμβολίζουμε το υποδέντρο σφάλματος που έχει ρίζα τον κόμβο c_j και ως $coeff(T_j)$ και $data(T_j)$ ορίζουμε τα σύνολα των συντελεστών (εσωτερικοί κόμβοι) και των αρχικών δεδομένων (φύλλα), αντίστοιχα, που ανήκουν στο T_j. Με $path(c_j)$ συμβολίζουμε το σύνολο των συντελεστών που ανήκουν στο μονοπάτι από τη ρίζα του δένδρου σφαλμάτων (c_0) ως τον κόμβο c_j (χωρίς αυτόν). Τέλος, με $M[j, b, S]$ συμβολίζουμε την ελάχιστη τιμή του μέγιστου σημειακού σφάλματος που εισέρχεται στην περίληψη επιλέγοντας b συντελεστές του T_j με την υπόθεση ότι έχουμε ήδη επιλέξει ένα σύνολο $S \subseteq path(c_j)$ μεγέθους το πολύ $min\{B - b, \log N + 1\}$. Έτσι, θεωρώντας ότι

δουλεύουμε με το σχετικό σφάλμα,

$$M[j,b,S] = \min_{S_j \subseteq coeff(T_j), \|S_j\| \leq b} \{ \max_{d_i \in data(T_j)} relErr_i \}$$

με

$$relErr_i = \frac{|d_i - \sum_{c_k \in path(d_i) \cap S_i \cup S} sign_{i,k} \cdot c_k|}{max\{|d_i|, s\}}.$$

Με παρόμοιο τρόπο αντιμετωπίζουμε και το απόλυτο σφάλμα.

Το επιθυμητό αποτέλεσμα δίνεται από την τιμή του $M[0, B, \emptyset]$. Η βάση της αναδρομής βρίσκεται στα φύλλα, δηλαδή τους κόμβους $c_j = d_{j-N+1}$, για $j \geq N$. Τα φύλλα φυσικά δεν ανήκουν στην περίληψη οπότε έχουμε $b = 0$.

$$M[j,0,S] = \frac{|d_{j-N+1} - \sum_{c_k \in S} sign_{j-N,k} \cdot c_k|}{max\{|d_{j-N+1}, s\}}$$

Για τους εσωτερικούς κόμβους του δένδρου σφάλματος ο αλγόριθμος εξετάζει δύο επιλογές: να κρατήσει τον κόμβο c_j στην περίληψη ή να τον απορρίψει. Αν τον απορρίψει, τότε το ελάχιστο μέγιστο σφάλμα για το T_j είναι το μεγαλύτερο από τα ελάχιστα μέγιστα σφάλματα για τα υποδένδρα T_{2j} και T_{2j+1}. Ο συνολικός χώρος περίληψης και οι προεπιλεγμένοι κόμβοι που μπορούν να εκμεταλλευτούν τα δύο υποδένδρα είναι ίδιοι με του T_j.

$$M_{drop}[j,b,S] = \min_{0 \leq \dot{b} \leq b} max\{M[2j, \dot{b}, S], M[2j+1, b-\dot{b}, S]\}$$

Αν από την άλλη κρατήσει τον κόμβο c_j, τότε κατά τον υπολογισμό του ελάχιστου μέγιστου σφάλματος των T_{2j} και T_{2j+1} ο κόμβος c_j προστίθεται στους προεπιλεγμένους κόμβους ενώ προσαρμόζεται και ο διαθέσιμος χώρος περίληψης.

$$M_{keep}[j,b,S] = \min_{0 \leq \dot{b} \leq b-1} max\{M[2j, \dot{b}, S \cup \{c_j\}], M[2j+1, b-\dot{b}-1, S \cup \{c_j\}]\}$$

Τελικά ο αλγόριθμος επιλέγει την καλύτερη από τις δύο επιλογές.

$$M[j,b,S] = min\{M_{drop}[j,b,S], M_{keep}[j,b,S]\}$$

Για ένα συγκεκριμένο κόμβο c_j επιπέδου l στο δένδρο σφάλματος, ο αλγόριθμος έχει να εξετάσει το πολύ $B+1$ περιπτώσεις όσον αφορά το πλήθος των συντελεστών που θα κρατήσει στο υποδέντρο T_j (συνυπολογίζοντας την περίπτωση να κρατήσει 0 συντελεστές). Ακόμα, για έναν κόμβο επιπέδου l υπάρχουν 2^l υποσύνολα προγόνων να εξεταστούν. Έτσι, στον κόμβο c_j αντιστοιχούν $O(B2^l)$ τιμές του πίνακα $M[]$. Αφού υπάρχουν 2^l κόμβοι επιπέδου l, ο συνολικός χώρος του αλγορίθμου είναι

$$\sum_{l=0}^{l=\log N} 2^l B 2^l = O(N^2 B).$$

Ακόμα, για να υπολογίσουμε το κάθε στοιχείο του πίνακα χρειαζόμαστε $O(\log B)$ χρόνο: αφού το $M[2j, \dot{b}, S]$ είναι φθίνουσα συνάρτηση του $\dot{b}$ ενώ το $M[2j+1, b-\dot{b}, S]$ είναι αύξουσα συνάρτηση του $\dot{b}$, μπορούμε να εκτελέσουμε δυαδική αναζήτηση για το $\dot{b}$, ώστε να βρούμε το σημείο που τα δύο σφάλματα γίνονται ίσα (και άρα το σφάλμα του κόμβου-γονέα γίνεται ελάχιστο). Όπως και προηγουμένως, υπολογίζουμε ότι ο συνολικός χρόνος του αλγορίθμου είναι $O(N^2 B \log B)$. Έτσι, προέκυψε ότι η πολυπλοκότητα του αλγορίθμου είναι $O(N^2 B \log B)$ σε χρόνο και $O(N^2 B)$ σε χώρο.

Οι παρατηρήσεις του Guha

Ο Guha στο [3] αποδεικνύει ότι ο αλγόριθμος των Garofalakis και Kumar μπορεί να βελτιωθεί ως προς την πολυπλοκότητα. Η βασική του παρατήρηση είναι ότι δοθέντος ενός εσωτερικού κόμβου c_j επιπέδου l -άρα με $l+1$ προγόνους - και διαθέσιμου χώρου περίληψης B, το μέγιστο πλήθος των συντελεστών του δένδρου T_j που μπορούν να προστεθούν στην περίληψη είναι $\min\{B, t\}$, όπου t το πλήθος των κόμβων που ανήκουν στο υποδέντρο με ρίζα το c_j, συμπεριλαμβανομένου του ιδίου. Είναι $t = t_0 = 0$ για τη ρίζα του δένδρου σφάλματος και $t = t_1 = 2^{\log N - l} - 1$ για τους υπόλοιπους κόμβους. Ο κόμβος καλείται $2^l < 2N$ φορές, όσα και τα διαφορετικά μονοπάτια απογόνων από τη ρίζα προς τον κόμβο, ανάλογα με το αν ένας πατρικός κόμβος έχει μείνει στην περίληψη ή όχι. Ο συνολικός χρόνος που απαιτείται για έναν κόμβο είναι, λοιπόν, $2^l min\{B, t\} \log \min\{B, t\}$ και αφού οι κόμβοι επιπέδου l είναι 2^l στο πλήθος (εκτός από το επίπεδο $l = 0$ έχουμε 2 κόμβους), ο συνολικός χρόνος του αλγορίθμου είναι

$$\sum_{l=0}^{\log N} (2^l 2^l \min\{B, t_1\} \log \min\{B, t_1\}) + \min\{B, t_0\} \log \min\{B, t_0\}.$$

Στη χειρότερη περίπτωση (worst case) είναι $B = N \geq t$. Ακόμα, ισχύει $2^l(t+1) = 2N \Rightarrow 2^l t = O(2N)$. Έτσι, η πολυπλοκότητα είναι (θέτοντας $r = \log N - l$)

$$O(\sum_{l=0}^{\log N} 2^l 2^l t_1 \log t_1 + t_0 \log t_0) = O(\sum_{l=0}^{\log N} (2^l 2^l (2^{\log N - l} - 1) \log (2^{\log N - l} - 1)) + N \log N) =$$

$$= O(\sum_{l=0}^{\log N} N 2^l \log \frac{N}{2^l} + N \log N) = O(\sum_{r=0}^{\log N} \frac{N^2 r}{2^r} + N \log N) = O(N^2 + N \log N) = O(N^2).$$

Όσον αφορά τη χωρική πολυπλοκότητα, για έναν κόμβο απαιτούνται $O(2^l \min\{B, 2^{\log N - l}\}) = O(\min\{B 2^l, N\})$ θέσεις στον πίνακα $M[]$. Συνολικά, ο χώρος που καταλαμβάνει ο αλγόριθμος είναι

$$O(\sum_{l=0}^{\log N} 2^l min\{B 2^l, N\}) \leq O(N \sum_{l=0}^{\log N} 2^l) = O(N^2)$$

2.4 Unrestricted Haar

Στην προηγούμενη παράγραφο εξετάσαμε τον απλό M/Σ Haar. Όπως είδαμε στον απλό M/Σ Haar οι συντελεστές του δέντρου Haar ορίζονται μονοσήμαντα με βάση τις ημιδιαφορές των συντελεστών του προηγούμενου επιπέδου, οπότε οι αλγόριθμοι γνωρίζοντας το δέντρο Haar επιλέγουν αν θα κρατήσουν ή θα απορρίψουν κάποιο συντελεστή του δέντρου. Οι Guha και Harb στα [4, 5] έκαναν την διαπίστωση ότι κρατώντας τους αρχικούς συντελεστές οδηγούμαστε σε μη βέλτιστες λύσεις για μία μεγάλη κατηγορία από μετρικές σφάλματος και εισήγαγαν νέους αλγορίθμους για την κατασκευή περίληψης κυματιδίων. Κάποιες επιπλέον βελτιώσεις στην πολυπλοκότητα των αλγορίθμων προκύπτουν από τη λύση του δυαδικού προβλήματος όπως φαίνεται στο [7]. Το αποτέλεσμα των παραπάνω μελετών ήταν η εισαγωγή ενός νέου μετασχηματισμού, του M/Σ unrestricted Haar που αποτελεί μία επέκταση του απλού μετασχηματισμού και έχει στόχο να προσφέρει μεγαλύτερη ευελιξία στις τιμές των συντελεστών του δέντρου. Πλέον οι τιμές των συντελεστών δεν είναι γνωστές εξαρχής. Κάθε κόμβος του δέντρου μπορεί να πάρει τιμές σε ένα διάστημα που ορίζεται με βάση την εισερχόμενη τιμή στον κόμβο και το ελάχιστο και μέγιστο των δεδομένων που υπάρχουν υπό την εμβέλεια του. Αυτό σημαίνει ότι για κάθε συντελεστή θα πρέπει να εξετάζονται όλες οι δυνατές τιμές που μπορεί να πάρει καθώς και η περίπτωση μηδενισμού του. Θα πρέπει να τονισθεί ότι η μόνη διαφορά σε σχέση με τον απλό M/Σ είναι το γεγονός ότι ο κάθε συντελεστής μπορεί να παίρνει περισσότερες από μία μηδενικές τιμές. Κατά τα αλλά η δομή και οι ιδιότητες του δέντρου παραμένουν ίδιες. Και σε αυτή την περίπτωση κάθε συντελεστής προσφέρει την τιμή του στο άθροισμα ανάκτησης των δεδομένων θετικά αν το στοιχείο που ανακτάται είναι στο αριστερό υπόδεντρο και αρνητικά αν το στοιχείο βρίσκεται στο δεξί υπόδεντρο.

Ως ένα απλό παράδειγμα εφαρμογής του M/Σ unrestricted Haar θεωρούμε ότι θέλουμε να βρούμε μία βέλτιστη σύνοψη που να κρατά ένα μόνο συντελεστή, δηλαδή $B = 1$ έχοντας ως διάνυσμα δεδομένων το $D = 2, 10, 12, 8$. Αν χρησιμοποιήσουμε τον M/Σ unrestricted Haar η περίληψη θα είναι ο ολικός μέσος όρος με τιμή 7 με μέγιστο απόλυτο σφάλμα 5. Σε περίπτωση που χρησιμοποιούσαμε τον απλό μετασχηματισμό θα λαμβάναμε ως βέλτιστη περίληψη και πάλι τον ολικό μέσο όρο αλλά με τιμή 8 και με μέγιστο απόλυτο σφάλμα 6. Από το απλό αυτό παράδειγμα φαίνεται ότι το γεγονός ότι οι συντελεστές μπορούν να παίρνουν διάφορες τιμές είναι ικανό να βελτιώσει το σφάλμα.

Επιλέγουμε να μην αναφερθούμε με μεγαλύτερη λεπτομέρεια σε αυτόν τον μετασχηματισμό γιατί ακολουθεί αναλυτική περιγραφή του M/Σ Haar$^+$ που είναι μία επέκταση του unrestricted Haar. Εφόσον το unrestricted Haar είναι υποπερίπτωση του Haar$^+$ οι τεχνικές που χρησιμοποιούνται για την κατασκευή περίληψης θα είναι ανάλογες.

2.5 Μετασχηματισμός Haar$^+$

2.5.1 Το Δέντρο Haar$^+$

Οι Karras και Mamoulis στο [6] εισήγαγαν τον M/Σ Haar$^+$ που ενισχύει και επεκτείνει τον απλό μετασχηματισμό Haar. Για να περιγράψουμε τον M/Σ, θεωρούμε το δέντρο Haar$^+$ στην περίπτωση που το σύνολο των δεδομένων αποτελείται από τέσσερα στοιχεία $\{d_0, d_1, d_2, d_3\}$. Η δομή που δημιουργείται είναι παρόμοια με το δέντρο του απλού μετασχηματισμού Haar με την διαφορά ότι κάθε συντελεστής C_i αποτελείται από τρεις επιμέρους συντελεστές $\{c_j, c_{j+1}, c_{j+2}\}$ που δημιουργούν με την σειρά τους ένα δυαδικό δέντρο τριών στοιχείων. Η ρίζα c_j συμπεριφέρεται όπως οι συντελεστές στον κλασσικό μετασχηματισμό κυματιδίων, δηλαδή συνεισφέρει θετικά την τιμή της στο αριστερό υπόδεντρο και αρνητικά στο δεξί. Οι δύο συμπληρωματικοί συντελεστές συνεισφέρουν θετικά την τιμή τους μόνο στο υπόδεντρο στο οποίο ανήκουν, δηλαδή ο αριστερός συντελεστής c_{j+1} στο αριστερό υπόδεντρο και αντίστοιχα ο δεξιός συντελεστής c_{j+2} στο δεξί υπόδεντρο.

Με βάση τον παραπάνω ορισμό μπορούμε να κάνουμε τις εξής σημαντικές παρατηρήσεις:

- Το Haar δέντρο είναι στην ουσία μία υποπερίπτωση του γενικού Haar$^+$, όπου οι συμπληρωματικοί συντελεστές είναι πάντοτε μηδενικοί. Οπότε ο μετασχηματισμός Haar$^+$ είναι οπωσδήποτε πιο ισχυρός και στην χειρότερη περίπτωση περιμένουμε εξίσου καλά αποτελέσματα με τον απλό μετασχηματισμό Haar.

- Παρόλο που ο κάθε συντελεστής αποτελείται από μία τριάδα τιμών, ο χώρος απεικόνισης των συντελεστών δεν χρειάζεται να αυξηθεί σε σχέση με τον απλό μετασχηματισμό. Αυτό συμβαίνει γιατί κάθε τριάδα αντιστοιχεί σε ένα συντελεστή, οπότε κάθε συντελεστής της τριάδας προσδιορίζεται μονοσήμαντα από την θέση του στην τριάδα (ρίζα, αριστερός/δεξιός συντελεστής) και τον δείκτη του συντελεστή που ανήκει. Με αυτόν τον τρόπο καταλήγουμε να χρειαζόμαστε $\lceil \log N \rceil$ bits για την κωδικοποίηση των N διαφορετικών δεικτών συντελεστών.

