Αριστοτέλειο Πανεπιστήμιο Θεσσαλονίκης
Πολυτεχνική Σχολή
Τμήμα Ηλεκτρολόγων Μηχανικών &
Μηχανικών Υπολογιστών
Τομέας Ηλεκτρονικής και Υπολογιστών
Εργαστήριο Επεξεργασίας Πληροφορίας και Υπολογισμών

Διπλωματική Εργασία

Πολυκατηγορική Ταξινόμηση με Μανθάνοντα Συστήματα Ταξινομητών

Εκπόνηση:
Αλέξανδρος Χ. Φιλοθέου
ΑΕΜ: 6209

Επίβλεψη:
Καθ. Περικλής Α. Μήτκας
Δρ. Φανή Α. Τζήμα

Θεσσαλονίκη, Ιούλιος 2013

We shall not cease from exploration
And the end of all our exploring
Will be to arrive where we started
And know the place for the first time.
T.S. Eliot

It took me four years to paint like Raphael,
but a lifetime to paint like a child.
Pablo Picasso

Ευχαριστιες

Η παρούσα διπλωματική εργασία δε θα ήταν δυνατό να πραγματοποιηθεί χωρίς την εμπιστοσύνη που έδειξε στο πρόσωπό μου ο καθηγητής Περικλής Μήτκας και χωρίς την πολύτιμη συνεισφορά και υποστήριξη της επιβλέπουσάς μου, Δρ. Φανής Τζήμα.

Αποτελεί το επιστέγασμα μίας πορείας οκτώ ετών, στον καθορισμό της οποίας συνεισέφεραν κάποιοι άνθρωποι τους οποίους ευγνωμονώ. Καταρχάς, θέλω να ευχαριστήσω τον Ανδρέα Συμεωνίδη, τη Θωμαή Σταθοπούλου, τον Αλέξανδρο Δημάρατο, το Βαγγέλη Μέμμο, τις Ηλιάνα και Κορίνα Παππή, τον Ιωάννη Παπαδόπουλο, το Μάνο Τσαρδούλια, το Χαράλαμπο Σερένη και τον Ηλία Τσιγγενόπουλο. Χρωστώ ευχαριστίες στους φίλους μου, Δημήτρη Μοσχοχωριάτη, Αστέρη Πλιάτσικα, Ανδρέα Καργάκο, Άριστείδη Γκουντάρα, Δημήτρη Κοκίδη και στην καθηγήτρια και φίλη πλέον, Μαρία Λύμπερ.

Το μεγαλύτερο ευχαριστώ και τον απεριόριστο σεβασμό μου οφείλω στους γονείς μου, Μαρίνα και Χρήστο, για όλα τα δύσκολα και εύκολα χρόνια, για όλες τις θυσίες, για το γεγονός ότι μου επέτρεψαν να γίνω ό,τι μπορούσα να γίνω και για ό,τι ένα παιδί μπορεί να οφείλει σε γονείς που το αγαπάνε άνευ όρων.

Η παρούσα διπλωματική αφιερώνεται σε όλους τους παραπάνω και στο φίλο μας Δημήτρη "jimi" Μπαλή, ο οποίος έφυγε από κοντά μας τον Αύγουστο του 2010.

Περίληψη

Με τον αυξανόμενο ρυθμό παραγωγής δεδομένων και πληροφοριών παγκοσμίως και τη στροφή της ανθρωπότητας προς την αυτοματοποίηση όλο και περισσότερων διαδικασιών της σύγχρονης ζωής, τις τελευταίες δεκαετίες έχει υπάρξει αυξημένο ενδιαφέρον για τη Μηχανική Μάθηση, έναν τομέα της Υπολογιστικής Νοημοσύνης. Ο τομέας αυτός ασχολείται με την ανάπτυξη μηχανών που μπορούν να μαθαίνουν από την εμπειρία, ώστε να αναλάβουν αυτές το γιγάντιο έργο της αυτοματοποίησης διεργασιών και δραστηριοτήτων - ένα έργο που πλέον καμία ανθρώπινη μονάδα δεν μπορεί να φέρει εις πέρας. Η αυτοματοποίηση αυτή αφορά στην πρόβλεψη, εξήγηση ή/και κατανόηση των υποκείμενων δεδομένων. Πολλά προβλήματα μπορούν να περιγραφούν από ένα σύνολο δεδομένων που συλλέχθηκαν κάποια συγκεκριμένη στιγμή και, έτσι, μία πληθώρα μεθόδων Μηχανικής Μάθησης μπορεί να βοηθήσει στην εξαγωγή εύκολα ερμηνεύσιμων μοντέλων, διαφόρων μέσων, όπως είναι οι κανόνες ή τα δένδρα αποφάσεων. Ωστόσο, ειδικά στην περίπτωση της πρόβλεψης, υπάρχουν ειδικές περιστάσεις, όπως όταν το πρόβλημα απαιτεί αλληλεπίδραση με κάποια άλλη οντότητα, όπου ο αριθμός των επιλογών μειώνεται και τα Μανθάνοντα Συστήματα Ταξινομητών γίνονται η καλύτερη, αν όχι η μόνη, λύση.

