3η Εργασία Εξόρυξης Δεδομένων

## Περιγραφή

Η επιστημονική δημοσίευση **MMDT: a multi-valued and multi-labeled decision tree classifier for data mining** προτείνει μια εξέλιξη του αλγορίθμου MMC, τον MMDT ώστε να επιτευχθεί μεγαλύτερη ακρίβεια ταξινόμησης σε δεδομένα που έχουν πολλές τιμές στις μεταβλητές του αλλά και ανήκουν σε πολλές κλάσεις. Το δέντρο απόφασης MMC αρχικά δημιουργήθηκε για να μπορεί να διαχειρίζεται και να ταξινομεί τέτοιου είδους δεδομένα. Για να πετύχει καλύτερη ομαδοποίηση και να χωρίσει τον κόμβο σε κόμβους-παιδιά, το MMC προτείνει ένα νέο μέτρο, την σταθμισμένη ομοιότητα (weighted similarity). Διαλέγει τις ετικέτες με τα μεγαλύτερα νούμερα για να αντιπροσωπεύουν τα φύλλα του δέντρου. Το δέντρο εξετάζεται όπως όλα τα δέντρα απόφασης αλλά αν η εξέταση καταλήγει σε πολλά φύλλα, το MMC ενώνει τα ετικέτες των αποτελεσμάτων για το αποτέλεσμα της ταξινόμησης.

Για να ελαχιστοποιηθεί η υπερπροσαρμογή του MMC, δημιουργείται ένα όριο δεδομένων το οποίο θα πρέπει να ξεπερνάνε σαν πλήθος όταν ορίζονται σε έναν κόμβο, και να μην εκλαμβάνεται μόνο η μέση ομοιότητα, αλλά και η μέση καταλληλότητα των ετικετών σε κάθε κόμβο-παιδί. Πάνω στα βήματα αυτά βασίστηκε ο MMDT.

Ο αλγόριθμος που βασίζεται ο MMC αρχικά χωρίζει τα δεδομένα με βάση ένα γνώρισμα για να καθορίσει τους κόμβους και τα κλαδιά του δέντρου. Ο διαχωρισμός αυτός αξιολογείται με βάση την σταθμισμένη ομοιότητα που υπολογίζεται από το άθροισμα της ομοιότητας μεταξύ των σετ ετικετών επί τον αριθμό των δεδομένων σε έναν κόμβο ανά τον συνολικό αριθμό δεδομένων. Ύστερα ορίζει τους τελικούς κόμβους (φύλλα) με βάση το ποσοστό των δεδομένων του κόμβου που χαρακτηρίζονται από μια ετικέτα. Εάν το ποσοστό της συγκεκριμένης ετικέτας είναι μεγαλύτερο από μια προκαθορισμένη τιμή (minsup), τότε το ονομάζουμε «μεγάλη» ετικέτα, αλλιώς «μικρή». Υπολογίζεται η διαφορά της μικρότερης τιμής από τις «μεγάλες» ετικέτες με την μεγαλύτερη τιμή των «μικρών» ετικετών και αν η τιμής της ξεπερνάει μια προκαθορισμένη τιμή (mindiff) τότε έχουμε έναν «καθαρό» κόμβο. Ένας κόμβος γίνεται τελικός α) αν είναι «καθαρός», και έχει ως σετ ετικετών όλες τις μεγάλες ετικέτες, β) αν δεν είναι καθαρός και οι τιμές του έχουν χρησιμοποιηθεί σε όλη την διαδρομή από την ρίζα έως αυτόν, αν έχει «μεγάλες» ετικέτες τότε θα τα έχει ως σετ ετικετών, αλλιώς θα έχει τη μεγαλύτερη ετικέτα από τις «μικρές» ετικέτες, γ) αν ο κόμβος δεν είναι καθαρός και τα δεδομένα είναι μικρότερα από μια προκαθορισμένη τιμή (minqty) τότε τηρούνται τα βήματα από το «β». Τηρώντας τα βήματα του MMC, ο MMDT χρησιμοποιεί αναλογία ετικετών με συντελεστή βαρύτητας για να επιλέγει τις μεταβλητές με βάση τις οποίες θα διαχωρίζει τα δεδομένα σε καλύτερες αναπαραστάσεις. Βασίζεται στην τιμή της ετικέτας, που δίνεται από τον βαθμό της εκπροσώπησης του κόμβου από αυτή και υπολογίζεται με την ομοιότητα των ετικετών επί το κομμάτι όλων των όμοιων ετικετών. Η ομοιότητα υπολογίζεται με βάση δυο στρατηγικές: α) με βάση την εμφάνιση, που βασίζεται στο πλήθος των ετικετών ίδιων κλάσεων μεταξύ δύο σετ ετικετών προς το πλήθος των ετικετών που υπάρχουν μόνο σε ένα από τα δύο σετ, β) με βάση την ομοιότητα όπως υπολογιζόταν και στον MMC. Η πρόβλεψη του MMDT γίνεται καθώς διασχίζεται το δέντρο από την ρίζα έως να καταλήξει σε ένα «φύλλο». Στην περίπτωση που βρεθούν περισσότερα «φύλλα» ως αποτελέσματα, ο αλγόριθμος τα ενώνει και παρουσιάζει μια πρόβλεψη πολλαπλών ετικετών.