Βασικές Έννοιες

Στη συνέχεια εισάγουμε τον συμβολισμό και αποδεικνύουμε τις βασικές ιδιότητες του δέντρου Haar$^+$. Έστω ότι έχουμε ένα σύνολο από $N = 2^d$ στοιχεία. Ένα δέντρο Haar$^+$ μπορεί να αναπαραστήσει τα δεδομένα με την βοήθεια ενός διανύσματος μεγέθους $N = 3 \times 2^d - 2$ αφού κάθε συντελεστής εκτός από την ρίζα θα αποτελείται από μία τριάδα. Για τη συνέχεια θα συμβολίζουμε με $a = P(b)$ αν ο κόμβος a είναι ο γονέας του κόμβου b. Επιπλέον αν $a \in Rleaves(b)$, τότε ο κόμβος a ανήκει στο δεξί υποδέντρο του κόμβου b, ενώ αν $a \in path(b)$, τότε ο κόμβος a βρίσκεται στο μονοπάτι που ξεκινά από την ρίζα και καταλήγει στον κόμβο b.

Στη συνέχεια μπορούμε να αναπαρίστανουμε με την βοήθεια των παραπάνω συμβολισμών την δομή του Haar$^+$ δέντρου διακρίνοντας τρεις διαφορετικές περιπτώσεις. Κάθε συντελεστής μπορεί να είναι είτε η ρίζα είτε ο αριστερός ή δεξιός συντελεστής. Μπορούμε να διακρίνουμε

αυτές τις περιπτώσεις ανάλογα με τον δείκτη του συντελεστή λαμβάνοντας υπόψιν ότι κάθε τριάδα δεικτοδοτείται από αριστερά προς τα δεξία (ποστ ορδερ). Συνοπτικά ισχύει ότι:

$$c_i = \begin{cases} P(c_1) & \text{if } i = 0 \\ P(c_{i+1}) \wedge P(c_{i+2}) & \text{if } (i-1) \bmod 3 = 0 \\ P(c_{2i}) & \text{if } (i-2) \bmod 3 = 0 \\ P(c_{2i+1}) & \text{if } i \bmod 3 = 0 \end{cases}$$

Κάθε στοιχείο των αρχικών δεδομένων ανακτάται με βάση το άθροισμα των τιμών των κόμβων που ανήκουν στο μονοπάτι που ξεκινά από τη ρίζα και καταλήγει στο συγκεκριμένο στοιχείο, δηλαδή συμβολικά έχουμε ότι $d_j = \sum_{i \in path(j)} d_{ij} c_i$. Σύμφωνα με τους κανόνες που έχουμε περιγράψει στην προηγούμενη ενότητα, κάποιος συντελεστής συνεισφέρει την τιμή του στο άθροισμα αρνητικά μόνο αν είναι η ρίζα της τριάδας και το φύλλο ανήκει στο δεξί υποδέντρο αυτού. Οπότε για τον συντελεστή d_{ij} ισχύει ότι:

$$d_{ij} = \begin{cases} -1 & \text{if } i - 1 \; mod \; 3 = 0 \wedge (d_j \in Rleaves\,(c_i)) \\ +1 & \text{otherwise} \end{cases}$$

Αντί να διακρίνουμε τους συντελεστές της τριάδας ανάλογα με τον δείκτη, μπορούμε να αναπαριστάνουμε κάθε τριάδα με ένα διάνυσμα τεσσάρων στοιχείων $[u, a, b, c]$, όπου u είναι η εισερχόμενη σε έναν κόμβο τιμή, a η τιμή της ρίζας της τριάδας, b η τιμή του αριστερού συμπληρωματικού συντελεστή και c η τιμή του δεξιού συμπληρωματικού συντελεστή. Με βάση τα παραπάνω κάθε συντελεστής συνεισφέρει $[u + a + b, u - a + c]$ στο αριστερό και δεξί υποδέντρο αντίστοιχα. Επιπλέον συμβολίζουμε με $\|H\|$ τον αριθμό των μη μηδενικών συντελεστών σε ένα Haar$^+$ δέντρο.

Κάθε Haar$^+$ δέντρο περιέχει το πολύ έναν από τους τρεις συντελεστές της κάθε τριάδας και για αυτό αναφερόμαστε στο αραιό Haar$^+$ δέντρο. Παρακάτω παραθέτουμε το θεώρημα που αποδεικνύει την βασική αυτή ιδιότητα του μετασχηματισμού Haar$^+$.

Θεώρημα 2.5.1. *Έστω* $\mathbf{H}$ *ένα Haar*$^+$ *δέντρο που αναπαριστά ένα σύνολο δεδομένων* $\mathbf{D}$ *και το οποίο περιέχει περισσότερους από έναν συντελεστή σε τουλάχιστον μία τριάδα. Τότε το διάνυσμα* $\mathbf{D}$ *μπορεί να παρασταθεί ισοδύναμα από τουλάχιστον ένα αραιό Haar*$^+$ *δέντρο* $\mathbf{H'}$*, του οποίου οι τριάδες περιέχουν το πολύ έναν μη μηδενικό συντελεστή και τέτοιο ώστε* $\|\mathbf{H'}\| \leq \|\mathbf{H}\|$.

Απόδειξη. Έστω μία τριάδα $C = [u, a, b, c]$ του δέντρου $\mathbf{H}$ η οποία περιέχει περισσότερους από έναν μη μηδενικούς συντελεστές. Η τριάδα αυτή μπορεί να αναπαρασταθεί ισοδύναμα με την τριάδα $C' = [u, 0, b + a, c - a]$. Προφανώς και στις δύο περιπτώσεις η έξοδος θα είναι $[u + a + b, u - a + c]$ για το αριστερό και δεξί υποδέντρο αντίστοιχα. Αν θέσουμε $p = b + a$ και$q = c - a$ η τριάδα C' γίνεται $C' = [u, 0, p, q]$ και αντίστοιχα η έξοδος είναι $[u + p, u - q]$. Παρατηρούμε ότι και αυτή η τριάδα μπορεί να μειωθεί αν αντικατασταθεί από την τριάδα $C'' = \left[u - \frac{p+q}{2}, \frac{p-q}{2}, 0, 0\right]$. Προφανώς η έξοδος και σε αυτήν την περίπτωση παραμένει ίδια και ισούται με $[u + p, u - q]$. Για να δημιουργηθεί όμως αυτή η τριάδα θα πρέπει

να αλλάξει η εισερχόμενη τιμή. Διακρίνουμε λοιπόν τις εξής δύο περιπτώσεις για την αλλαγή της εισερχόμενης τιμής:

- Αν ο γονέας της τριάδας C είναι η ρίζα του δέντρου τότε απλά προσθέτουμε $\frac{p+q}{2}$ στον κόμβο αυτό. Προφανώς αφού η έξοδος είναι ίδια το δέντρο που προκύπτει είναι ισοδύναμο με το αρχικό.

- Αν ο γονέας της τριάδας C δεν είναι η ρίζα του δέντρου τότε προσθέτουμε $\frac{p+q}{2}$ στον συμπληρωματικό συντελεστή που δείχνει προς τη μεριά της τριάδας C. Σε περίπτωση που προκύπτει τριάδα-γονέας με περισσότερους του ενός μη μηδενικούς συντελεστές επαναλαμβάνουμε την διαδικασία από την αρχή θεωρώντας ως τριάδα C, την τριάδα-γονέα. Η διαδικασία αυτή μπορεί να επαναλαμβάνεται μέχρι να φτάσουμε στην ρίζα του δέντρου, οπότε ισχύει η πρώτη περίπτωση.

Παραπάνω αποδείξαμε ότι κάθε τριάδα C του δέντρου $\mathbf{H}$ που περιέχει παραπάνω από έναν μη μηδενικό συντελεστή μπορεί να αναπαρασταθεί ισοδύναμα με μία τριάδα C'' που περιέχει έναν μόνο μη μηδενικό συντελεστή χωρίς να δημιουργεί σε καμία άλλη τριάδα του δέντρου επιπλέον μη μηδενικούς συντελεστές. Με αυτόν τον τρόπο ο αριθμός των μη μηδενικών συντελεστών μειώνεται ή στην χειρότερη περίπτωση παραμένει ίδιος. Η χειρότερη περίπτωση είναι όταν η τριάδα γονέας έχει μόνο μηδενικούς συντελεστές, οπότε κάθε τριάδα που περιείχε δύο μη μηδενικούς συντελεστές αντικαθίσταται από δύο τριάδες που περιέχουν από έναν μη μηδενικό συντελεστή, δηλαδή ο συνολικός αριθμός των μη μηδενικών συντελεστών παραμένει ο ίδιος αν και αλλάζει η σύνθεση των τριάδων. Αφού εφαρμόσουμε λοιπόν την παραπάνω μέθοδο σε όλες τις τριάδες του δέντρου $\mathbf{H}$ που περιέχουν πλεονάζοντες συντελεστές, προκύπτει ένα δέντρο $\mathbf{H'}$ που έχει το πολύ έναν μη μηδενικό συντελεστή σε κάθε τριάδα και έχει το πολύ ίδιο αριθμό μη μηδενικών συντελεστών, δηλαδή $\|\mathbf{H'}\| \leq \|\mathbf{H}\|$.

2.5.2 Σύνοψη Haar$^+$ για κατανεμημένες μετρικές σφάλματος

2.5.3 Μοντελοποιήση της λύσης

Σε αυτήν την ενότητα θα εισάγουμε τις βασικές δομές και το θεωρητικό υπόβαθρο που θα βοηθήσει στην κατανόηση της λύσης. Σκοπός του προβλήματος σε μαθηματική μοντελοποίηση είναι να γεμίσουμε έναν πίνακα $Q(i, u, b)$ που θα αποθηκεύει την βέλτιστη επιλογή συντελεστή στην τριάδα C_i δεδομένου ότι η εισερχόμενη τιμή είναι u και ότι ο αριθμός των συντελεστών που θα κρατηθούν σε όλο το υποδέντρο που ξεκινά από την τριάδα C_i είναι b. Σε αυτό το σημείο μπορεί να γίνει μία βελτιστοποίηση της λύσης αν λάβουμε υπόψη μας ότι το επίπεδο που βρίσκεται η τριάδα καθορίζει τον αριθμό των συντελεστών-τριάδων που υπάρχουν στο υποδέντρο που ορίζει ως κορυφή. Δεδομένου ότι σε κάθε τριάδα όπως έχει αποδειχθεί αρκεί να κρατηθεί το πολύ ένας συντελεστής, το μέγιστο πλήθος των συντελεστών που μπορούν να χρησιμοποιηθούν στην περίληψη ισούται με τον αριθμό των τριάδων στο υποδέντρο. Οπότε δεν έχει νόημα να εξετάζεται η κατασκευή της περίληψης για b μεγαλύτερο από αυτόν τον αριθμό.Συμπερασματικά, θα έχουμε ότι $b \in \{0, 1, \ldots, \min\{B, 2^{l_i} - 1\}\}$, όπου l_i το επίπεδο που ανήκει η τριάδα με αναφορά το χαμηλότερο επίπεδο του δέντρου. Επιπλέον προς

υπολογισμό των τιμών του πίνακα θα πρέπει να ορίσουμε κάποιο επιτρεπόμενο πεδίο τιμών για την εισερχόμενη τιμή. Εφόσον έχουμε καθορίσει αυτό το πεδίο τιμών μπορούμε στην συνέχεια να το διαμερίσουμε με βάση κάποιο βήμα d και να υπολογίσουμε τις τιμές του πίνακα για όλο το εύρος τιμών της εισερχόμενης τιμής. Παρακάτω παραθέτουμε κάποιες προτάσεις που θα βοηθήσουν στον ορισμό των του εύρους της εισερχόμενης τιμής.

Πρόταση 2.5.1. *Για εισερχόμενη τιμή u στην τριάδα C_i υπάρχουν ανακατασκευασμένες τιμές $\hat{d}_k$ και $\hat{d}_l$ τέτοιες ώστε $\hat{d}_k \leq u$ και $\hat{d}_l \geq u$.*

Πρόταση 2.5.2. *Αν η τριάδα C_i έχει ένα μη μηδενικό συντελεστή ρίζα z_h, τότε η εισερχόμενη τιμή u στην τριάδα C_i βρίσκεται στο διάστημα (m_i, M_i). Συμβολικά, $z_h \neq 0 \Rightarrow u \in (m_i, M_i)$. Ή αντίστροφα, $u \notin (m_i, M_i) \Rightarrow z_h = 0$.*

Απόδειξη. Έστω ότι $u \notin (m_i, M_i)$ και $z_h \neq 0$. Θα αποδείξουμε ότι •κάτι τέτοιο δεν οδηγεί σε βέλτιστη περίληψη. Θεωρούμε χωρίς βλάβη της γενικότητας ότι $u > M_i$). Αφού $z_h \neq 0$ θα αφαιρείται η τιμή του συντελεστή-ρίζα από την εισερχόμενη τιμή για το δεξί υποδέντρο και αντίστοιχα θα προστίθεται για το αριστερό. Με αυτόν τον τρόπο όμως θα προκύψει στο αριστερό υποδέντρο τιμή που ξεπερνά την μέγιστη τιμή που στοχεύουμε να ανακατασκευάσουμε κάτι το οποίο προφανώς δεν είναι χρήσιμο. Εναλλακτικά θα μπορούσαμε να κρατήσουμε τον αριστερό συμπληρωματικό συντελεστή z_l με τιμή $-z_h$ οπότε θα έχουμε την ίδια έξοδο προς τα αριστερά χωρίς να υπερβαίνουμε την μέγιστη τιμή στο δεξί υποδέντρο. Αυτή η λύση είναι προφανώς καλύτερη αφού εισάγει σε κάθε περίπτωση λιγότερο σφάλμα στην περίληψη. Οπότε καταλήγουμε ότι το z_h να είναι μηδενικό. Ομοίως αποδεικνύεται ότι και στην περίπτωση που η εισερχόμενη τιμή u είναι μικρότερη του m_i θα πρέπει $z_h = 0$.

Πρόταση 2.5.3. *Μία εισερχόμενη τιμή u στην τριάδα δεν μπορεί να οδηγήσει σε καλύτερη προσέγγιση των δεδομένων από μία τιμή u', τέτοια ώστε $u < u' < m_i (u > u' > M_i)$ για κάθε περίληψη με τον ίδιο αριθμό μη μηδενικών συντελεστών στο υποδέντρο που ορίζεται με ρίζα την τριάδα C_i.*

Απόδειξη. Έστω $u < m_i$. Σύμφωνα με την Πρόταση 2 ο πρώτος μη μηδενικός συντελεστής που θα υπάρχει στο υποδέντρο θα είναι ένας συμπληρωματικός συντελεστής. Αν αλλάξουμε την τιμή αυτού του συντελεστή κατάλληλα ώστε να καλύπτει την διαφορά που εισέρχεται από μία εισερχόμενη τιμη u' τέτοια ώστε $u < u' < m_i$, θα προκύψει η ίδια έξοδος. Όμως στις τιμές (αν υπάρχουν αυτές) που παίρνουν αυτούσια την εισερχόμενη τιμή ως την ανακατασκέυασμένη τιμή, λόγω του ότι όλοι οι ενδιάμεσοι συντελεστές είναι μηδενικοί, θα πετύχουμε μικρότερο σφάλμα αν επιλέξουμε την εισερχόμενη τιμή u'. Έτσι προκύπτει ότι μια τιμή u τέτοια ώστε $u < u' < m_i$ πετυχαίνει χειρότερη ή το πολύ ίδια ακρίβεια στα ανακατασκευασμένα δεδομένα με μία εισερχόμενη τιμή u'.