Τα Μανθάνοντα Συστήματα Ταξινομητών (ΜαΣΤ) ανήκουν σε μία κλάση συστημάτων Μηχανικής Μάθησης Βασισμένης στη Γενετική (ΜΜΒΓ), τα οποία είναι σχεδιασμένα ώστε να μπορούν να αντιμετωπίσουν τόσο σειριακά, όσο και ενός-βήματος προβλήματα απόφασης, χρησιμοποιώντας κανόνες ταξινόμησης. Η παρούσα εργασία εστιάζει σε προβλήματα ταξινόμησης και, πιο συγκεκριμένα, χρησιμοποιεί ΜαΣΤ τύπου Michigan για να αντιμετωπίσει προβλήματα πολυκατηγορικής φύσης.

Η πολυκατηγορική ταξινόμηση είναι μία διαδικασία Εξόρυξης Δεδομένων όπου κάθε δείγμα ενός συνόλου δεδομένων συσχετίζεται με περισσότερες από μία κατηγορίες που ονομάζονται ετικέτες. Πολυκατηγορικά δεδομένα εμφανίζονται σε αφθονία σε πραγματικά προβλήματα, όπως οι ιατρικές διαγνώσεις, οι κατηγοριοποιήσεις εγγράφων ή η συσχέτιση γονιδίων με βιολογικές λειτουργίες.

Η παρούσα εργασία βασίζεται και επεκτείνει την προσέγγιση της πολυκατηγορικής ταξινόμησης του αλγορίθμου GMl-ASLCS, ο οποίος με τη σειρά του επεκτείνει το πλαίσιο εποπτευόμενης μάθησης AS-LCS στον πολυκατηγορικό χώρο. Εδώ πρέπει να σημειώσουμε πως, από όσο γνωρίζουμε, η προσέγγιση της πολυκατηγορικής ταξινόμησης με χρήση ΜαΣΤ είναι από τις πρώτες στον αντίστοιχο χώρο.

Βασιζόμενοι στις παραπάνω μεθόδους, η προσέγγισή μας κινείται σε τρεις άξονες: i) την εμβάθυνση στις λειτουργίες ενός (πολυκατηγορικού) ΜαΣΤ, μελετώντας και αναλύοντας τις εσωτερικές του διαδικασίες, ii) την προσέγγιση του υπό μελέτη προβλήματος περισσότερο από τη σκοπιά του μηχανικού και λιγότερο από αυτή της Επιστήμης Υπολογιστών (με την έννοια ότι μελετούμε ευρύτερα τη συμπεριφορά διαφορετικών τμημάτων ενός ΜαΣΤ και των μεταβολών που επιφέρουν στη συμπεριφορά του οι επιμέρους μεταβολές των παραμέτρων που τη διέπουν), και iii) τη βελτίωση της συνολικής συμπεριφοράς του GMl-ASLCS βάσει των δύο παραπάνω αξόνων, τόσο ως προς τις μετρικές αξιολόγησης που χρησιμοποιούνται, όσο και ως προς τη συμπεριφορά των επιμέρους τμημάτων του.

Η προσήλωση μας στους τρεις παραπάνω άξονες έχει ως αποτέλεσμα την εξαγωγή μίας σειράς από παρατηρήσεις και την επινόηση και υιοθέτηση διορθωτικών και διαρθρωτικών δράσεων:

1. Εμβαθύνοντας στο Γενετικό Αλγόριθμο που χρησιμοποιούν τα ΜαΣΤ, προτείνουμε την υιοθέτηση ενός νέου τελεστή Διασταύρωσης, τον τελεστή Διασταύρωσης Δύο Τμημάτων, ο οποίος προσιδιάζει στη φύση των πολυκατηγορικών προβλημάτων

2. Για τη διεύρυνση του αριθμού των δειγμάτων ενός συνόλου δεδομένων που μπορούν να ταξινομήσουν οι κανόνες του ΜαΣΤ με ακρίβεια, προτείνουμε την εισαγωγή ενός νέου τμήματος διαγραφής κανόνων που εφαρμόζεται σε επιμέρους σύνολα κανόνων αντί για το σύνολο του πληθυσμού τους.

3. Αναλύοντας εσωτερικά τη συμπεριφορά του πρώτου GMl-ASLCS, ανακαλύπτουμε τις σοβαρές συνέπειες της διατήρησης κανόνων στον πληθυσμό του ΜαΣΤ οι οποίοι είναι ανίκανοι να ταξινομήσουν δείγματα που τους παρουσιάζονται και προτείνουμε την εφαρμογή μίας μεθοδολογίας που τις εξαλείφει.

4. Παρατηρούμε την αντίρροπη δράση της υπερ-συσσώρευσης μη σαφών αποφάσεων για ετικέτες στο τμήμα απόφασης των κανόνων και υιοθετούμε μία προσέγγιση που μειώνει σε λογικό βαθμό αυτές τις αποφάσεις.

Τέλος, κάνουμε παρατηρήσεις πάνω στις διαφορετικές λειτουργίες ενός ΜαΣΤ (που μπορούν να χρησιμοποιηθούν για την εξαγωγή συμπερασμάτων, τόσο όσον αφορά σε πολυκατηγορικά όσο και σε μονοκατηγορικά ΜαΣΤ) και προτείνουμε μερικές τροποποιήσεις στον πλήρη ορισμό του GMl-ASLCS, με σκοπό την αύξηση των επιδόσεών του, όπως για παράδειγμα την αρχικοποίηση του πληθυσμού χρησιμοποιώντας την ομαδοποίηση των δειγμάτων του πολυκατηγορικού συνόλου με το οποίο εκπαιδεύεται ο GMl-ASLCS.