Συγκρίθηκαν οι δύο αλγόριθμοι με σύνολο δεδομένων από πελάτες τουριστικής εταιρίας μεγέθους 5000 εγγραφών. Η σύγκρισή τους έγινε με διαφορετικές τιμές στις παραμέτρους μεγέθους δεδομένων εκπαίδευσης, ελάχιστη τιμή minsup, ελάχιστη διαφορά mindiff, ελάχιστος αριθμός δεδομένων στον κόμβο minqty και τιμή μεγέθους του δέντρου ub. Ο MMDT αλγόριθμος που βασιζόταν στον αριθμό εμφάνισης των ετικετών, και ο MMDT αλγόριθμος που βασιζόταν στην ομοιότητα των ετικετών, παρουσίαζαν καλύτερη ακρίβεια από τον αλγόριθμο MMC (64.20% , 63.86% και 62.56% αντίστοιχα), ενώ ο MMC δημιουργούσε μικρότερο δέντρο.

Κατά την αλλαγή των τιμών που προαναφέρθηκαν παρατηρήθηκε ότι ο MMDT όταν το μέγεθος του συνόλου δεδομένων ξεπερνούσε ένα συγκεκριμένο κατώφλι, η ακρίβεια μειωνόταν γιατί αυξανόταν και ο θόρυβος έναντι των πληροφοριών στα δεδομένα. Με την αύξηση της τιμής minsup αυξανόταν η ακρίβεια αλλά οι κανόνες μειωνόντουσαν γιατί οι ετικέτες σε κάθε «φύλλο» έχουν μεγαλύτερη ομοιότητα και αναλογία, αλλά αυξάνεται επίσης η πιθανότητα ένας κόμβος να σταματήσει να αυξάνεται. Η αύξηση της τιμής της διαφοράς για τον διαχωρισμό των «καθαρών» και «μη καθαρών» κόμβων οδηγούσε στην αύξηση της ακρίβειας πρόβλεψης και των κανόνων καθώς τα όρια μεγαλώνουν και έχουμε πιο ακριβή διαχωρισμό μεταξύ «μεγάλων» και «μικρών» ετικετών που οδηγεί στην αύξηση της ομοιότητας των ετικετών σε ένα «φύλλο». Ακόμα, με την αύξηση του αριθμού δεδομένων σε έναν κόμβο minqty παρατηρείται μειώνονται οι κανόνες του δέντρου γιατί μπορεί τα δεδομένα να είναι λιγότερα από το όριο και ένας κόμβος να σταματήσει να μεγαλώνει νωρίτερα. Τέλος με την αύξηση του ορίου μεγέθους του δέντρου ub παρατηρείται τάση μείωσης του αριθμού κανόνων διότι μπορεί ένας κόμβος με πολλά κλαδιά να έχει μικρότερη ποσότητα δεδομένων από έναν κόμβο με λιγότερα κλαδιά αυξάνοντας την τάση για δημιουργία μικρότερων δέντρων και για μείωση των κανόνων.