Συνεχίζουμε αποδεικνύοντας τα θεωρήματα που θέτουν το εύρος τιμών για όλες τις παραμέτρους του προβλήματος.

Θεώρημα 2.5.2. *Έστω m_i, M_i το ελάχιστο και μέγιστο αντίστοιχα του υποδέντρου που έχει ως ρίζα την τριάδα C_i και ότι η εισερχόμενη τιμή u στην C_i ανήκει στο διάστημα (m_i, M_i). Αν θέσουμε μη μηδενική στο z_h τότε αυτή θα βρίσκεται στο διάστημα $\max\{M_i - u, u - m_i\}$.*

Απόδειξη. Εφόσον $z_h \neq 0$ η τιμή που θα προωθηθεί στο αριστερό και δεξί υποδέντρο θα είναι αντίστοιχα $u + z_h, u - z_h$. Χωρίς βλάβη της γενικότητας θεωρούμε ότι $z_h > 0$ και ότι οι εξερχόμενες τιμές υπερβαίνουν τα όρια που τίθενται από την έλαχιστη και μέγιστη τιμή του υποδέντρου, δηλαδή $u + z_h > M_i, u - z_h < m_i$. Αντίστοιχα αν θεωρούσαμε $z_h < 0$, θα υποθέταμε $u - z_h > M_i, u + z_h < m_i$. Σύμφωνα με την Πρόταση 3 μπορούμε να πετύχουμε μικρότερο σφάλμα αν διαλέξουμε εισερχόμενη τιμή τέτοια ώστε τουλάχιστον μία από τις εξόδους να βρίσκεται στο διάστημα $[m_i, M_i]$. Από την διαπίστωση αυτή προκύπτουν δύο ανισότητες από τις οποίες μπορούμε να πάρουμε τα όρια για το z_h. Δηλαδή έχουμε ότι:

$$\begin{gathered} m_i \leq u + z_h \leq M_i \vee m_i \leq u - z_h \leq M_i \Leftrightarrow \\ m_i - u \leq z_h \leq M_i - u \vee u - M_i \leq z_h \leq u - m_i \Leftrightarrow \\ z_h \in [\min\{u - M_i, m_i - u\}, \max\{M_i - u, u - m_i\}] \Leftrightarrow \\ z_h < \max\{M_i - u, u - m_i\} \end{gathered}$$

Ομοίως μπορούμε πλέον εύκολα με βάση την Πρόταση 3 όπως παραπάνω να θέσουμε το εύρος τιμών για τους συμπληρωματικούς συντελεστές και την ρίζα του δέντρου. Το παρακάτω θεώρημα καθορίζει τα όρια αυτά.

Θεώρημα 2.5.3. *Έστω $m_l, M_l (m_r, M_r)$ το ελάχιστο και μέγιστο αντίστοιχα του υποδέντρου που έχει ως ρίζα τον αριστερό z_l(δεξιό z_r) συμπληρωματικό συντελεστή της τριάδας C_i. Αν θέσουμε μη μηδενική τιμή στο αριστερό z_l(δεξιό z_r) συμπληρωματικό συντελεστή, τότε θα πρέπει να ισχύει $z_l \in [m_l - u, M_l - u] (z_r \in [m_r - u, M_r - u])$. Επιπλέον αν η ρίζα του δέντρου έχει μη μηδενική τιμή, τότε αυτή θα πρέπει να βρίσκεται στο διάστημα $[m, M]$, όπου m, M το ελάχιστο και μέγιστο αντίστοιχα.*

Στη συνέχεια θα εκφράσουμε τις εισερχόμενες τιμές σε κάθε κόμβο με βάση το ολικό ελάχιστο και μέγιστο.

Θεώρημα 2.5.4. *Η εισερχόμενη τιμή στην τριάδα C_i βρίσκεται στο διάστημα $[m - D, M + D]$ όπου $D = M - m$.*

Απόδειξη. Από το Θεώρημα 3 έχουμε ότι η ρίζα αν έχει μη μηδενική τιμή θα βρίσκεται στο διάστημα $[m, M]$. Επιπλέον η τιμή των συμπληρωματικών συντελεστών, εφόσον κάποιος από αυτούς είναι μη μηδενικός, θα βρίσκεται στο διάστημα $[-D, D]$. Το κάτω όριο προκύπτει αν θεωρήσουμε ότι η ρίζα παίρνει την ανώτατη τιμή M οπότε ο συμπληρωματικός συντελεστής θα βρίσκεται στο διάστημα $[-D, 0]$ και αντίστοιχα το άνω όριο αν θεωρήσουμε ότι η ρίζα παίρνει την κατώτατη τιμή M οπότε ο συμπληρωματικός συντελεστής θα βρίσκεται στο διάστημα $[0, D]$, δηλαδή συνολικά για κάθε περίπτωση m. Όπως είναι φανερό η τιμή της εισερχόμενης τιμής για τον απόγονο-συντελεστή έστω C_i θα βρίσκεται στο διάστημα $[m - D, M + D]$. Για την περίπτωση που κάποια τριάδα έχει μη μηδενικό συντελεστή ρίζα z_h με βάση την Πρόταση 2

η εισερχόμενη τιμή u θα βρίσκεται στο διάστημα $[m, M]$. Επιπλέον σύμφωνα με το Θεώρημα 2 θα ισχύει ότι $z_h < \max\{M_i - u, u - m_i\}$. Οπότε συνδυάζοντας τα παραπάνω η εισερχόμενη τιμή στο αριστερό και δεξί υποδέντρο θα είναι αντίστοιχα $u \pm z_h = [2m - M, 2M - m]$ ή διαφορετικά $[m - D, M + D]$. Οπότε καταλήξαμε ότι σε κάθε περίπτωση η εισερχόμενη τιμή για όλες τις τριάδες θα βρίσκεται στο διάστημα $[m - D, M + D]$.

Με βάση το παραπάνω Θεώρημα μπορούμε να υπολογίσουμε το μέγεθος του συνόλου που περιέχει όλες τις δυνατές τιμές των εισερχόμενων τιμών και των τιμών των συντελεστών κάθε τριάδας. Το εύρος των εισερχόμενων τιμών όπως προκύπτει από το Θεώρημα 4 είναι $3D$. Αν θεωρήσουμε ένα βήμα d για την εύρεση των πιθανών τιμών που βρίσκονται σε αυτό το διάστημα προκύπτει ένα σύνολο S μέγιστου μήκους $\left\lfloor \frac{3D}{d} \right\rfloor + 1 = O\left(\frac{D}{d}\right)$. Επιπλέον αν υποθέσουμε ότι τα $S^u_{i,H} \subset \Re, S^u_{i,L} \subset \Re, S^u_{i,R} \subset \Re$ συμβολίζουν τα σύνολα από όλες τις δυνατές τιμές που παίρνουν οι τρεις συντελεστές z_h, z_l, z_r μιας τριάδας C_i δεδομένου κάποιας εισερχόμενης τιμής u, σύμφωνα με τα Θεωρήματα 2 και 3 και την παραπάνω παρατήρηση, συμπεραίνουμε ότι και η χωρική πολυπλοκότητα αυτών των συνόλων θα είναι επίσης $O\left(\frac{D}{d}\right)$.

2.5.4 Εξαγωγή της λύσης

Στόχος όπως δείξαμε και παραπάνω είναι να δημιουργήσουμε τον πίνακα $Q(i, u, b)$. Θα χρειαστεί ένας αλγόριθμος δυναμικού προγ4ραμματισμού που θα εξετάζει όλες τις δυνατές συνόψεις αναδρομικά, θα υπολογίζει το σφάλμα για κάθε περίπτωση και θα επιλέγει την βέλτιστη λύση. Θεωρούμε ότι ο πίνακας $A(u, b)$ αποθηκεύει τις βέλτιστες λύσεις για κάθε επιτρεπόμενη εισερχόμενη τιμή u και για υπολειπόμενο χώρο περίληψης b_l. Κάθε στοιχείο του πίνακα θα έχει τα εξής πεδία:

1. z_h, z_l, z_r οι τιμές που παίρνει κάθε συντελεστής της τριάδας.
2. $E(i, u, b)$ το μικρότερο σφάλμα που μπορεί να επιτευχθεί με βάση την συγκεκριμένη εισερχόμενη τιμή και τον περιορισμό χώρου b.
3. b_l τον χώρο που αποδίδεται σύμφωνα με την βέλτιστη λύση στο αριστερό υποδέντρο.

Το σφάλμα υπολογίζεται όπως αναφέραμε και προηγουμένως αναδρομικά με βάση τα ελάχιστα σφάλματα των παιδιών του κόμβου C_i, μέχρι να φτάσουμε στην βάση της αναδρομής, οπότε υπολογίζουμε εύκολα το σφάλμα από την εισερχόμενη και την αρχική τιμή. Παρακάτω φαίνεται συμβολικά ο υπολογισμός του σφάλματος:

$$E(0,0,B) = \min_{z \in S^0_{0,H}} \{E(1, z, B - (z \neq 0))\}$$

$$E(i,u,B) = \min \begin{cases} \min_{z_h \in S^u_{i,H}, b' \in D_i} \{E(i_l, u + z_h, b') + E(i_r, u - z_h, b - b' - \{z_h \neq 0\}) \\ \min_{z_l \in S^u_{i,L}, b' \in D_i} \{E(i_l, u + z_l, b') + E(i_r, u, b - b' - \{z_l \neq 0\}) \\ \min_{z_r \in S^u_{i,R}, b' \in D_i} \{E(i_l, u, b') + E(i_r, u + z_r, b - b' - \{z_r \neq 0\}) \end{cases}$$

Όπως βλέπουμε χρησιμοποιήθηκαν όλες οι ιδιότητες που αποδείξαμε παραπάνω. Κάθε φορά υπολογίζουμε το σφάλμα δεδομένου ότι ένας μόνο από τους συντελεστές είναι μη μηδενικός και τέλος συνδυάζουμε τα αποτελέσματα για να βρούμε την βέλτιστη λύση. Στη βάση της

αναδρομής υπολογίζουμε το σφάλμα σύμφωνα με την διαφορά της εισερχόμενης τιμής από την επιθυμητή.

2.5.5 Ανάλυση Πολυπλοκότητας Εύρεσης Πίνακα Σφαλμάτων

Για κάθε αναδρομή εξετάζουμε όλες τις δυνατές εισερχόμενες τιμές και με βάση αυτές όλες τις πιθανές τιμές των τριών συντελεστών. Όπως αποδείχθηκε παραπάνω καθένα από αυτά τα σύνολα έχει χωρική πολυπλοκότητα $O\left((\frac{D}{d})^2\right)$ για δεδομένο b. Οπότε η χρονική πολυπλοκότητα θα ισούται με το μήκος αυτών των συνόλων ανά δύο(για δεδομένο u υπολογίζουμε τις δυνατές τιμές των συνόλων $S^u_{i,H}, S^u_{i,L}, S^u_{i,R}$ για όλες τις δυνατές τιμές του b,δηλαδή για $b \in \min\left\{B, 2^{l_i}-1\right\}$. Οπότε συνολικά η χρονική πολυπλοκότητα θα ισούται με $O\left((\frac{D}{d})^2 \sum_{i=1}^{n} \min\left\{B, 2^{l_i}-1\right\}^2\right) = O\left(\frac{D}{d} nB\right)$. Επιπλέον στην ειδική περίπτωση που η μετρική σφάλματος αφορά το μέγιστο από τα επιμέρους σφάλματα και όχι το άθροισμα αυτών ο παράγοντας B γίνεται $\log^2 B$ λόγω της δυαδικής αναζήτησης στο σύνολο των δυνατών τιμών. Σε ό,τι αφορά την χωρική πολυπλοκότητα, δεδομένου ότι το μέγιστο πλήθος των πινάκων που πρέπει να είναι αποθηκευμένοι κάθε στιγμή ισούται με $\log n + 1$(δηλαδή ένας πίνακας το πολύ ανά επίπεδο), η χωρική πολυπλοκότητα θα είναι $O\left(\frac{D}{d} \sum_i = 1^{\log n+1} \min\left\{B, 2^{l_i}-1\right\}\right) = O\left(\frac{D}{d} \log \frac{n}{B}\right)$.

2.5.6 Δημιουργία Περίληψης

Η κατασκευή της περίληψης μπορεί να γίνει με δύο διαφορετικές προσεγγίσεις. Η μία είναι περισσότερο αποδοτική ως προς τον χώρο αλλά υστερεί σε πολυπλοκότητα χρόνου και η άλλη αντίστροφα είναι αποδοτική ως προς τον χρόνο και μη αποδοτική σε χώρο.

Space-Efficient Λύση

Σε αυτή την περίπτωση αφού λύσουμε το πρόβλημα και εξάγουμε τον πίνακα E για την ρίζα, βρίσκουμε την εγγραφή του πίνακα που ελαχιστοποιεί το ολικό σφάλμα. Στη συνέχεια ξανατρέχουμε τον αλγόριθμο θεωρώντας ως ρίζα τα παιδιά του. Με αυτόν τον τρόπο γνωρίζοντας την εισερχόμενη τιμή (από την επιλογή της εγγραφής με το ελάχιστο σφάλμα) και το μέγεθος της περίληψης του υποδέντρου (από το πεδίο της εγγραφής που αποθηκεύει το πλήθος των συντελεστών που κατανέμονται στο αριστερό υποδέντρο), βρίσκουμε τους συντελεστές του κόμβου και επαναλαμβάνουμε την ίδια διαδικασία για τα παιδιά του μέχρι να φτάσουμε στα φύλλα οπότε έχει προσδιορισθεί το δέντρο πλήρως. Η χρονική πολυπλοκότητα θα ισούται με το άθροισμα του χρόνου όλων των κλήσεων της μεθοόδου που υπολογίζει τους πίνακες E. Όπως αναφέρθηκε παραπάνω η χρονική πολυπλοκότητα για δεδομένα μεγέθους n και περίληψη μεγέθους B είναι $O\left(\frac{D}{d} nB\right)$. Άρα επειδή σε κάθε κλήση το μέγεθος των δεδομένων θα ισούται με το πλήθος των δεδομένων που ανήκουν στο υποδέντρο του κόμβου που λαμβάνεται ως ρίζα, η συνολική χρονική πολυπλοκότητα θα είναι $O\left(\frac{D}{d} B \sum_{l=0}^{\log n} 2^{\log n - l}\right) = O\left(\frac{D}{d} nB \log n\right)$ όπου l το επίπεδο του κάθε συντελεστή κόμβου με αναφορά τη ρίζα, το οποίο παίρνει τιμές προφανώς στο διάστημα $[0, \log n]$. Η χωρική πολυπλοκότητα θα ισούται με τον χώρο που χρειάζεται η κλήση για την ρίζα του

δέντρου(οπότε απαιτείται μέγιστη δέσμευση χώρου) $O\left(\frac{D}{d}\log\frac{n}{B}\right)$ και το χώρο n που απαιτείται για την αποθήκευση των δεδομένων, δηλαδή συνολικά $O\left(\frac{D}{d}\log\frac{n}{B}+n\right)$.

Time-Efficient Λύση

Η αποδοτική ως προς τον χρόνο λύση θα έχει χρονική πολυπλοκότητα $O\left(\frac{D}{d}nB\right)$, δηλαδή ίση με μία κλήση της μεθόδου για την ρίζα του δέντρου. Στη συνέχεια υπάρχουν οι εξής δύο περιπτώσεις για την κατασκευή του δέντρου:

- Να κρατήσουμε όλους τους πίνακες των εσωτερικών κόμβων του δέντρου οπότε ύστερα θα βρίσκουμε τους συντελεστές με απλή πρόσβαση στους πίνακες χωρίς να χρειάζεται να ξανατρέχουμε τον αλγόριθμο. Σε κάθε επίπεδο ο χώρος που απαιτείται για την αποθήκευση του πίνακα θα είναι $O\left(\left(\frac{D}{d}\right)\min\left\{B,2^l\right\}\right)$. Για να έχουμε συνολική εκτίμηση του χώρου θα πρέπει να αθροίσουμε τον χώρο που απαιτείται σε κάθε επίπεδο του δέντρου. Έτσι σε επίπεδο l δημιουργούνται 2^{l-1} πίνακες οπότε συνολικά θα ισχύει ότι $O\left(\left(\frac{D}{d}\right)\sum_{i=l-1}^{\log n}\min\left\{B,2^i\right\}\right)=O\left(\frac{D}{d}n\log B\right)$.