Μία αρχική αξιολόγηση του ανανεωμένου GMl-ASLCS πραγματοποιείται σε τρία πολυκατηγορικά τεχνητά προβλήματα, σχεδιασμένα ώστε να δοκιμάσουν τις επιδόσεις του σε διαφορετικά περιβάλλοντα, λιγότερο πολύπλοκα από ότι τα πραγματικά σύνολα δεδομένων. Όσον αφορά στα πραγματικά σύνολα πολυκατηγορικών δεδομένων, οι επιδόσεις του GMl-ASLCS συγκρίνονται με αυτές του πρωταρχικού GMl-ASLCS και των διαδεδομένων μεθόδων πολυκατηγορικής ταξινόμησης RAkEL-J48, MlkNN και BR-J48. Σύμφωνα με τα αποτελέσματα της πειραματικής διαδικασίας, η παρούσα έκδοση του GMl-ASLCS κατατάσσεται πρώτη ανάμεσά τους, επιδεικνύοντας, επιπλέον, στατιστικά σημαντικές διαφορές σε σχέση με τον προκάτοχό του.

Τέλος, εξετάζουμε μεμονωμένα την επίδραση που είχαν οι τέσσερις λειτουργίες που μεταβάλαμε στην αρχική έκδοση του GMl-ASLCS στα τεχνητά πολυκατηγορικά προβλήματα που προαναφέραμε και καταγράφουμε τις επιδόσεις των τροποποιημένων εκδόσεων GMl-ASLCS στα πραγματικά σύνολα δεδομένων που χρησιμοποιήσαμε. Από αυτές, ξεχωρίζουμε μία έκδοση του GMl-ASLCS που χρησιμοποιεί ομαδοποίηση για την αρχικοποίηση του πληθυσμού των κανόνων του και μία που δεν λαμβάνει υπόψη τις μη σαφείς αποφάσεις κανόνων στον υπολογισμό της καταλληλότητάς τους.

Συνολικά, η παρούσα εργασία έχει ως αποτέλεσμα τη βελτίωση της συμπεριφοράς επιμέρους τμημάτων του αλγορίθμου GMl-ASLCS, αλλά και της συνολικής συμπεριφοράς και επίδοσής του, όσον αφορά στις μετρικές που χρησιμοποιούνται για την αξιολόγηση του μοντέλου που αναπτύσσει και την αύξηση του αριθμού των δειγμάτων που μπορεί να ταξινομήσει με ακρίβεια.

Το κείμενο της παρούσας εργασίας βρίσκεται αναρτημένο στο διαδίκτυο στον παρακάτω σύνδεσμο: https://github.com/li9i/auth_thesis. Η κωδικοποίηση του αναφερθέντων πλαισίων και οι προσωμοιώσεις της συμπεριφοράς τους έγιναν σε JAVA και είναι αναρτημένες στον παρακάτω σύνδεσμο: https://github.com/li9i/mlaslcs.

Title

Multi-label Classification with Learning Classifier Systems

Abstract

Due to the rising growth of data production worldwide and the turn of mankind to process automation, in the last decades there has been a rising interest in Machine Learning, a branch of Computational Intelligence that deals with the construction and study of machines that can learn from experience, so as to tackle the immense task of automation - a task that can be matched by no human being. Said automation takes the form of prediction, explanation and/or comprehension of the underlying data of a target problem. If the problem at hand can be described by a set of data collected at some point in time, there is a plethora of machine learning techniques that can induce easily comprehensible models that employ various means of classification, like classification rules or decision trees. However, there are a number of occasions, for example when the problem presupposes interaction with other (external) entities, that impose restrictions on the number and sort of applicable methods, making Learning Classifier Systems the best (if not only) option.

Learning Classifier Systems (LCS) belong to a class of Genetics-Based Machine Learning (GBML) systems, designed to work for both sequential and single-step problems, using classification rules. The present Diploma Thesis focuses on classification problems and, more specifically, uses the LCS framework to tackle multi-label classification problems.

Multi-label classification is a Data Mining task in which a data instance is assigned multiple target labels. Multi-label data are very common in real world problems, such as medical diagnosis, document categorization and gene association with biological functions.

The current Diploma Thesis is based on and extends the multi-label Michigan LCS, GMl-ASLCS, which in turn extends the supervised learning scheme of AS-LCS on the realm of multiple labels. It is important to note that, to our knowledge, this multi-label approach with LCS is one of the first in the field.

Based on the aforementioned frame of reference, our approach moves on the tracks of: i) gaining insight into the operations of a (multi-label) LCS, by studying and analyzing its internal functions, ii) approaching the problem and the studied LCS through an engineering viewpoint rather than that of a Computer Science one (in the sense of studying the broader behaviour of the different components that a LCS consists of and the changes in its behaviour brought by modifying its individual running parameters) and, iii) the overall improvement of GMl-ASLCS's behaviour, in light of the above statements, with respect to the evaluation metrics employed and the behaviour of its individual components.

Following this approach results in a series of remarks and the invention and adoption of a number of correctional and structural actions.

1. By delving deeper into the function of the Genetic Algorithm that is part of an LCS, we propose the adoption of a new crossover operator, the Two Segment Crossover operator, that pertains to the multi-label nature of the classification problem.