- Να αποθηκεύουμε σε κάθε εγγραφή του πίνακα μαζί με το ελάχιστο σφάλμα και την τιμή και τον δείκτη όλων των μη μηδενικών κόμβων που θα πρέπει να υπάρχουν στο υποδέντρο ώστε να επιτευχθεί η βέλτιστη αυτή προσέγγιση. Σε αυτήν την περίπτωση και πάλι αρκεί να έχουμε το πολύ $\log n+1$ ανοιχτούς πίνακες με την διαφορά ότι κάθε πίνακας, επειδή θα περιέχει και ένα διάνυσμα με τους μη μηδενικούς συντελεστές που έχουν επιλεχθεί στα κατώτερα επίπεδα, θα έχει χωρική πολυπλοκότητα $O\left(\left(\frac{D}{d}\right)\min\left\{B,2^l\right\}^2\right)$. Οπότε συνολικά θα έχουμε ότι $O\left(\frac{D}{d}\sum_{i=1}^{\log n+1}\min\left\{B,2^{l_i}-1\right\}^2\right)=O\left(\frac{D}{d}B^2\log\frac{n}{B}\right)$.

2.6 Ενιαίος Συμβολισμός

Στις προηγούμενες παραγράφους αναλύσαμε τους τρεις τύπους του μετασχηματισμού Haar. Σε αυτήν την ενότητα επιδιώκουμε να τονίσουμε τα ενιαία χαρακτηριστικά των τριών αυτών μετασχηματισμών και να εισάγουμε ένα ενιαίο συμβολισμό που θα περιλαμβάνει σε αφαιρετικό επίπεδο και τις τρεις αυτές περιπτώσεις.

Τα δεδομένα σε κάθε μετασχηματισμό είναι τα φύλλα του αντίστοιχου δέντρου που δημιουργείται ενώ η προσεγγιστική τιμή του κάθε στοιχείου είναι το άθροισμα όλων των συντελεστών που υπάρχουν στο μονοπάτι από τη ρίζα του δέντρου μέχρι το φύλλο στο οποίο βρίσκεται το στοιχείο. Συμβολικά για το στοιχείο έστω d_j των δεδομένων η ανακατασκευασμένη τιμή θα είναι$d_j=\sum_{i\in path(j)}d_{ij}c_i$. Το d_{ij} συμβολίζει το πρόσημο της συνεισφοράς της τιμής του κάθε συντελεστή στο άθροισμα. Η παράμετρος αυτή είναι $+1$ για τα στοιχεία που βρίσκονται στο αριστερό υποδέντρο και -1 για τα στοιχεία που βρίσκονται στο δεξί για τον απλό M/Σ Haar και τον M/Σ unrestricted Haar. Αντίστοιχα για τον M/Σ Haar$^+$ η παράμετρος αυτή παίρνει τις τιμές που περιγράφονται στην παράγραφο 2.4.2.

Επιπλέον όταν τρέχουμε έναν αλγόριθμο από κάτω προς τα πάνω θα χρειάζεται να γνωρίζουμε το αρχικό μονοπάτι S που καταλήγει σε αυτό τον κόμβο. Το μονοπάτι αυτό

θα εκφράζεται ως το σύνολο των μη μηδενικών συντελεστών που έχουν επιλεχθεί γα τον M/Σ Haar και ως μία εισερχόμενη τιμή u για τον M/Σ Haar$^+$ και τον M/Σ unrestricted Haar.

Κεφάλαιο 3

Συνόψεις για Συναθροιστικές Ερωτήσεις

3.1 Διατύπωση του προβλήματος

Σε αυτή την παράγραφο θα διατυπώσουμε με μαθηματικό συμβολισμό το γενικό πρόβλημα που περιλαμβάνει ως υποπεριπτώσεις όλα τα είδη μετασχηματισμών κυματιδίων. Κάθε μετασχηματισμός κυματιδίων στηρίζεται στην εξής γενική μεθοδολογία:

- Θεωρώντας ότι έχουμε ένα σύνολο από N στοιχεία. Ορίζουμε τον αριθμό των στοιχείων του κατώτατου επιπέδου του δέντρου που προκύπτει από τον μετασχηματισμό ίσο με N. Τα στοιχεία αυτά αναπαριστούν τα ανεκτημένα στοιχεία με βάση τον μετασχηματισμό.
- Αναπτύσσουμε το δέντρο προς τα επάνω σύμφωνα με τους κανόνες του μετασχηματισμού και υπολογίζουμε για κάθε δυνατή σύνοψη τη μετρική σφάλματος που θέλουμε να βελτιστοποιήσουμε.

Σε αυτό σημείο θα πρέπει να εξηγήσουμε επιπλέον τον τρόπο με τον οποίο υπολογίζεται το σφάλμα. Σε κάθε περίπτωση σκοπός είναι να επιλέξουμε ένα μικρό αριθμό από συντελεστές, έστω B από τους συνολικούς N (ο αριθμός των εσωτερικών κόμβων του δέντρου) συντελεστές του δέντρου του μετασχηματισμού. Οι συντελεστές c_i οι οποίοι δεν έχουν επιλεχθεί, θεωρούνται μηδενικοί κατά την ανάκτηση των δεδομένων. Κάθε στοιχείο των αρχικών δεδομένων ανακτάται με βάση το άθροισμα των τιμών των κόμβων που ανήκουν στο μονοπάτι που ξεκινά από τη ρίζα και καταλήγει στο συγκεκριμένο στοιχείο, δηλαδή συμβολικά έχουμε ότι $d_j = \sum_{i \in path(j)} d_{ij} c_i$, όπου d_{ij} σταθερά που ορίζεται από το είδος του μετασχηματισμού. Η απόλυτη διαφορά της πραγματικής τιμής από την ανεκτημένη, ορίζει το σημειακό απόλυτο σφάλμα.

Αν συμβολίσουμε με $s_{ij} = \sum_{k=i}^{j} d_k$ το άθροισμα των τιμών των δεδομένων που έχουν δείκτη i μέχρι και j, παίρνουμε την τιμή της αντίστοιχης συναθροιστικής ερώτησης. Η απόλυτη διαφορά της πραγματικής τιμής της συναθροιστικής αυτής ερώτησης (που προκύπτει από τα πραγματικά δεδομένα) και της ανακατασκευασμένης (που προκύπτει από τα φύλλα του δέντρου) ορίζεται ως το συναθροιστικό απόλυτο σφάλμα.

Πλέον σύμφωνα με τα παραπάνω μπορούμε να διατυπώσουμε το πρόβλημα προς επίλυση ως εξής:

Δεδομένου ενός συνόλου N στοιχείων (όπου N δύναμη του 2) , αναζητούμε την περίληψη με B από τους συνολικούς N συντελεστές, η οποία ελαχιστοποιεί το μέγιστο απόλυτο σφάλμα όλων των δυνατών συναθροιστικών ερωτήσεων που προκύπτουν από τα δεδομένα.

3.2 Θεωρητικό Υπόβαθρο

Σε αυτή την ενότητα θα εισάγουμε κάποια βασικά θεωρήματα που θα βοηθήσουν στη επίλυση του προβλήματος. Αρχικά θα πρέπει να ορίσουμε τον απαραίτητο συμβολισμό. Θεωρούμε το σύνολο $e_1, e_2, \ldots, e_n$ με τα προσημασμένα σημειακά σφάλματα που αντιστοιχούν στα $d_1, d_2, \ldots, d_n$ στοιχεία του συνόλου των δεδομένων. Συμβολίζουμε με $s(k,l) = \sum_{i=k}^{l} e_i$ το σφάλμα της συναθροιστικής ερώτησης που περιέχει τα στοιχεία με δείκτη από k μέχρι l, το οποίο προκύπτει από το άθροισμα των $k-l$ διαδοχικών σφαλμάτων που οι δείκτες τους ξεκινούν από k και καταλήγουν σε l. Επιπλέον αν $d(k,l) = d_k, \ldots, d_m, \ldots, d_l$ ένα υποδιάστημα των δεδομένων, θα συμβολίζουμε με $M(k,l) = \max_{k \le i \le m \le j \le l)} s(i,j)$ το μέγιστο σφάλμα όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχουν στοιχεία και από τα δύο διαδοχικά υποδιαστήματα $d(k,m), d(m,l)$. Αντίστοιχα ορίζεται το $m(k,l)$ ως το ελάχιστο σφάλμα όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχουν στοιχεία και από τα δύο διαδοχικά υποδιαστήματα $d(k,m), d(m,l)$. Τέλος με τους όρους $R_M(k,l) = \max_{k \le i \le l} s(i,l)$ και $L_M(k,l) = \max_{(k \le j \le l)} s(k,j)$ συμβολίζουμε το μέγιστο όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχονται στο υποδιάστημα $d(k,l)$ και καταλήγουν στο δεξιό άκρο ή ξεκινούν από το αριστερό άκρο αντίστοιχα. Ομοίως ορίζονται οι αντίστοιχοι συμβολισμοί για τα ελάχιστα. Αναλυτικά οι παραπάνω συμβολισμοί φαίνονται στον Πίνακα 3.1.

Το πρώτο λήμμα αναφέρεται στην εύρεση του μεγίστου σφάλματος ενός διαστήματος των δεδομένων, έχοντας κάποιες πληροφορίες για δύο επιμέρους τμήματά του. Τονίζουμε ότι η παρακάτω απόδειξη αναφέρεται στο μέγιστο προσημασμένο σφάλμα και όχι κατά απόλυτη τιμή. Το μέγιστο απόλυτο σφάλμα προκύπτει από την σύγκριση ελάχιστου και μέγιστου σφάλματος, όπως θα δείξουμε παρακάτω.

Λήμμα 3.2.1. *Έστω $d(k,l) = d_k, \ldots, d_j, \ldots, d_l$ ένα υποδιάστημα των δεδομένων που χωρίζεται με βάση το στοιχείο d_j σε δύο διαδοχικά υποδιαστήματα, μπορούμε να υπολογίσουμε το μέγιστο σφάλμα $M(k,j,l)$ όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχουν δεδομένα και από τα δύο υποδιαστήματα, αν γνωρίζουμε τα μεγέθη $R_M(k,j)$,$L_M(j+1,l)$.*

Απόδειξη. Θέλουμε να υπολογίσουμε το μέγιστο σφάλμα όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχουν δεδομένα και από τα δύο υποδιαστήματα $d(k,j)$ και $d(j+1,l)$. Το σφάλμα αυτό προκύπτει, όπως φαίνεται από τον ορισμό του, από το άθροισμα των επιμέρους προσημασμένων σημειακών σφαλμάτων.Επειδή το άθροισμα αφορά μόνο διαδοχικούς αριθμούς θα πρέπει να συνδυάσουμε αθροίσματα που τελειώνουν στο στοιχείο d_j με άλλα που αρχίζουν

Σύμβολα Συναθροιστικών Ερωτήσεων	
$I(k,l)$	Μέγιστο απόλυτο συναθροιστικό σφάλμα που βρίσκεται εντός των ορίων k,l
$L_M(k,l)$	Μέγιστο προσημασμένο συναθροιστικό σφάλμα που ξεκινά από το στοιχείο k και τελειώνει πριν το l
$L_m(k,l)$	Ελάχιστο προσημασμένο συναθροιστικό σφάλμα που ξεκινά από το στοιχείο k και τελειώνει πριν το l
$R_M(k,l)$	Μέγιστο προσημασμένο συναθροιστικό σφάλμα που καταλήγει στο l και ξεκινά μετά το k
$R_m(k,l)$	Ελάχιστο προσημασμένο συναθροιστικό σφάλμα που καταλήγει στο l και ξεκινά μετά το k
$m(k,j,l)$	Ελάχιστο προσημασμένο συναθροιστικό σφάλμα που ξεκινά μεταξύ k,j και τελείωνει μεταξύ $j+1,l$
$M(k,j,l)$	Μέγιστο προσημασμένο συναθροιστικό σφάλμα που ξεκινά μεταξύ k,j και τελείωνει μεταξύ $j+1,l$
$s(k,l)$	Σφάλμα της συναθροιστικής ερώτησης με στοιχεία από k μέχρι l δείκτη

Πίνακας 3.1: Σύμβολα Συναθροιστικών Ερωτήσεων

από το d_{j+1}. Αν συνδυάσουμε το $R_M(k,j)$ με το $L_M(j+1,l)$ δεν χρειάζεται να εξετάσουμε οποιαδήποτε άλλα ενδιάμεσα αθροίσματα γιατί έτσι θα έχουμε πάρει το μέγιστο δυνατό ενδιάμεσο άθροισμα. Αυτό αποδεικνύεται εύκολα με απαγωγή σε άτοπο.

Έστω $R(k,j)$, ένα από τα σφάλματα που συνδυάζουμε, το οποίο ενώ δεν είναι το μέγιστο από τα αριστερά συνδυαζόμενο με κάποιο $L(j+1,l)$ δίνει το μέγιστο ενδιάμεσο σφάλμα. Τότε θα υπάρχει τουλάχιστον ένα σφάλμα το οποίο θα είναι ο συνδυασμός του $R_M(k,j)$ με το ίδιο $L(j+1,l)$ και το οποίο θα είναι προφανώς μεγαλύτερο από το μέγιστο ενδιάμεσο άθροισμα που έχουμε υποθέσει. Άρα καταλήξαμε σε άτοπο, οπότε η υπόθεση δεν ισχύει, δηλαδή για την εύρεση του μέγιστου ενδιάμεσου αθροίσματος αρκεί να συνδυάσουμε μόνο τα μέγιστα από αριστερά και δεξιά.

□

Ομοίως αποδεικνύεται ότι το ελάχιστο σφάλμα $m(k,j,l)$ του διαστήματος βρίσκεται γνωρίζοντας τα μεγέθη $R_m(k,j)$,$L_m(j+1,l)$. Δηλαδή ισχύει το παρακάτω λήμμα.

Λήμμα 3.2.2. *Έστω $d(k,l) = d_k,\ldots,d_j,\ldots,d_l$ ένα υποδιάστημα των δεδομένων που χωρίζεται με βάση το στοιχείο d_j σε δύο διαδοχικά υποδιαστήματα, μπορούμε να υπολογίσουμε το ελάχιστο σφάλμα $m(k,j,l)$ όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχουν δεδομένα και από τα δύο υποδιαστήματα, αν γνωρίζουμε τα μεγέθη $R_m(k,j)$,$L_m(j+1,l)$.*

Στη συνέχεια θα παραθέσουμε ένα θεώρημα που χρησιμοποιεί τα δύο παραπάνω λήμματα για την εύρεση του μεγίστου απολύτου συναθροιστικού σφάλματος.

Θεώρημα 3.2.1. *Έστω $d(k,l) = d_k,\ldots,d_j,\ldots,d_l$ ένα υποδιάστημα των δεδομένων που χωρίζεται με βάση το στοιχείο d_j σε δύο διαδοχικά υποδιαστήματα, μπορούμε να υπολογίσουμε το μέγιστο απόλυτο σφάλμα $I(k,l)$ όλων των δυνατών συναθροιστικών*

ερωτήσεων που περιέχονται στο υποδιάστημα $d(k,l)$, αν γνωρίζουμε τα μεγέθη $I(k,j)$,$I(j+1,l)$,$m(k,j,l)$,$M(k,j,l)$.