2. We introduce a new deletion mechanism, that is applied on individual rule sets rather than the population, for the purpose of augmenting the number of instances that an LCS can accurately classify.

3. By analyzing the internal behaviour of the initial GMl-ASLCS, we discover the grave repercussions of preserving rules that are unable of classifying even a single instance and we suggest the adoption of a mechanism that eliminates them.

4. We also observe the impact of the overaccumulation of non-explicit decisions about labels in the rules' consequent part and adopt a mechanism for mitigating it.

Finally, we study and remark on the different functions that an LCS employes (that can be used to deduce valuable conclusions on multi-label and single-label LCS) and we suggest a number of variations in the definition of GMl-ASLCS for increasing its performance. These variations include a population initialization component by means of clustering the instances of the data-set that GMl-ASLCS is being trained with.

A preliminary evaluation of the final version of GMl-ASLCS is performed on three multi-label testbed problems, designed to challenge GMl-ASLCS's perormance in environments that are less complex than real-world multi-label data-sets/problems. Regarding these real-world multi-label problems, we test GMl-ASLCS on six of them and GMl-ASLCS's performance is compared with that of the initial GMl-ASLCS and three state-of-the-art algorithms used in multi-label classification, namely RAkEL-J48, MlkNN and BR-J48. The results show that GMl-ASLCS ranks first among them, revealing that there is no statistically significant performance differences between the current version of GMl-ASLCS and its three rival non-LCS-based methods. In contrast, the performance of the current version of GMl-ASLCS exhibits a statistically significant performance difference with respect to the initial GMl-ASLCS.

Finally, we examine the individual impact of the four major changes we introduced to the initial GMl-ASLCS framework, on the aforementioned multi-label test-beds and evaluate the performance of the above variations of GMl-ASLCS on the same set of real-world problems. Among these variations, we distinguish one that uses clustering to initialize the LCS's population and one that "discards" non-explicit decisions of rules about labels, by not allowing them to have any effect on the fitness calculation method.

Overall, the present thesis results in an improvement of the behaviour of the individual components that GMl-ASLCS consists of and of the overall behaviour of GMl-ASLCS itself. These improvements vary and concern the accuracy of the model that GMl-ASLCS builds and the increase of the number of instances GMl-ASLCS can accurately classify.

Alexandros Philotheou
Intelligent Systems & Software Engineering Lab,
Electrical & Computer Engineering Department,
Aristotle University of Thessaloniki, Greece
July 2013

Περιεχόμενα

Κατάλογος σχημάτων

Κατάλογος πινάκων

Κατάλογος Αλγορίθμων

1

Εισαγωγή

Η κατηγοριοποίηση και οργάνωση διαφορετικών οντοτήτων, αν και υπήρχε από την εποχή του Αριστοτέλη, τον τελευταίο αιώνα έχει γίνει ανάγκη της καθημερινότητας και ένα εργαλείο της. Η ραγδαία σύγκλιση των επιστημονικών εξελίξεων στις περιοχές της πληροφορικής και των επικοινωνιών, όμως, στις τελευταίες δεκαετίες, έχει οδηγήσει στην κατακόρυφη αύξηση των διαθέσιμων και εύκολα προσβάσιμων δεδομένων, λόγω και της μείωσης του κόστους αποθήκευσής τους. Αυτό είχε ως αποτέλεσμα τη συσσώρευση δεδομένων που, λόγου του όγκου τους, δεν είναι πλέον εύκολα διαχειρίσιμα από τους ανθρώπους. Παράλληλα, η αύξηση της επεξεργαστικής ισχύος των υπολογιστών οδήγησε στην ανάπτυξη προσεγγιστικών, αλλά ταυτόχρονα μεγάλης ακρίβειας, τρόπων επίλυσης προβλημάτων από αυτούς, μεταφέροντας το βάρος από τον Άνθρωπο στη Μηχανή.

Το έργο της κατηγοριοποίησης, συνεπώς, ήταν φυσιολογικό να γίνει αντικείμενο ενδελεχούς έρευνας, δίνοντας ώθηση για την ανάπτυξη του κλάδου της Μηχανικής Μάθησης, τομέας του οποίου είναι η Εξόρυξη Δεδομένων. Η Εξόρυξη Δεδομένων, έχει ως αντικείμενο την εξαγωγή έμμεσης, προηγουμένως άγνωστης και ενδεχομένως χρήσιμης πληροφορίας από δεδομένα, με σκοπό τη δημιουργία υπολογιστικών προγραμμάτων τα οποία διατρέχουν αυτόματα τα περιεχόμενα κάθε είδους βάσεων δεδομένων, αναζητώντας κανονικότητες ή πρότυπα. Τα πρότυπα αυτά, εφόσον ανακαλυφθούν, είναι δυνατόν να γενικευτούν, ώστε να χρησιμοποιηθούν για την πρόβλεψη μελλοντικών καταστάσεων. Η εξαγωγή ισχυρών και γενικεύσιμων προτύπων είναι μία αναπόφευκτα ανακριβής διαδικασία: τα ανακαλυπτόμενα πρότυπα μπορεί να είναι κοινότυπα ή άνευ ενδιαφέροντος, είτε να εξαρτώνται από τυχαίες συμπτώσεις, εγγενείς στα χρησιμοποιούμενα σύνολα δεδομένων. Επιπλέον, εφόσον τα πραγματικά δεδομένα είναι από τη φύση τους ατελή, περιέχοντας συχνά τμήματα που είναι αλλοιωμένα ή ελλειπή, οποιαδήποτε κανονικότητα ανακαλύπτεται σε αυτά έχει περιορισμένη ακρίβεια. Σε κάθε περίπτωση, η Μηχανική Μάθηση έχει

ως απώτερο στόχο την επίλυση προβλημάτων που είναι σε τέτοιο βαθμό σύνθετα ώστε δεν είναι επιλύσιμα από τον Άνθρωπο ή, τουλάχιστον την παροχή οδηγιών για την επίλυσή τους [WF05].