Απόδειξη. Το μέγιστο απόλυτο σφάλμα $I(k,l)$ όλων των δυνατών συναθροιστικών ερωτήσεων που περιέχονται στο υποδιάστημα $d(k,l)$ θα ανήκει οπωσδήποτε σε μία από τις παρακάτω τρεις περιπτώσεις:

- Το μέγιστο απόλυτο σφάλμα προκύπτει εξολοκλήρου από στοιχεία με μέγιστο δείκτη j οπότε δίνεται απευθείας από την τιμή του $I(k,j)$ όπως προκύπτει και από τον ορισμό του συμβολισμού που ορίσαμε παραπάνω.

- Το μέγιστο σφάλμα προκύπτει εξολοκλήρου από στοιχεία με ελάχιστο δείκτη $j+1$ οπότε δίνεται απευθείας από την τιμή του $I(j+1,l)$.

- Το μέγιστο σφάλμα αντιστοιχεί σε συναθροιστική ερώτηση που περιέχει στοιχεία και από τα δύο σύνολα. Σύμφωνα με τα Λήμματα 1,2 οι τιμές $m(k,j,l)$,$M(k,j,l)$ μας δίνουν το ελάχιστο και μέγιστο προσημασμένο σφάλμα για ερώτηση που περιέχει στοιχεία και από τα δύο υποδιαστήματα. Άρα σε αυτή την περίπτωση το μέγιστο απόλυτο σφάλμα θα είναι το μέγιστο κατά απόλυτη τιμή των μεγεθών $m(k,j,l)$,$M(k,j,l)$.

Αποδείξαμε ότι το μέγιστο απόλυτο σφάλμα θα είναι σε κάθε περίπτωση ένα από αυτά τα τέσσερα μεγέθη, οπότε αρκεί η σύγκριση των απολύτων τιμών των $I(k,j)$,$I(j+1,l)$,$m(k,j,l)$,$M(k,j,l)$ για την εύρεση του μεγίστου απολύτου συναθροιστικού σφάλματος. □

Στη συνέχεια θα παραθέσουμε ένα Λήμμα για την δημιουργία των μέτρων σφάλματος στα άκρα.

Λήμμα 3.2.3. *Έστω $d(k,l) = d_k,\ldots,d_j,\ldots,d_l$ ένα υποδιάστημα των δεδομένων που χωρίζεται με βάση το στοιχείο d_j σε δύο διαδοχικά υποδιαστήματα, μπορούμε να υπολογίσουμε το μέγιστο σφάλμα $L_M(k,l)$ όλων των συναθροιστικών ερωτήσεων που ξεκινούν από το στοιχείο d_k, αν γνωρίζουμε τα μεγέθη $L_M(k,j)$, $s(k,j)$,$L_M(j+1,n)$.*

Απόδειξη. Για το μέγιστο σφάλμα $L_M(k,l)$ όλων των συναθροιστικών ερωτήσεων που ξεκινούν από το στοιχείο d_k θα ισχύει μία από τις παρακάτω δύο περιπτώσεις:

- Όλα τα στοιχεία να έχουν μέγιστο δείκτη j οπότε το ζητούμενο άθροισμα δίνεται απευθείας από το $L_M(1,j)$.

- Το ζητούμενο μέγιστο άθροισμα περιέχει στοιχεία και από τα δύο υποδιαστήματα. Σε αυτή την περίπτωση θα πρέπει να περιέχει όλα τα στοιχεία του πρώτου υποδιαστήματος (αφού ξεκινά από το αριστερό άκρο και περιλαμβάνει και κάποια στοιχεία από το δεξί υποδιάστημα). Άρα αρκεί να συνδυάσουμε το άθροισμα όλων των σφαλμάτων του αριστερού υποδιαστήματος με το μέγιστο σφάλμα από τα δεξιά $L_M(j+1,n)$. Πράγματι δεν χρειάζεται να υπολογίσουμε τα υπόλοιπα αθροίσματα γιατί όλα θα έχουν τον δεύτερο όρο του αθροίσματος μικρότερο (εφόσον έχουμε θεωρήσει ότι παίρνουμε το μέγιστο).

□

Ομοίως αποδεικνύεται ότι με αντίστοιχο τρόπο μπορεί να υπολογιστεί το $R_M(k,l)$ καθώς και τα αντίστοιχα μέτρα για τα ελάχιστα αθροίσματα. Ενδεικτικά αναφέρουμε το παρακάτω λήμμα χωρίς απόδειξη.

Λήμμα 3.2.4. *Έστω $d(k,l) = d_k, \ldots, d_j, \ldots, d_l$ ένα υποδιάστημα των δεδομένων που χωρίζεται με βάση το στοιχείο d_j σε δύο διαδοχικά υποδιαστήματα, μπορούμε να υπολογίσουμε το μέγιστο σφάλμα $R_M(k,l)$ όλων των συναθροιστικών ερωτήσεων που ξεκινούν από το στοιχείο d_k, αν γνωρίζουμε τα μεγέθη $R_M(k,j)$, $s(j+1,k)$,$R_M(j+1,n)$.*

Τα παραπάνω λήμματα και θεωρήματα θα χρησιμοποιηθούν για την εύρεση της βέλτιστης λύσης ώστε να μην υπολογίζονται εξαντλητικά τα σφάλματα όλων των δυνατών συναθροιστικών ερωτήσεων παρά μόνο των μέγιστων τιμών αυτών, που είναι και η ποσότητα προς ελαχιστοποίηση.

3.3 Γενικός Αλγόριθμος RangeHaar

3.3.1 Εισαγωγή

Σε αυτό το κεφάλαιο θα παρουσιάσουμε τον γενικό αλγόριθμο για την επίλυση του προβλήματος των συναθροιστικών ερωτήσεων. Ο γενικός αλγόριθμος λύνει το πρόβλημα αφαιρετικά για κάθε μετασχηματισμό Haar χωρίς να υπεισέρχεται σε λεπτομέρειες που αφορούν τον τρόπο με το οποίο λειτουργεί ο κάθε μετασχηματισμός. Οι επιμέρους λεπτομέρειες μπορούν εύκολα να προστεθούν στη γενική λογική του αλγορίθμου ώστε να προκύψουν οι αλγόριθμοι που τελικά θα χρησιμοποιηθούν στην πράξη. Επιλέγουμε να παρουσιάσουμε τον γενικό αυτό αλγόριθμο με αυτόν τον τρόπο για να δείξουμε την ενιαία λογική του προβλήματος των συναθροιστικών ερωτήσεων ανεξάρτητα από τον μετασχηματισμό και να προσφέρουμε μία όσο το δυνατόν πιο γενική λύση του προβλήματος.

Αρχικά επανερχόμαστε στον ορισμό του προβλήματος ώστε να αναλύσουμε και να μοντελοποιήσουμε τη λύση του. Αν θεωρήσουμε ότι ο αλγόριθμος που θα προτείνουμε λειτουργεί από κάτω προς τα πάνω (bottom-up), τότε σε κάθε κόμβο η βέλτιστη σύνοψη επιλέγεται ώστε να βελτιστοποιούνται όχι μόνο οι συναθροιστικές ερωτήσεις που βρίσκονται υπό την εμβέλεια του κόμβου αυτού αλλά και όσες περιέχουν κάποιο κομμάτι από τα δεδομένα του υποδέντρου και επεκτείνονται εκτός αυτού. Όπως θα φανεί και παρακάτω το πρόβλημα αυτό είναι ιδιαίτερα δύσκολο διότι δεν μπορεί να αντιμετωπισθεί με τις κλασσικές τεχνικές του δυναμικού προγραμματισμού.

Επειδή όπως φαίνεται το πρόβλημα εξαρτάται από τα μέγιστα σφάλματα σε ανώτερο επίπεδο, η τοπικά βέλτιστη λύση (δηλαδή η βέλτιστη λύση για το υποδέντρο) δεν μας καλύπτει γιατι μπορεί να υπάρχει κάποια άλλη σύνοψη που ενδεχομένως να μην είναι βέλτιστη για το υποδέντρο αλλά να βελτιστοποιεί κάποιο σφάλμα σε ανώτερο επίπεδο. Αυτό προφανώς δεν μπορούμε να το γνωρίζουμε εκ των προτέρων αν δεν υπολογίσουμε τα σφάλματα σε όλο το δέντρο.

Σύμβολα του αλγορίθμου RangeHaar	
$I[i]$	Μέγιστο απόλυτο συναθροιστικό σφάλμα του υποδέντρου που έχει ως ρίζα τον συντελεστή c_i
$L_M[i]$	Μέγιστο προσημασμένο συναθροιστικό σφάλμα που ξεκινά από το αριστερό άκρο του υποδέντρου c_i
$L_m[i]$	Ελάχιστο προσημασμένο συναθροιστικό σφάλμα που ξεκινά από το αριστερό άκρο του υποδέντρου c_i
$R_M[i]$	Μέγιστο προσημασμένο συναθροιστικό σφάλμα που καταλήγει από το δεξί άκρο του υποδέντρου c_i
$R_m[i]$	Ελάχιστο προσημασμένο συναθροιστικό σφάλμα που καταλήγει από το δεξί άκρο του υποδέντρου c_i
$m[i]$	Ελάχιστο προσημασμένο συναθροιστικό σφάλμα που περιέχει στοιχεία και από τα δύο υπόδεντρα-παιδιά του c_i
$M[i]$	Μέγιστο προσημασμένο συναθροιστικό σφάλμα που περιέχει στοιχεία και από τα δύο υπόδεντρα-παιδιά του c_i
$r[i]$	Άθροισμα όλων των πραγματικών τιμών των στοιχείων που περιέχονται στο υπόδεντρο c_i
$s[i]$	Σφάλμα της συναθροιστικής ερώτησης που περιέχει όλα τα στοιχεία του υποδέντρου c_i

Πίνακας 3.2: Σύμβολα του αλγορίθμου RangeHaar

3.3.2 Μοντελοποίηση Προβλήματος

Στο προηγούμενο εδάφιο αποδείξαμε ότι για την βελτιστοποίηση του ολικού σφάλματος χρειάζεται να γνωρίζουμε τα ελάχιστα και μέγιστα σφάλματα στα άκρα και το μέγιστο σφάλμα εντός του υποδέντρου κατά απόλυτη τιμή. Μοντελοποιώντας το πρόβλημα θα μπορούσαμε να πούμε ότι το στιγμιότυπο κάθε σύνοψης ορίζεται μονοσήμαντα από τιμή στον κόμβο c_i, το αρχικό μονοπάτι S και τις τιμές στους συντελεστές-απογόνους του κόμβου c_i δεδομένου ότι απομένει χώρος b. Το στιγμιότυπο αυτό μπορεί να αντιστοιχισθεί σε ένα διάνυσμα σφάλματος $V = [I[i], L_M[i], L_m[i], R_M[i], R_m[i]]$ που περιέχει το μέγιστο σφάλμα εντός του υποδέντρου, το μέγιστο και ελάχιστο σφάλμα στα αριστερό και δεξί άκρο του υποδέντρου αντίστοιχα. Όπως φαίνεται από τον ορισμό πολλά διαφορετικά στιγμιότυπα συνόψεων μπορούν να αντιπροσωπεύονται από το ίδιο διάνυσμα σφάλματος αφού μπορεί να προκαλούν ίδιες τιμές στα μετρούμενα σφάλματα.

Πριν προχωρήσουμε στην περιγραφή του αλγορίθμου αναφέρουμε συγκεντρωτικά στον Πίνακα 3.2 τον συμβολισμό που θα χρησιμοποιήσουμε για την δημιουργία του αλγορίθμου.

3.3.3 Περιγραφή Αλγορίθμου RangeHaar

Αρχικά θα αναλύσουμε τα βασικά σημεία του αλγορίθμου και τον τρόπο με τον οποίο προέκυψαν με βάση την θεωρητική μελέτη που έχει προηγηθεί. Όπως έχει ήδη αναφερθεί ο αλγόριθμος λειτουργεί από κάτω προς τα πάνω(bottom-up) και σε κάθε επίπεδο υπολογίζει για κάθε στιγμιότυπο μίας σύνοψης το διάνυσμα σφάλματος. Στη συνέχεια αποφασίζει αν θα κρατήσει το συγκεκριμένο διάνυσμα σφάλματος ανάλογα με το αν πληρούνται οι προϋποθέσεις

σύμφωνα με το κριτήριο μερικής διάταξης που θα αναλύσουμε παρακάτω των διανυσμάτων. Εάν τηρείται το κριτήριο προσθέτει στο αντίστοιχο στιγμιότυπο της περίληψης που ανήκει.

Αναδρομικός Υπολογισμός Διανυσμάτων Σφάλματος

Ξεκινώντας θα πρέπει να αναφέρουμε την αντιστοιχία ανάμεσα στους συμβολισμούς που χρησιμοποιήσαμε κατά την θεωρητική μελέτη και τους συμβολισμούς που χρησιμοποιούμε κατά την περιγραφή του αλγορίθμου. Στην θεωρητική ανάλυση του προβλήματος θεωρούσαμε ένα υποδιάστημα των δεδομένων $d(k,l) = d_k, \ldots, d_j, \ldots, d_l$ που χωρίζεται με βάση το στοιχείο d_j σε δύο διαδοχικά υποδιαστήματα και αντίστοιχα ορίζαμε π.χ. ως $L_M(k,j)$ το μέγιστο σφάλμα όλων των ερωτήσεων που βρίσκονται μέσα στο διάστημα $d(k,j)$ και ξεκινούν από το αριστερό του άκρο. Στον αντίστοιχο συμβολισμό του αλγορίθμου χρησιμοποιούμε ένα μόνο δείκτη για να δηλώσουμε το αντίστοιχο σφάλμα $L_M[i]$. Αυτό συμβαίνει γιατί κάθε δέντρο $Haar$ είναι ένα δυαδικό δέντρο, οπότε κάθε κόμβος του δέντρου (που δηλώνεται μονοσήμαντα από τον δείκτη του) έχει υπό την εμβέλεια του ένα συγκεκριμένο υποδιάστημα των δεδομένων που είναι τα φύλλα όλων των απογόνων του κόμβου. Άρα ουσιαστικά όπως φαίνεται ο συμβολισμός που χρησιμοποιείται για την περιγραφή του αλγορίθμου παριστάνει μία ειδική περίπτωση των όσων έχουν αναφερθεί κατά τη θεωρητική μελέτη όπου ο κάθε συντελεστής του δυαδικού δέντρου ορίζει ένα υποδιάστημα των δεδομένων. Επιπλέον παρατηρούμε την ιδιότητα ότι αν κάποιος κόμβος c_i ορίζει ένα διάστημα των δεδομένων τότε τα παιδιά του θα διαμερίζουν το διάστημα αυτό σε δύο διαδοχικά διαστήματα ίσου μεγέθους.