Η αναζήτηση και προσέγγιση του βέλτιστου μοντέλου που αναπτύσσουν οι αλγόριθμοι μηχανικής μάθησης είναι η βασική προϋπόθεση κάθε μορφής υπολογιστικής νοημοσύνης και, συνεπώς, για την αποτελεσματικότερη επίλυση προβλημάτων, η αναζήτηση θα πρέπει να είναι καθολική και να μην περιορίζεται σε τοπικά βέλτιστες λύσεις. Ταυτόχρονα, μία καθολική αναζήτηση όλων των πιθανών καταστάσεων και λύσεων ενδέχεται να είναι απαγορευτική από άποψη χρόνου, ανάλογα με το μέγεθος του προβλήματος.

Επειδή το πρόβλημα της καθολικής αναζήτησης, λοιπόν, είναι δύσκολο και, επιπρόσθετα, σπάνια υπάρχουν σαφείς μαθηματικές λύσεις στα πραγματικά προβλήματα, οι μηχανικοί και οι επιστήμονες της πληροφορικής στράφηκαν προς τη Φύση για έμπνευση, ακολουθώντας την προσέγγιση της μίμησης της εξέλιξης των ειδών. Τα βιολογικά όντα δρουν στο (και σε σχέση με το) περιβάλλον τους, δοκιμαζόμενα σε διάφορες συνθήκες, ενώ τα γονίδιά τους διασταυρώνονται και μεταλλάσσονται αδιάκοπα στο πέρασμα του χρόνου. Παράλληλα, τα γονίδια που καθιστούν τα επιμέρους άτομα ενός είδους κατάλληλα για επιβίωση στο περιβάλλον τους διατηρούνται και κληροδοτούνται στις μελλοντικές γενιές, με μεγαλύτερη πίεση σε σχέση με αυτά που καθιστούν την επιβίωση (με την ευρεία έννοια) δυσχερέστερη, ακριβώς λόγω των μεγαλύτερων δυσκολιών των ατόμων που τα κατέχουν να επιβιώσουν και, συνεπώς, να αναπαραχθούν και να τα μεταλαμπαδεύσουν στους απογόνους τους. Το παραπάνω τμήμα της θεωρίας της εξέλιξης των ειδών που πρότεινε ο Δαρβίνος αποτελεί αναπόσπαστο κομμάτι του πλαισίου λειτουργίας των Γενετικών Αλγορίθμων (Genetic Algorithms).

Όπως και η θεωρία της εξέλιξης των ειδών, έτσι και οι Γενετικοί Αλγόριθμοι θεωρούν έναν πληθυσμό από άτομα-μοντέλα, η καταλληλότητα των οποίων εξαρτάται από την ικανότητα επιβίωσής τους στο περιβάλλον ενός προβλήματος, σε αντιστοιχία με το περιβάλλον των βιολογικών ειδών. Στους Γενετικούς Αλγορίθμους, τα μοντέλα αυτά τοποθετούνται στο εικονικό περιβάλλον-οικοσύστημα του προβλήματος, όπου εξελίσσονται με τον ίδιο τρόπο που εξελίσσονται οι οργανισμοί στη φύση. Ένα τέτοιο περιβάλλον είναι τα Μανθάνοντα Συστήματα Ταξινομητών (ΜαΣΤ), τα οποία δημιουργούν έναν πληθυσμό από κανόνες ταξινόμησης (άτομα), που εξελίσσονται μέσα σε ένα δεδομένο πλαίσιο μάθησης, με βάση τα χαρακτηριστικά του προβλήματος-στόχου προς επίλυση. Τα ΜαΣΤ χρησιμοποιούν κανόνες ταξινόμησης για την παραγωγή εύληπτων και κατανοητών λύσεων, από ειδικούς ή/και υπεύθυνους λήψης αποφάσεων, ακόμα και χωρίς κάποια εκπαίδευση στους υπολογιστές.

Βασισμένη στα παραπάνω, η παρούσα εργασία ασχολείται με ζητήματα Εξόρυξης Δεδομένων και το συνδυασμό μεθόδων Μηχανικής Μάθησης για τη δημιουργία προβλεπτικών μοντέλων. Πιο συγκεκριμένα, χρησιμοποιεί Μανθάνοντα Συστήματα Ταξινομητών, τα οποία ενσωματώνουν τους Γενετικούς Αλγορίθμους, για την καθολική αναζήτηση κανόνων ταξινόμησης δειγμάτων με βάση σύνολα δεδομένων και την κατηγοριοποίηση δεδομένων (δηλαδή των δειγμάτων) σε περισσότερες από μία κατηγορίες, οι οποίες ονομάζονται ετικέτες, ώστε να υλοποιήσει πολυκατηγορική ταξινόμηση.

Το φάσμα των εφαρμογών της πολυκατηγορικής ταξινόμησης σε πραγματικά προβλήματα είναι ευρύ, και αυτές κυμαίνονται από την κατηγοριοποίηση εγγράφων και πολυμεσικών οντοτήτων έως τη συσχέτιση βιολογικών λειτουργιών σε γονίδια και τη διάγνωση ιατρικών προβλημάτων.

Η χρήση ΜαΣΤ τύπου Michigan για την επίλυση πολυκατηγορικών προβλημάτων επιλέχθηκε λόγω των πλεονεκτημάτων τους έναντι των εναλλακτικών μεθόδων ταξινόμησης. Τα ΜαΣΤ δημιουργούν και μεταχειρίζονται κανόνες, των οποίων η μορφή είναι αντιληπτή και κατανοητή από τον άνθρωπο, αλλά και η ικανότητα γενίκευσής τους οδηγεί σε συμπαγείς χαρτογραφήσεις του χώρου γνωρισμάτων σε σχέση με το χώρο των κατηγοριών/ετικετών. Παράλληλα, τα ΜαΣΤ διαθέτουν μία ευέλικτη αναπαράσταση γνώσης, η οποία μπορεί να προσαρμοστεί εύκολα για την αντιμετώπιση νέων τύπων δεδομένων, ενώ κατασκευάζουν τα μοντέλα τους online, γεγονός που είναι κρίσιμης σημασίας για την επιτυχία τους σε προβλήματα με μεγάλο όγκο δεδομένων ή με δεδομένα που γίνονται σταδιακά διαθέσιμα, με μορφή ροών. Τέλος, τα ΜαΣΤ αποτελούν παραλληλοποιήσιμα συστήματα, με αποτέλεσμα να είναι εύκολη η επέκτασή τους σε μεγάλη κλίμακα, χωρίς ιδιαίτερα προβλήματα.

Σκοπος της Διπλωματικης Εργασιας

Η παρούσα διπλωματική εργασία καταπιάνεται με το πρόβλημα της πολυκατηγορικής ταξινόμησης με Μανθάνοντα Συστήματα Ταξινομητών (ΜαΣΤ) τύπου Michigan (Παρ. 3.6.1). Πιο συγκεκριμένα, αναζητά τρόπους με τους οποίους μπορεί να επεκταθεί η υπάρχουσα υλοποίηση του GMl-ASLCS [Mil11], του πρώτου αλγορίθμου που επιχειρεί την επίλυση πολυκατηγορικών προβλημάτων με χρήση ΜαΣΤ. Η επέκταση έχει ως απώτερο στόχο την αύξηση της προβλεπτικής ικανότητας αυτού του αλγορίθμου, μέσω της τροποποίησης των επιμέρους τμημάτων του και της επινόησης νέων λειτουργιών, που προσιδιάζουν στη φύση των ΜαΣΤ, αλλά και της πολυκατηγορικής ταξινόμησης.

Στην εποπτευόμενη μάθηση, η ικανότητα ταξινόμησης ενός μοντέλου προσδιορίζεται από την ικανότητά του για ακριβή αντιστοίχιση ενός συνόλου κατηγοριών σε κάθε δείγμα ενός συνόλου δεδομένων, τα περιεχόμενα του οποίου είναι άγνωστα στο μοντέλο. Με άλλα λόγια πρόκειται για δείγματα με τα οποία το μοντέλο δεν έχει εκπαιδευτεί. Για να επιτύχει την αντιστοίχιση ένα ΜαΣΤ, θα πρέπει καθένας από τους κανόνες που χρησιμοποιεί να είναι σε θέση να ενεργοποιείται από ένα σύνολο άγνωστων προς αυτόν δειγμάτων και να τα ταξινομεί με ακρίβεια. Συνεπώς, το μοντέλο στο σύνολό του θα πρέπει να μπορεί να γενικεύει πάνω στο σύνολο δειγμάτων με το οποίο εκπαιδεύεται. Η παρούσα εργασία, λοιπόν, κινείται σε δύο άξονες: α) στοχεύει στην αύξηση του αριθμού των δειγμάτων που είναι ικανό να κατηγοριοποιήσει το μοντέλο που παράγει ο GMl-ASLCS, χωρίς να μειώνεται η προβλεπτική του ικανότητα, και β) στην αύξηση της ακρίβειας κατηγοριοποίησής του, αλλά και τη βελτίωση των υπολοίπων μετρικών που χρησιμοποιούνται για την αξιολόγηση των επιδόσεών του.

Μεθοδολογια Εργασιας

Η εκπόνηση της παρούσας εργασίας ξεκίνησε τον Απρίλιο του 2012 και ολοκληρώθηκε το Μάιο του 2013. Σε πρώτο στάδιο, στόχος υπήρξε η εξοικείωση με την ευρύτερη περιοχή του γνωστικού πεδίου της εργασίας. Για αυτό το σκοπό μελετήθηκε ένα σύνολο βιβλίων σχετικά με τον τομέα της Μηχανικής Μάθησης και των Γενετικών Αλγορίθμων. Η δεύτερη φάση περιελάμβανε την εξοικείωση με το ήδη υλοποιημένο λογισμικό, σε γλώσσα Java, την οργάνωση του κώδικά του και τις διάφορες λειτουργίες των τμημάτων που το απαρτίζουν.