Πλέον μπορούμε να εφαρμόσουμε τα λήμματα και τα θεωρήματα που αποδείξαμε για την δημιουργία του διανύσματος σφάλματος. Θεωρούμε ότι ο κάθε κόμβος παίρνει από τα παιδιά του μία λίστα από όλα τα πιθανά βέλτιστα στιγμιότυπα περιλήψεων και τα συνδυάζει ώστε να φτιάξει την δική του λίστα από πιθανά βέλτιστα στιγμιότυπα. Κάθε στιγμιότυπο θα εξετάζεται από τον κόμβο με βάση το διάνυσμα σφάλματος του παιδιού του. Ατό σημαίνει ότι τελικά ο κόμβος θα κάνει όλους τους δυνατούς συνδυασμούς των διανυσμάτων σφάλματος των παιδιών του. Αν θεωρήσουμε ότι ο κόμβος c_i ορίζει ένα διάστημα των δεδομένων, τότε τα παιδιά του όπως έχουμε αναφέρει θα διαμερίζουν το διάστημα αυτό σε δύο διαδοχικά διαστήματα. Η εισερχόμενη τιμή έστω u θα ισούται με τη μέση ανακατασκευασμένη τιμή στο υποδέντρο στο οποίο εισέρχεται οπότε το σφάλμα της ερώτησης που περιέχει όλα τα στοιχεία του διαστήματος που ορίζει ο πατέρας-κόμβος μπορεί να υπολογιστεί απευθείας αν αφαιρέσουμε από το άθροισμα των πραγματικών τιμών των στοιχείων που περιέχονται στο υπόδεντρο την εισερχόμενη στον κόμβο τιμή πολλαπλασιαζόμενη με τον αριθμό των στοιχείων που περιέχονται στο υπόδεντρο. Επιπλέον τα αντίστοιχα σφάλματα των κόμβων-παιδιών θα δίνονται με την ίδια λογική από την εισερχόμενη τιμή στα αντίστοιχα υπόδεντρα-παιδιά, η οποία καθορίζεται με βάση την εισερχόμενη τιμή στον κόμβο-πατέρα και την τιμή του συντελεστή c_i και υποδιπλασιάζοντας τον αριθμό των στοιχείων. Αυτό σημαίνει ότι μπορούμε εύκολα να υπολογίσουμε τα εν λόγω σφάλματα από τις παρακάτω σχέσεις:

$$s[i] = r[i] - ku$$

$$s[2i] = s[i] + (k/2)c_i$$

$$s[2i+1] = s[i] - (k/2)c_i$$

Όπου k ο αριθμός των στοιχείων που περιλαμβάνονται στο υπόδεντρο. Στη συνέχεια θα υπολογίσουμε με βάση τα Λήμματα 3.2.3-3.2.4 το διάνυσμα σφάλματος του πατέρα-κόμβου για ένα συγκεκριμένο στιγμιότυπο από τα διανύσματα σφάλματος των παιδιών. Οπότε θα έχουμε ότι:

$$L_M[i] = \max(L_M[2i], s[2i] + L_M[2i+1])$$

$$L_m[i] = \min(L_m[2i], s[2i] + L_m[2i+1])$$

$$R_M[i] = \max(R_M[2i] + s[2i+1], R_M[2i+1])$$

$$R_m[i] = \min(R_m[2i] + s[2i+1], R_m[2i+1])$$

Για να βρούμε το μέγιστο και ελάχιστο ενδιάμεσο σφάλμα έστω $M[i], m[i]$ χρησιμοποιούμε τα Λήμματα 3.2.1-3.2.2 οπότε θα έχουμε:

$$M[i] = \max(I[2i], I[2i+1], R_M[2i] + L_M[2i+1])$$

$$m[i] = R_m[2i] + L_m[2i+1]$$

Έπειτα για να βρούμε το μέγιστο απόλυτο σφάλμα εντός του υποδέντρου χρησιμοποιούμε το Θεώρημα 3.2.1. Οπότε παίρνουμε την παρακάτω σχέση:

$$I[i] = \max(I[2i], I[2i+1], |R_M[2i] + L_M[2i+1]|, |R_m[2i] + L_m[2i+1]|)$$

Σύμφωνα με τα παραπάνω θα έχουμε υπολογίσει επιτυχώς το διάνυσμα σφάλματος του πατέρα έστω $V = [I[i], L_M[i], L_m[i], R_M[i], R_m[i]]$ από τα διανύσματα σφάλματος των κόμβων-παιδιών.

Με βάση τα παραπάνω μπορούμε ως μία πρώτη προσέγγιση στο πρόβλημα να προτείνουμε τον αναδρομικό υπολογισμό των διανυσμάτων σφάλματος κάθε περίληψης από τα διανύσματα σφάλματος των κόμβων-παιδιών. Αυτό σημαίνει ότι για δεδομένο αρχικό μονοπάτι S και ελεύθερο χώρο b θα πρέπει να αποθηκεύουμε σε κάθε κόμβο i όλες τις δυνατές συνόψεις στις οποίες συμμετέχει, οι οποίες ενδεχομένως να οδηγήσουν στην βέλτιστη καθολική λύση. Ελλείψει κάποιου κριτηρίου, θα πρέπει να αποθηκεύσουμε όλες τις δυνατές συνόψεις και να τις συνδυάζουμε στο ανώτερο επίπεδο (συνδυάζοντας τα αντίστοιχα διανύσματα σφάλματος όπως περιγράφτηκε παραπάνω). Αυτή η λύση είναι φανερό ότι είναι ιδιαίτερα μη αποδοτική αφού απαιτεί εκθετικό χώρο και χρόνο. Αυτό συμβαίνει γιατί ξεκινώντας αρχικά από το κατώτατο επίπεδο εξετάζουμε ένα πεπερασμένο αριθμό συνόψεων (που αντιστοιχούν σε έναν αριθμό διανυσμάτων σφάλματος) έστω a τις οποίες αποθηκεύουμε. Στο αμέσως ανώτερο επίπεδο κάθε κόμβος πρέπει να συνδυάσει τις a συνόψεις του κάθε κόμβου-παιδιού (συνδυάζοντας τα διανύσματα σφάλματος στα οποία αντιστοιχούν) οπότε θα προκύψουν a^2 συνόψεις οι οποίες αποθηκεύονται προς χρήση του επόμενου κόμβου-πατέρα. Στο αμέσως

επόμενο επίπεδο θα δημιουργηθούν a^4 συνόψεις, ώσπου τελικά στη ρίζα θα έχουμε a^N συνόψεις. Δηλαδή καταλήγουμε να χρειαζόμαστε εκθετικό χώρο για να αποθηκεύσουμε όλες τις πιθανές βέλτιστες περιλήψεις και εκθετικό χρόνο για να τις συνδυάσουμε σε κάθε βήμα και να εξετάσουμε τελικά ποια είναι η καλύτερη.

3.3.4 Μερική Διάταξη Στιγμιοτύπων

Οι Deligiannakis και Roussopoulos στο [1] χρησιμοποίησαν έναν αλγόριθμος μερικής διάταξης για να επιλύσουν το πρόβλημα βέλτιστης σύνοψης σε πολυδιάστατα δεδομένα. Αντίστοιχα στο πρόβλημα που εξετάζουμε έχοντας ως στόχο να μειώσουμε την πολυπλοκότητα του αλγορίθμου προσπαθούμε να εισάγουμε κάποιο κριτήριο μερικής διάταξης των στιγμιότυπων με βάση το οποίο θα μπορούμε να απορρίπτουμε τα στιγμιότυπα ορισμένων συνόψεων εφόσον γνωρίζουμε ότι δεν οδηγούν σε καμία περίπτωση στην βέλτιστη λύση. Η έννοια της μερικής διάταξης απορρέει από το γεγονός ότι ορισμένα διανύσματα θα υπερτερούν σε κάποια πεδία και θα υστερούν σε κάποια άλλα έναντι ορισμένων άλλων διανυσμάτων.

Το κριτήριο θα εφαρμοστεί ως εξής. Κάθε κόμβος θα κρατά ένα σύνολο από διανύσματα σφάλματος που έχουν επιλεχθεί σύμφωνα με το κριτήριο μερικής διάταξης. Αρχικά το σύνολο αυτό είναι κενό οπότε το πρώτο διάνυσμα θα κρατηθεί χωρίς τον έλεγχο του κριτηρίου. Στη συνέχεια για κάθε νέο διάνυσμα σφάλματος θα γίνεται σύγκριση με τα διανύσματα που υπάρχουν στο σύνολο με βάση το κριτήριο μερικής διάταξης. Αρχικά θεωρούμε ότι πρέπει να κρατήσουμε το διάνυσμα αυτό. Η κατάσταση αυτή αλλάζει αν βρεθεί τουλάχιστον ένα διάνυσμα μέσα στο σύνολο το οποίο να υπερέχει σύμφωνα με το κριτήριο μερικής διάταξης. Το διάνυσμα συγκρίνεται με όλα τα διανύσματα του συνόλου έτσι ώστε να βρεθούν αν υπάρχουν διανύσματα που υπολείπονται με βάση το κριτήριο. Άρα αν βρεθεί κάποιο διάνυσμα το οποίο να υπολείπεται έναντι του νέου διανύσματος, τότε αυτό διαγράφεται άμεσα από το σύνολο. Συνοπτικά θα λέγαμε ότι κάθε νέο διάνυσμα συγκρίνεται με κάθε διάνυσμα του συνόλου, διαγράφοντας όλα τα διανύσματα έναντι των οποίων υπερέχει και αλλάζοντας την κατάσταση του σε μη αποθηκεύσιμο αν βρει κάποιο διάνυσμα που είναι καλύτερο από αυτό. Όταν ο αλγόριθμος έχει σαρώσει όλο το σύνολο αποθηκεύει ή διαγράφει το νέο διάνυσμα ανάλογα με την κατάσταση που έχει τεθεί. Σε περίπτωση που θα πρέπει να αποθηκευθεί το διάνυσμα σφάλματος, τότε αποθηκεύουμε και το στιγμιότυπο της σύνοψης στο οποίο ανήκει και ενημερώνουμε αν χρειάζεται τη μεταβλητή που αποθηκεύει το μέγιστο από τα μέγιστα απόλυτα σφάλματα του στιγμιότυπου.

Όπως αναφέραμε στην προηγούμενη παράγραφο το στιγμιότυπο κάθε σύνοψης αντιπροσωπεύεται από ένα διάνυσμα σφάλματος. Στην απλή περίπτωση των σημειακών ερωτήσεων το διάνυσμα σφάλματος αποτελείται από μία μόνο τιμή (το μέτρο του σημειακού σφάλματος που θέλουμε να βελτιστοποιήσουμε), οπότε μπορούμε εύκολα να διατάξουμε τα διανύσματα σφάλματος που προκύπτουν από τα διάφορα στιγμιότυπα περιλήψεων σε αύξουσα σειρά και να επιλέξουμε αυτό που βελτιστοποιεί (ελαχιστοποιεί) το ζητούμενο σφάλμα. Αντίθετα στην περίπτωση των συναθροιστικών ερωτήσεων το διάνυσμα σφάλματος αποτελείται από έξι σφάλματα εκ των οποίων μας ενδιαφέρει να βελτιστοποιήσουμε το μέγιστο εσωτερικό

απόλυτο σφάλμα εντός του υποδέντρου και το σφάλμα στα άκρα που επηρεάζει το μέγιστο απόλυτο συναθροιστικό σφάλμα σε ανώτερα επίπεδα. Στόχος λοιπόν του κριτηρίου είναι με βάση τη φύση του προβλήματος να μπορέσουμε να διατάξουμε μερικώς (υπό την έννοια ότι ορισμένα διανύσματα θα υπερτερούν σε κάποια πεδία και θα υστερούν σε κάποια άλλα έναντι ορισμένων άλλων διανυσμάτων) το σύνολο των διανυσμάτων σφάλματος. Σε κάθε περίπτωση θα υπάρχουν ορισμένα διανύσματα τα οποία υπολείπονται σε όλα τα πεδία και τα οποία θα πρέπει να απορρίπτονται γιατί δεν θα οδηγούν ποτέ σε βέλτιστη λύση. Απορρίπτοντας ορισμένα από τα διανύσματα ως μη βέλτιστα, απορρίπτουμε συγχρόνως τα στιγμιότυπα των περιλήψεων που συνδέονται με αυτά και τελικά καταλήγουμε σε μία μερική διάταξη των στιγμιοτύπων των συνόψεων. Αρχικά διατυπώνουμε τον παρακάτω ορισμό:

Ορισμός. *Θα λέμε ότι μία πλειάδα* $V = [I, L_M, L_m, R_M, R_m]$ *υπερτερεί μίας άλλης πλειάδας* $V' = [I', L'_M, L'_m, R'_M, R'_m]$, *όπου* $I < I'$,$L_M < L'_M$, $L_m > L'_m$,$R_M < R'_M$ *και* $R_m > R'_m$.

Έχοντας ορίσει την έννοια της υπεροχής ενός διανύσματος σφάλματος θα αποδείξουμε ότι αν η πλειάδα V υπερτερεί μίας άλλης πλειάδας V'τότε η πλειάδα V δημιουργεί ένα διάνυσμα σφάλματος στην ρίζα του δέντρου V_{root} που είναι σε κάθε περίπτωση καλύτερο (δηλαδή έχει μικρότερο απόλυτο μέγιστο σφάλμα $I_M[i]$) από το διάνυσμα σφάλματος V'_{root} που δημιουργείται από την πλειάδα V'. Παρακάτω παραθέτουμε το αντίστοιχο Θεώρημα.

Θεώρημα Μερικής Διάταξης Στιγμιοτύπων. *Αν η πλειάδα* V *υπερτερεί μίας άλλης πλειάδας* V'*τότε η πλειάδα* V *δημιουργεί ένα διάνυσμα σφάλματος στην ρίζα του δέντρου* V_{root} *που είναι σε κάθε περίπτωση καλύτερο (δηλαδή έχει μικρότερο απόλυτο μέγιστο σφάλμα* $I_M[i]$*) από το διάνυσμα σφάλματος* V'_{root} *που δημιουργείται από την πλειάδα* V'.

Απόδειξη. Για να αποδείξουμε το παραπάνω θεώρημα θα πρέπει να δείξουμε ότι το διάνυσμα που υπερτερεί ανεξάρτητα με ποιο διάνυσμα θα συνδυαστεί, θα δώσει σίγουρα ένα καλύτερο διάνυσμα (δηλαδή ένα διάνυσμα που θα υπερτερεί έναντι του άλλου) στο ανώτερο επίπεδο σε σχέση με το διάνυσμα που θα δώσει η υπολειπόμενη πλειάδα. Η διαδικασία αυτή θα επαναλαμβάνεται, δηλαδή το διάνυσμα που υπερέχει θα γεννά διανύσματα που υπερέχουν σε ανώτερο επίπεδο ώσπου να φτάσουμε στη ρίζα οπότε θα έχουμε αποδείξει πλήρως το θεώρημα. Για τα παρακάτω θεωρούμε ότι θέλουμε να παράγουμε ένα διάνυσμα σφάλματος ενός κόμβου-πατέρα έστω $V = [I, L_M, L_m, R_M, R_m]$ από τα διανύσματα σφάλματος των παιδιών έστω $V_L = [I^L, L^L_M, L^L_m, R^L_M, R^L_m], V_R = [I^R, L^R_M, L^R_m, R^R_M, R^R_m]$ του αριστερού και δεξιού υποδέντρου αντίστοιχα. Χωρίς βλάβη της γενικότητας θεωρούμε ένα άλλο διάνυσμα $V'_L = [I'^L, L'^L_M, L'^L_m, R'^L_M, R'^L_m]$ το οποίο υπολείπεται έναντι του V_R ή συμβολικά $\text{dominate}(V^L, V'^L)$ και υπολογίζουμε για το τυχαίο διάνυσμα V_L το αντίστοιχο παραγόμενο διάνυσμα του κόμβου-πατέρα $V'_1 = [I', L'_M, L'_m, R'_M, R'_m]$. Αντίστοιχα μπορούμε να αποδείξουμε ότι καταλήγουμε στα ίδια συμπεράσματα αν θεωρήσουμε $\text{dominate}(V^R, V'^R)$. Θέλουμε να δείξουμε ότι το V_1 θα υπερέχει έναντι του V'_1. Για κάθε σφάλμα θα πρέπει να ισχύει η ανισότητα που τίθεται στον ορισμό. Αρχικά θα αποδείξουμε ότι $I < I'$. Το μέγιστο απόλυτο σφάλμα θα βρίσκεται είτε εντός των δύο υποδέντρων είτε θα περιέχει στοιχεία και από τα δύο υπόδεντρα. Για την περίπτωση που το μέγιστο σφάλμα βρίσκεται

μέσα στο αριστερό υποδέντρο θα ισχύει ότι $I < I'$ εξαιτίας της σχέσης $I^L < I'^L$. Προφανώς στην περίπτωση που το μέγιστο απόλυτο σφάλμα βρίσκεται στο δεξί υποδέντρο το διάνυσμα V_L δεν θα επηρεάσει αυτό το μέγεθος σφάλματος. Συνεχίζουμε θεωρώντας την τελευταία περίπτωση όπου το μέγιστο συναθροιστικό σφάλμα προκύπτει από στοιχεία και των δύο υποδέντρων. Σύμφωνα με τα όσα έχουμε πει για τον συνδυασμό των διανυσμάτων σφάλματος στην προηγούμενη παράγραφο θα πρέπει να κάνουμε τους συνδυασμούς $L_M^R + R_M^L$ και $L_m^R + R_m^L$ ώστε να λάβουμε το μέγιστο απόλυτο σφάλμα I για τον κόμβο-πατέρα. Πρέπει να αποδείξουμε ότι το διάνυσμα σφάλματος που υπολείπεται με βάση το κριτήριο μερικής διάταξης δεν δίνει ποτέ καλύτερη λύση. Για τα σφάλματα στα άκρα που συνδυάζονται σε ανώτερο επίπεδο θα ισχύει μία από τις παρακάτω περιπτώσεις:

(α) Τα μέγιστα σφάλματα που συνδυάζονται να είναι ομόσημα θετικά και τα αντίστοιχα ελάχιστα σφάλματα να είναι ομόσημα αρνητικά. Προφανώς το κριτήριο οδηγεί σε βέλτιστο μέγιστο απόλυτο σφάλμα I (όπου $I < I'$), αφού το άθροισμα δύο θετικών αριθμών $L_M^R + R_M^L$ θα είναι ελάχιστο εφόσον το L_M^R θεωρείται σταθερό και R_M^L έχει μικρότερη τιμή ($R_M^L < R_M'^L$). Αντίστοιχα το άθροισμα των ελαχίστων ελαχιστοποιείται αφού το L_m^R είναι σταθερός αρνητικός αριθμός και το R_m^L θα είναι ένας μεγαλύτερος αρνητικός αριθμός ($R_m^L > R_m'^L$).

(β) Τα μέγιστα σφάλματα να είναι ομόσημα θετικά και τα ελάχιστα σφάλματα να είναι ομόσημα θετικά. Σε αυτή την περίπτωση η σχέση που ισχύει για τα μέγιστα $R_M^L < R_M'^L$ θα ελαχιστοποιεί το σφάλμα που προκύπτει από το συνδυασμό των μέγιστων σφαλμάτων. Αντίθετα η σχέση που θέλουμε να ισχύει για τα ελάχιστα θα αυξάνει το σφάλμα που προκύπτει από τα ελάχιστα γιατί θα προτιμάμε όσο το δυνατόν μεγαλύτερες θετικές τιμές λόγω της σχέσης $R_m^L > R_m'^L$. Εδώ και πάλι το κριτήριο μερικής διάταξης δεν θα χάσει καμία λύση γιατί το σφάλμα που αυξάνει λόγω της σχέσης που θα πρέπει να ισχύει για τα ελάχιστα θα είναι σίγουρα μικρότερο από το σφάλμα που βελτιστοποιείται με βάση των σχέσεων για τα μέγιστα. Αυτό συμβαίνει γιατί εφόσον αναφερόμαστε σε θετικούς αριθμούς το άθροισμα των μεγίστων θα δίνει σίγουρα μεγαλύτερο σφάλμα από το σφάλμα των ελαχίστων. Δηλαδή ισχύει η παρακάτω σχέση:

$$L_m^R < L_M^R, R_m^L < R_M^L \wedge L_m^R, L_M^R, R_m^L, R_M^L \geq 0 \Rightarrow L_m^R + R_m^L < L_M^R + R_M^L = I$$

$$L_m^R < L_M^R, R_m'^L < R_M'^L \wedge L_m^R, L_M^R, R_m'^L, R_M'^L \geq 0 \Rightarrow L_m^R + R_m'^L < L_M^R + R_M'^L = I'$$

$$R_M^L < R_M'^L \Rightarrow L_M^R + R_M^L = I_1 < L_M^R + R_M'^L = I'$$

Με βάση τα παραπάνω έχουμε αποδείξει ότι και σε αυτή την περίπτωση το κριτήριο που έχουμε θέσει θα ελαχιστοποιεί το μέγιστο σφάλμα οπότε ισχύει και πάλι η σχέση $I < I'$.

(γ) Τα μέγιστα σφάλματα να είναι ομόσημα αρνητικά και τα ελάχιστα σφάλματα να είναι ομόσημα αρνητικά. Η απόδειξη ότι και πάλι το κριτήριο βελτιστοποιεί το μέγιστο σφάλμα είναι παρόμοια με την παραπάνω περίπτωση. Εδώ επειδή αναφερόμαστε σε

αρνητικούς αριθμούς, το μέγιστο κατά απόλυτη τιμή σφάλμα θα προέρχεται από τον συνδυασμό των ελαχιστών, οπότε θα θέλουμε να αυξήσουμε όσο το δυνατόν την τιμή των προσημασμένων ελαχίστων ($R_m^L > R_m'^L$). Οι αντίστοιχες σχέσεις για τα μέγιστα αυξάνουν το σφάλμα αλλά αυτο δεν μπορεί να οδηγήσει σε επιλογή χειρότερης λύσης αφού σίγουρα αυτό το σφάλμα θα είναι μικρότερο από αυτό που προκύπτει από το συνδυασμό των ελαχίστων. Με μαθηματικό συμβολισμό θα ισχύουν τα εξής:

$$L_m^R < L_M^R, R_m^L < R_M^L \wedge L_m^R, L_M^R, R_m^L, R_M^L \leq 0 \Rightarrow L_M^R + R_M^L < L_m^R + R_m^L = I$$

$$L_m^R < L_M^R, R_m'^L < R_M'^L \wedge L_m^R, L_M^R, R_m'^L, R_M'^L \leq 0 \Rightarrow L_M^R + R_M'^L < L_m^R + R_m'^L = I'$$

$$R_m^L > R_m'^L \Rightarrow \left|L_m^R + R_m^L\right| = I_1 < \left|L_m^R + R_m'^L\right| = I'$$

(δ) Τα μέγιστα σφάλματα να είναι ετερόσημα και ομοίως τα ελάχιστα σφάλματα να είναι ετερόσημα. αυτή η περίπτωση είναι λίγο πιο πολύπλοκη από τις προηγούμενες. Χωρίς βλάβη της γενικότητας θεωρούμε την περίπτωση που τα δεξιά σφάλματα που συνδυάζονται είναι θετικά και τα αριστερά αρνητικά, δηλαδή ισχύει ότι $R_m^L, R_M^L \geq 0 \wedge L_m^R, L_M^R \leq 0$. Αφού τα αριστερά σφάλματα είναι αρνητικά θα ισχύουν τα εξής για τις απόλυτες τιμές τους:

$$L_m^R < L_M^R \leq 0 \Rightarrow \left|L_m^R\right| > \left|L_M^R\right|$$

Στη συνέχεια θεωρούμε την περίπτωση όπου $R_M^L < \left|L_M^R\right|$, ακριβώς με αντίστοιχο τρόπο μπορούμε αποδείξουμε την ισχύ του κριτηρίου αν $L_M^R < R_M^L$. Το σφάλμα R_M^L τείνει να μειωθεί λόγω της σχέσης $R_M^L < R_M'^L$ τείνει να μειωθεί ενώ το σφάλμα $\left|L_M^R\right|$ παραμένει σταθερό. Αυτό σημαίνει ότι η διαφορά $\left|R_M^L - \left|L_M^R\right|\right|$ που θα αποτελεί το παραγόμενο σφάλμα τείνει να αυξηθεί σε απόλυτη τιμή. Παράλληλα για τα ελάχιστα θα ισχύει $R_m^L < R_M^L < \left|L_M^R\right| < \left|L_m^R\right|$). Είναι φανερό από την προηγούμενη ανισότητα ότι το σφάλμα που προκύπτει από τα ελάχιστα ($R_m^L - \left|L_m^R\right| = I$)θα είναι οπωσδήποτε μεγαλύτερο από το σφάλμα που παράγεται από τα μέγιστα αφού διατάσσονται πιο μακρία στον θετικό αμιάξονα. Όμως με βάση το κριτήριο η τιμή R_m^L τείνει να αυξηθεί λόγω της σχέσης $R_m^L > R_m'^L$, ενώ η τιμή $|L_m|$ παραμένει σταθερή. Αυτό συνεπάγεται ότι το σφάλμα που παράγεται από τα ελάχιστα, δηλαδή η διαφορά $R_m^L - \left|L_m^R\right| = I$ τείνει να ελαχιστοποιηθεί που σημαίνει ότι $I < I'$. Οπότε σύμφωνα με τα παραπάνω αποδείξαμε ότι και σε αυτή την περίπτωση το κριτήριο εξασφαλίζει την βελτιστότητα του μέγιστου απολύτου σφάλματος.

(ε) Τα μέγιστα σφάλματα είναι ομόσημα θετικά και τα ελάχιστα σφάλματα είναι ετερόσημα. Χωρίς βλάβη της γενικότητας θεωρούμε ότι τα αριστερά σφάλματα είναι θετικά ενώ τα δεξιά είναι ετερόσημα δηλαδή ισχύει ότι $L_m^R, L_M^R, R_M^L \geq 0 R_m^L \leq 0$. Το άθροισμα των μέγιστων σφαλμάτων θα ελαχιστοποιείται εξορισμού αφού και οι δύο όροι του αθροίσματος είναι θετικοί και ο ένας από τους δύο R_M^L τείνει να μειωθεί λόγω της σχέσης $R_M^L < R_M'^L$. Για τα ελάχιστα σφάλματα το κριτήριο θα βελτιστοποιεί το παραγόμενο σφάλμα αν $L_m^R < \left|R_m^L\right|$ αφού βάσει του κριτηρίου το σφάλμα $\left|R_m^L\right|$ θα τείνει να μειωθεί

ενώ το σφάλμα L_m^R παραμένει σταθερό. Στην περίπτωση που ισχύει ότι $\left|R_m^L\right| < L_m^R$ το παραγόμενο σφάλμα τείνει να αυξηθεί για τον ίδιο λόγο. Όμως εξορισμού ισχύει ότι $\left|R_m^L\right| < L_m^R < L_M^R$. Το σφάλμα που προκύπτει από τα μέγιστα θα είναι οπωσδήποτε μεγαλύτερο του L_M^R αφού σε αυτό προστίθεται ένας θετικός αριθμός. Επιπλέον το σφάλμα που προκύπτει από τα ελάχιστα θα είναι οπωσδήποτε μικρότερο από L_m^R αφού προκύπτει από την αφαίρεση της τιμής του $\left|R_m^L\right|$ από το L_m^R. Συνοπτικά θα έχουμε ότι:

$$\left|R_m^L\right| < L_m^R \Rightarrow L_m^R > L_m^R - \left|R_m^L\right| > 0$$

$$I = L_M^R + R_M^L > L_M^R > L_m^R > L_m^R - \left|R_m^L\right| > 0$$

$$\left|R_m^L\right| < L_m^R \Rightarrow L_m^R > L_m^R - \left|R_m^L\right| > 0$$

$$I' = L_M^R + R_M'^L > I = L_M^R + R_M^L > L_M^R > L_m^R > L_m^R - \left|R_m'^L\right| > 0$$

Δηλαδή αποδείξαμε ότι το σφάλμα που θα παράγεται από τα μέγιστα και το οποίο ελαχιστοποιείται με βάση το κριτήριο είναι μεγαλύτερο από το σφάλμα που προκύπτει από τα ελάχιστα σφάλματα. Οπότε και σε αυτήν περίπτωση το κριτήριο λειτουργεί επιτυχώς.

(ζ) Τα μέγιστα σφάλματα είναι ομόσημα αρνητικά και τα ελάχιστα σφάλματα είναι ετερόσημα. Χωρίς βλάβη της γενικότητας θεωρούμε ότι τα αριστερά σφάλματα είναι αρνητικά ενώ τα δεξιά είναι ετερόσημα δηλαδή ισχύει ότι $L_m^R, L_M^R, R_m^L \leq 0 R_M^L \geq 0$. Το άθροισμα των ελαχίστων σφαλμάτων θα ελαχιστοποιείται εξορισμού αφού και ο ένας όρος είναι σταθερός L_m^R και ο άλλος τείνει να αυξηθεί ($L_m'^R < L_m^R$). Το άθροισμα των μεγίστων σφαλμάτων θα ελαχιστοποιείται με βάση το κριτήριο αν ισχύει ότι $\left|L_M^R\right| < R_M^L$ αφού το σφάλμα R_M^L θα μειώνεται ενώ το σφάλμα $\left|L_M^R\right|$ παραμένει σταθερό. Αντίθετα αν $\left|L_M^R\right| > R_M^L$ το σφάλμα με βάση το κριτήριο θα αυξάνει. Σε αυτή την περίπτωση όμως θα ισχύουν τα εξής:

$$R_M^L < \left|L_M^R\right| < \left|L_m^R\right|$$

$$\left|L_m^R\right| + \left|R_m^L\right| = I > \left|L_m^R\right| > R_M^L > R_M^L - \left|L_M^R\right| > 0$$

$$R_M^L < R_M'^L < \left|L_M^R\right| < \left|L_m^R\right|$$

$$I' = \left|L_m^R\right| + \left|R_m'^L\right| > I = \left|L_m^R\right| + \left|R_m^L\right|$$

Από την παραπάνω ανισότητα φαίνεται ότι το σφάλμα που προκύπτει από τα ελάχιστα και το οποίο βελτιστοποιείται είναι οπωσδήποτε μεγαλύτερο κατά απόλυτη τιμή από το σφάλμα που προκύπτει από τα μέγιστα.

Με βάση τα παραπάνω αποδείξαμε ότι το $I < I'$ για κάθε V^R dominates V'^R. Στη συνέχεια θα πρέπει να αποδείξουμε ότι και τα υπόλοιπα σφάλματα που παράγονται θα είναι βέλτιστα με βάση το κριτήριο. Πράγματι εύκολα μπορούμε να αποδείξουμε ότι τα σφάλματα L_M, L_m, R_M, R_m θα είναι βέλτιστα ως προς το κριτήριο έναντι των L'_M, L'_m, R'_M, R'_m αν θεωρήσουμε τις σχέσεις

από τις οποίες παράγονται, οπότε θα ισχύουν τα παρακάτω:

$$L_M = \max(L_M^L, s^L + L_M^R) \wedge L'_M = \max(L_M^L, s^L + L_M'^R) \wedge R_M^L < R_M'^L \Rightarrow L_M < L'_M$$

$$L_m = \min(L_m^L, s^L + L_m^R) \wedge L'_m = \min(L_m^L, s^L + L_m'^R) \wedge R_m^L > R_m'^L \Rightarrow L_m > L'_m$$

$$R_M = \max(R_M^R, s^R + R_M^L) \wedge R'_M = \max(R_M'^R, s'^R + R_M^L) \wedge s^R = s'^R \wedge R_M^R < R_M'^R \Rightarrow R_M < R'_M$$

$$R_m = \min(R_m^R, s^R + R_m^L) \wedge R'_m = \min(R_m'^R, s'^R + R_m^L) \wedge s^R = s'^R \wedge R_m^R > R_m'^R \Rightarrow R_m > R'_m$$

Το σφάλμα s είναι το σφάλμα της συναθροιστικής ερώτησης που περιέχει όλα τα στοιχεία του υποδέντρου. Υπενθυμίζουμε ότι το σφάλμα s εξαρτάται μόνο από την εισερχόμενη τιμή και άρα για δεδομένα (S, b) που εξετάζονται τα δύο διανύσματα V^L και V'^L θα είναι σταθερό. Η ιδιότητα αυτή ισχύει πάντα μόνο για τους M/Σ Haar και unrestricted Haar αφού κάθε συντελεστής προστίθεται και αφαιρείται στα δεδομένα του υποδέντρου. Αντίθετα για τον M/Σ Haar$^+$ δεν μπορούμε να κάνουμε αυτή την απλοποίηση γιατί οι συμπληρωματικοί συντελεστές προστίθενται σε ορισμένα μόνο στοιχεία διαφοροποιώντας το συναθροιστικό σφάλμα από την εισερχόμενη τιμή. Οπότε το θεώρημα μερικής διάταξης για το M/Σ Haar$^+$ θα περιλαμβάνει ακόμα και την συνθήκη $s = s'$. Τελικά σύμφωνα με τα όλα τα παραπάνω αποδείξαμε ότι για τους M/Σ Haar και unrestricted Haar, αν dominate(V_L, V'_L) τότε για τα διανύσματα στον κόμβο-πατέρα θα ισχύει dominate(V, V') . Αν τώρα πάρω ως διανύσματα V^L και V'^L, τα διανύσματα του κόμβου-πατέρα και εφαρμόσω την παραπάνω ιδιότητα θα πάρω στο πιο πάνω επίπεδο dominate(V_2, V'_2) . Η διαδικασία αυτή μπορεί να συνεχιστεί μέχρι να φτάσουμε στην κορύφη όπου θα ισχύει dominate(V_{root}, V'_{root}) , δηλαδή αποδείξαμε πλήρως τον ισχυρισμό μας ότι αν dominate(V, V'), τότε μπορούμε να απορρίψουμε το υπολειπόμενο διάνυσμα αφού δε θα μας οδηγήσει σε καμία περίπτωση σε καλύτερη λύση.