Παράλληλα, μελετήθηκε βιβλιογραφία γύρω από τα Μανθάνοντα Συστήματα Ταξινομητών για την περίπτωση της μονοκατηγορικής ταξινόμησης και η εργασία που κυοφόρησε την απαρχή των ΜαΣΤ στον πολυκατηγορικό χώρο. Για την περαιτέρω εμβάθυνση στις εσωτερικές διεργασίες του λογισμικού, μελετήθηκαν τα εσωτερικά του, ξεχωριστά, τμήματα, ως προς την αρχή και τον τρόπο λειτουργίας τους.

Η επόμενη φάση περιελάμβανε την ανάγνωση επιστημονικών άρθρων σχετικά με την πολυκατηγορική ταξινόμηση που, λόγω της απουσίας βιβλιογραφίας της αντιμετώπισής της με ΜαΣΤ, κινούνταν στον αιτιοκρατικό χώρο. Εκεί ξεκίνησε και η μεγαλύτερη συγκέντρωση προς το λογισμικό. Στην πρώτη φάση της εργασίας, διενεργούνταν πειράματα πάνω στα τεχνητά σύνολα δεδομένων, λόγω του μικρού χρόνου εκτέλεσής τους, και στη συνέχεια ακολούθησαν πειράματα με τα πραγματικά σύνολα δεδομένων. Σε αυτό το στάδιο άρχισε και η εργασία της βελτιστοποίησης του GMl-ASLCS και της εύρεσης αποτελεσματικών διαφοροποιήσεων και καινούριων λειτουργιών. Κάθε προσθήκη ή τροποποίηση ελεγχόταν ως προς τη συνεισφορά της πάνω στο σύνολο των πραγματικών συνόλων δεδομένων και, στη συνέχεια, των τεχνητών.

Αφού ολοκληρώθηκε η φάση της βελτιστοποίησης και συγκεντρώθηκαν στον τελικό GMl-ASLCS όλες οι τροποποιημένες και επιπρόσθετες λειτουργίες, ξεκίνησε το τελικό στάδιο εκπαίδευσης του GMl-ASLCS πάνω στα έξι πραγματικά σύνολα δεδομένων που εξετάστηκαν σε αυτή την εργασία. Στη συνέχεια, διενεργήθηκαν πειράματα πάνω σε τέσσερα τεχνητά σύνολα δεδομένων, δοκιμάζοντας τη συμπεριφορά του GMl-ASLCS, όσο και των βασικών αλλαγών που επιφέραμε σε αυτόν. Στο τελικό στάδιο, πραγματοποιήθηκε καταγραφή της ανάλυσης και των συμπερασμάτων που προέκυψαν κατά τη διάρκεια της εκπόνησης αυτής της εργασίας.

Διαρθρωση της Αναφορας

Η παρούσα Διπλωματική Εργασία αποτελείται από τρία μέρη:

Στο **Μέρος Ι** περιγράφονται οι βασικές γνώσεις και μέθοδοι των γνωστικών πεδίων στα οποία στηρίχθηκε η εργασία. Πιο συγκεκριμένα, γίνεται αναφορά στη Μηχανική Μάθηση (Κεφάλαιο 2) και αναλύονται οι βασικές έννοιες και προσεγγίσεις της Εξόρυξης Δεδομένων που έχουν αναπτυχθεί τόσο για την απλή, όσο και για την πολυκατηγορική ταξινόμηση. Στο Κεφάλαιο 3 καταγράφονται οι βασικές έννοιες των Γενετικών Αλγορίθμων, όπως η έννοια του πληθυσμού, της καταλληλότητας και της φυσικής επιλογής, ενώ στο Κεφάλαιο 4 περιγράφεται η δομή και η

ιστορία των Μανθανόντων Συστημάτων Ταξινομητών (ΜαΣΤ). Πιο συγκεκριμένα, γίνεται αναφορά στα πρώτα ΜαΣΤ που χρησιμοποιούν Ενισχυτική Μάθηση και σε μεταγενέστερα που χρησιμοποιούνται σε προβλήματα Εποπτευόμενης Μάθησης.

Στο **Μέρος ΙΙ** καταγράφεται πλήρως ο τρόπος με τον οποίο επιχειρήθηκε να αντιμετωπιστεί το πρόβλημα της πολυκατηγορικής ταξινόμησης με Μανθάνοντα Συστήματα Ταξινομητών, έχοντας ως αφετηρία τον αλγόριθμο GMl-ASLCS, τον οποίο στο πλαίσιο της εργασίας ονοματοδοτούμε ως GMl-ASLCS$_0$. Αρχικά, στο Κεφάλαιο 5 παρουσιάζεται επιγραμματικά η δομή του GMl-ASLCS$_0$ και δίνεται έμφαση σε μερικές από τις λειτουργίες του, ώστε να γίνουν κατανοητοί οι λόγοι και η σημασία των αλλαγών που επιφέραμε σε αυτόν. Στο Κεφάλαιο 6 παρουσιάζεται η δομή του GMl-ASLCS, δηλαδή του τροποποιημένου GMl-ASLCS$_0$, ο κύκλος εκπαίδευσής του και αναλύονται μερικές από τις λειτουργίες των επιμέρους τμημάτων του, οι οποίες τροποποιήθηκαν. Παράλληλα, προτείνονται μερικές περαιτέρω τροποποιήσεις. Στο Κεφάλαιο 7 αξιολογούνται οι επιδόσεις του GMl-ASLCS και η επίδραση των τεσσάρων βασικών αρθρωτών αλλαγών που επιφέραμε στη λειτουργία του GMl-ASLCS$_0$, σε τέσσερα τεχνητά σύνολα δεδομένων. Στο Κεφάλαιο 8 αξιολογούνται οι επιδόσεις του GMl-ASLCS, αλλά και των προτεινόμενων τροποποιήσεών του σε έξι πραγματικά σύνολα πολυκατηγορικών δεδομένων.