□

,

Στο Σχήμα 3.1 φαίνεται ο γενικός αλγόριθμος σε ψευδοκώδικα. Όπως φαίνεται η κύρια μέθοδος θα επιστρέψει ένα διδιάστατο πίνακα έστω E που για κάθε αρχικό μονοπάτι S και υπολειπόμενο χώρο b θα αποθηκεύει ένα σύνολο από συνόψεις που πιθανόν οδηγούν σε βέλτιστη λύση. Με τη μέθοδο dominate εξετάζεται το κριτήριο μερικής διάταξης και απορρίπτονται όσα στιγμιότυπα δεν οδηγούν σίγουρα στην βέλτιστη λύση.

3.3.5 Εύρεση Βέλτιστης Σύνοψης

Είναι φανερό ότι για να οριστεί πλήρως η βέλτιστη σύνοψη χρειάζεται μία επιπλέον επεξεργασία του πίνακα E που επιστρέφει η μέθοδος που περιγράψαμε στην προηγούμενη παράγραφο. Ο πίνακας E για την ρίζα του δέντρου θα περιέχει για κάθε στιγμιότυπο S, b όλες τις πιθανές βέλτιστες συνόψεις. Τα διανύσματα σφάλματος στην ρίζα του δέντρου θα αντιστοιχούν τα μέτρα σφάλματος που υπολογίζουμε σε κάθε βήμα με τα στιγμιότυπα των συνόψεων. Φτάνοντας στην κορυφή θα θέλαμε να ελαχιστοποιήσουμε το μέγιστο απόλυτο συναθροιστικό σφάλμα I του διανύσματος σφάλματος της ρίζας. Οπότε σαρώνουμε τον

πίνακα και βρίσκουμε την ελάχιστη τιμή του σφάλματος αυτού. Αυτή η ελάχιστη τιμή θα αντιπροσωπεύει την βέλτιστη σύνοψη, όποτε έχοντας βρει τους μη μηδενικούς συντελεστές του δέντρου που δίνουν το ελάχιστο μέγιστο απόλυτο σφάλμα για κάθε συναθροιστική ερώτηση έχουμε λύσει πλήρως το πρόβλημα.

3.4 Επέκταση Αλγορίθμου για τον M/Σ Unrestricted Haar

Στο προηγούμενο εδάφιο παρουσιάσουμε τον γενικό αλγόριθμο για τη λύση του προβλήματος. Στο παρόν κεφάλαιο θα προσαρμόσουμε τον αλγόριθμο αυτό στον μετασχηματισμό Unrestricted Haar. Ο αλγόριθμος θα περιλαμβάνει δύο βασικά στάδια:

1. την προεπεξεργασία των δεδομένων, όπου καθορίζονται τα επιτρεπτά όρια της εισερχόμενης τιμής u.

2. την κύρια μέθοδο, που ακολουθεί πιστά τα βήματα του αλγορίθμου RangeHaar προσαρμόζοντας όπου χρειάζεται τον αλγόριθμο στα χαρακτηριστικά του μετασχηματισμού.

3.4.1 Προεπεξεργασία Δεδομένων

Οι εισερχόμενες τιμές στο δέντρο unrestricted Haar θα πρέπει να αλλάξουν για να προσαρμοστούν στο νέο κριτήριο βελτιστοποίησης που είναι η ελαχιστοποίηση σφάλματος για συναθροιστικά και όχι για σημειακά ερωτήματα. Στο παραδοσιακό δέντρο unrestricted Haar οι δυνατές τιμές της εισερχόμενης τιμής περιοριζόταν στο εύρος του ελαχίστου m_i και του μεγίστου M_i των δεδομένων που περιέχονταν στο υποδέντρο που ορίζει ένας συντελεστής C_i. Κατα αντιστοιχία με αυτήν την λογική πλέον η εισερχόμενη τιμή θα πρέπει να βρίσκεται ανάμεσα στο ελάχιστο και μέγιστο όλων των δυνατών αθροισμάτων που προκύπτουν από τα δεδομένα του υποδέντρου. Ο υπολογισμός των μέτρων αυτών μπορεί να γίνει αναδρομικά με την βοήθεια των θεωρημάτων που έχουμε παραθέσει στο κεφάλαιο 2 θεωρώντας ως διατεταγμένο σύνολο τα δεδομένα-φύλλα του δέντρου. Έτσι κάθε εσωτερικός κόμβος θα αντιστοιχίζεται με ένα μέγιστο και ελάχιστο των συναθροιστικών ερωτήσεων που βρίσκονται στην εμβέλεια του. Η διαδικασία αυτή αποτελεί μία προεπεξεργασία των δεδομένων που είναι απαραίτητη για να λειτουργήσει ο αλγόριθμος όπως περιγράφεται παρακάτω.

3.4.2 Αλγόριθμος RangeHaarUnrestricted

Στην ουσία αυτό που θα πρέπει να προσαρμόσουμε στο γενικό αλγόριθμο είναι ο τρόπος με τον οποίο εξετάζονται όλες οι δυνατές τιμές του κόμβου. Οπότε τα βασικά βήματα του αλγορίθμου θα επαναλαμβάνονται για κάθε δυνατή (εντός των επιτρεπόμενων ορίων) εισερχόμενη τιμή και για κάθε δυνατή τιμή του συντελεστή του κόμβου ξεχωριστά. Στη συνέχεια θα δημιουργείται βάσει του αλγορίθμου το σύνολο των περιλήψεων που επικρατούν

έναντι των υπολοίπων με βάση τα κριτήρια που έχουν οριστεί στην παράγραφο 2.3.2. Τέλος η εύρεση της λύσης γίνεται με την διαδικασία που έχει περιγραφτεί στην παράγραφο 3.2.

3.4.3 Μελέτη Πολυπλοκότητας Αλγορίθμου

Όπως έχει εξηγηθεί και στην Παράγραφο 3.2 , το πρόβλημα είναι εκθετικού χώρου και χρόνου. Επιλέγοντας σε κάθε επανάληψη του αλγορίθμου τις επικρατέστερες λύσεις και αφαιρώντας τις υπολειπόμενες, πετυχαίνουμε δραστική μειώση της πολυπλοκότητας του αλγορίθμου. Ωστόσο θεωρητικά ο αλγόριθμος παραμένει εκθετικός. Ουσιαστικά θα πρέπει να γίνουν πολλά πειράματα με ετερογένη δεδομένα κάθε φορά ώστε να αποφανθούμε για την μέση απόδοση του αλγορίθμου στην πράξη.

```
     Input: Array D[1..N] of data, space budget B
     Output: 2-dimensional Array E of all optimal synopses-1st dimension for the path S, 2nd dimension for the space
             budget b
 1   if index==0 then                                                                              /* root */
 2   |  return E
 3   endif
 4   else if index < N then
 5   |  forall possible values S do
 6   |  |  forall space budget b do
 7   |  |  |  forall possible values of c_i do
 8   |  |  |  |  forall elements of W(2i) do
 9   |  |  |  |  |  forall elements of W(2i+1) do
10   |  |  |  |  |  |  s[2i] = r[i] − ku + (k/2)c_i
11   |  |  |  |  |  |  s[2i+1] = r[i] − ku − (k/2)c_i
12   |  |  |  |  |  |  L_M[i] = max(L_M[2i], s[2i] + L_M[2i+1])
13   |  |  |  |  |  |  L_m[i] = min(L_m[2i], s[2i] + L_m[2i+1])
14   |  |  |  |  |  |  R_M[i] = max(R_M[2i] + s[2i+1], R_M[2i+1])
15   |  |  |  |  |  |  R_m[i] = min(R_m[2i] + s[2i+1], R_m[2i+1])
16   |  |  |  |  |  |  I[i] = max(I[2i], I[2i+1], |R_M[2i] + L_M[2i+1]| , |R_m[2i] + L_m[2i+1]|)
17   |  |  |  |  |  |  V = [I[i], L_M[i], L_m[i], R_M[i], R_m[i]]
18   |  |  |  |  |  |  forall elements V' of W(i) do
19   |  |  |  |  |  |  |  if dominate(V',V)==1 then                                    /* V' dominates V */
20   |  |  |  |  |  |  |  |  drop V
21   |  |  |  |  |  |  |  |  break
22   |  |  |  |  |  |  |  endif
23   |  |  |  |  |  |  |  else if dominate(V',V)==2 then                               /* V dominates V' */
24   |  |  |  |  |  |  |  |  drop V'
25   |  |  |  |  |  |  |  endif
26   |  |  |  |  |  |  |  keep V
27   |  |  |  |  |  |  endfall
28   |  |  |  |  |  endfall
29   |  |  |  |  endfall
30   |  |  |  endfall
31   |  |  endfall
32   |  endfall
33   endif
34   else                                                                                  /* last wavelet level */
35   |  forall possible values S do
36   |  |  forall space budget b do
37   |  |  |  forall possible values of c_i do
38   |  |  |  |  s[2i] = r[i] − ku + (k/2)c_i
39   |  |  |  |  s[2i+1] = r[i] − ku − (k/2)c_i
40   |  |  |  |  L_M[2i] = |s[2i] − d[2i]|
41   |  |  |  |  L_m[2i] = |s[2i] − d[2i]|
42   |  |  |  |  R_M[2i] = |s[2i] − d[2i]|
43   |  |  |  |  R_m[2i] = |s[2i] − d[2i]|
44   |  |  |  |  L_M[2i+1] = |s[2i+1] − d[2i]|
45   |  |  |  |  L_m[2i+1] = |s[2i+1] − d[2i]|
46   |  |  |  |  R_M[2i+1] = |s[2i+1] − d[2i]|
47   |  |  |  |  R_m[2i+1] = |s[2i+1] − d[2i]|
48   |  |  |  |  L_M[i] = max(L_M[2i], s[2i] + L_M[2i+1])
49   |  |  |  |  L_m[i] = min(L_m[2i], s[2i] + L_m[2i+1])
50   |  |  |  |  R_M[i] = max(R_M[2i] + s[2i+1], R_M[2i+1])
51   |  |  |  |  R_m[i] = min(R_m[2i] + s[2i+1], R_m[2i+1])
52   |  |  |  |  I[i] = max(I[2i], I[2i+1], |R_M[2i] + L_M[2i+1]| , |R_m[2i] + L_m[2i+1]|)
53   |  |  |  |  V = [I[i], L_M[i], L_m[i], R_M[i], R_m[i]]
54   |  |  |  |  forall elements V' of W(i) do
55   |  |  |  |  |  if dominate(V',V)==1 then                                          /* V' dominates V */
56   |  |  |  |  |  |  drop V
57   |  |  |  |  |  |  break
58   |  |  |  |  |  else if dominate(V',V)==2 then                                     /* V dominates V' */
59   |  |  |  |  |  |  drop V'
60   |  |  |  |  |  keep V
61   |  |  |  |  endfall
62   |  |  |  endfall
63   |  |  endfall
64   |  endfall
65
```

Σχήμα 3.1: HaarRange Algorithm

Κεφάλαιο 4

Επίλογος

4.1 Συνοπτικές Παρατηρήσεις

Συνόπτικα θα λέγαμε ότι αναλύσαμε και παρουσιάσαμε μία πρώτη προσέγγιση στο δύσκολο πρόβλημα κατασκευής βέλτιστης σύνοψης για συναθροιστικά σφάλματα. Καταφέραμε να λύσουμε το πρόβλημα με έναν αναδρομικό αλγόριθμο μερικής διάταξης που υπολογίζει τα μέτρα ορισμένων σφαλμάτων και τα βελτιστοποιεί με βάση το κριτήριο μερικής διάταξης. Επιπλέον παρείχαμε μία εμπεριστατωμένη θεωρητική μελέτη του προβλήματος και αποδείξαμε τις βασικές ιδιότητες στις οποίες στηρίχθηκε ο αλγόριθμος που προτείναμε.

4.2 Μελλοντική Εργασία

Ο γενικός αλγόριθμος που περιγράφεται στο Κεφάλαιο 3 μπορεί εύκολα να εξειδικευθεί στους επιμέρους μετασχηματισμούς κυματιδίων. Συμπερασματικά θα λέγαμε ότι λύσαμε πλήρως ένα δύσκολο πρόβλημα που δεν μπορεί να αντιμετωπιστεί με την κλασική τεχνική του δυναμικού προγραμματισμού προτείνοντας έναν αλγόριθμο μερικής διάταξης. Ωστόσο θεωρητικά η πολυπλοκότητα παραμένει εκθετική. Σαν μελλοντική εργασία αναφέρουμε την ανάγκη πειραματικής μελέτης ώστε να μετρηθεί η μέση απόδοση του αλγόριθμου αφού το κριτήριο μερικής διάταξης πρακτικά μειώνει το χώρο αναζήτησης κάθε κόμβου καθώς και την κατασκευή ενός πιο πρακτικού (με χαμηλότερο υπολογιστικό κόστος) ευριστικού αλγορίθμου.

Βιβλιογραφία

[1] Antonios Deligiannakis, Minos Garofalakis και Nick Roussopoulos. Extended wavelets for multiple measures. *ACM Trans. Database Syst.*, 32(2), 2007.

[2] Minos Garofalakis και Amit Kumar. Deterministic wavelet thresholding for maximum-error metrics. Στο *Proceedings ACM Principles of Database Systems (PODS)*, σελίδες 166–176, 2004.

[3] Sudipto Guha. Space efficiency in synopsis construction algorithms. Στο *VLDB '05: Proceedings of the 31st international conference on Very large data bases*, σελίδες 409–420. VLDB Endowment, 2005.

[4] Sudipto Guha και Boulos Harb. Wavelet synopsis for data streams: minimizing non-euclidean error. Στο *KDD '05: Proceeding of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, σελίδες 88–97, New York, NY, USA, 2005. ACM Press.

[5] Sudipto Guha και Boulos Harb. Approximation algorithms for wavelet transform coding of data streams. Στο *Proceedings ACM-SIAM Symposium on Discrete Algorithms (SODA)*, 2006.

[6] Panagiotis Karras και Nikos Mamoulis. The haar+ tree: A refined synopsis data structure. Στο *ICDE*, 2007.

[7] Panagiotis Karras, Dimitris Sacharidis και Nikos Mamoulis. Exploiting duality in summarization with deterministic guarantees. Στο *KDD*, 2007.