Τέλος, στο **Μέρος ΙΙΙ** καταγράφονται τα συμπεράσματα που προέκυψαν από την παρούσα εργασία (Κεφάλαιο 9), καθώς και οι πιθανές μελλοντικές τροποποιήσεις και επεκτάσεις του GMl-ASLCS (Κεφάλαιο 10).

ΜΕΡΟΣ Ι

Ανασκόπηση Γνωστικού Πεδίου

2

Μηχανική Μάθηση και Κατηγοριοποίηση

Εισαγωγη

Ονομάζοντας το είδος μας *Homo Sapiens* - Άνθρωπος ο σοφός - λόγω της σαφούς διανοητικής μας διαφοράς από τα προηγούμενα ανθρώπινα είδη, αυτόματα τοποθετούμε τις ίδιες μας τις νοητικές ικανότητες σε ένα πεδίο ιδιαίτερης σημασίας για την ανθρωπότητα. Αυτές οι ικανότητες αποτέλεσαν ανά τους αιώνες, και ακόμα αποτελούν, αντικείμενο έρευνας με στόχο την κατανόηση του πώς σκεφτόμαστε, πώς δηλαδή όντα σαν εμάς αντιλαμβάνονται, καταλαβαίνουν, προβλέπουν και (μετα-)χειρίζονται το εξωτερικό περιβάλλον. Ένα περιβάλλον το οποίο είναι πολυπλοκότερο από τον ίδιο τον ανθρώπινο εγκέφαλο.

Για να αντιμετωπίσει τα προβλήματα που εγείρονταν μπροστά του και να ικανοποιήσει την εγγενή περιέργειά του, σε κάθε πεδίο της δραστηριότητάς του, ο Άνθρωπος ανέπτυξε τις επιστήμες. Παράλληλα, δημιούργησε *μηχανές* για την αυτοματοποίηση διαδικασιών και την επίλυση προβλημάτων. Στα μέσα του 20ού αιώνα είχε φτάσει σε τεχνολογικό και νοητικό επίπεδο τέτοιο, ώστε να εφεύρει την *Επιστήμη των Υπολογιστών* (Computer Science), δημιουργώντας έναν από τους πυλώνες της πραγματικότητάς μας. Λίγο αργότερα, ο Άνθρωπος προσπάθησε να κάνει το μεγάλο βήμα. Να συγκεράσει το πάθος του για κατανόηση με την τεχνολογική πρόοδο που ο ίδιος είχε επιφέρει. Να υπερβεί τον εαυτό του και να τον βοηθήσει να γίνει κοινωνός μίας "Αναγέννησης". Εγένετο τότε ένας νέος κλάδος της επιστήμης των υπολογιστών, αυτό που ονομάζουμε σήμερα *Τεχνητή Νοημοσύνη* (Artificial Intelligence).

Όπως είναι φυσικό, έχει δοθεί αφθονία ορισμών για το *τί* είναι η Τεχνητή Νοημοσύνη (ΤΝ). Ανεξαιρέτως, όμως, όλοι οι ορισμοί αναφέρονται σε δύο νοηματικούς άξονες: α) στη διαδικασία "σκέψης" και τη λογική - ορθολογικότητα και β) στη συμπεριφορά. Ο Alan Turing έθεσε τα θεμέλια της ΤΝ, θέτοντας τις προϋποθέ-

σεις ώστε μία μηχανή να μη μπορεί να διαχωριστεί από αναμφιβόλως σκεπτόμενα όντα. Χρησιμοποιώντας το μετασχηματισμό του ερωτήματος *"Μπορούν οι μηχανές να σκεφτούν;"* στο *"Μπορούν οι μηχανές να κάνουν ό,τι εμείς (ως σκεπτόμενα όντα) μπορούμε;"*, ο Turing υποστήριξε πως απαιτούνται τέσσερις ικανότητες από μία μηχανή για να μας πείσει ότι έχει τις ίδιες νοητικές ικανότητες με μας:

1. **Επεξεργασία φυσικής γλώσσας** για την επικοινωνία με τον άνθρωπο.

2. **Αναπαράσταση γνώσης** για την αποθήκευση αυτών που γνωρίζει.

3. **Αυτοματοποιημένη συλλογιστική** για τη χρήση αποθηκευμένων πληροφοριών, την απάντηση σε ερωτήματα και την εξαγωγή συμπερασμάτων.

4. **Μηχανική μάθηση** για την προσαρμογή της (μηχανής) σε νέες συνθήκες, την ανίχνευση και την προέκταση προτύπων γνώσης.

Οι τρεις πρώτες απαιτήσεις βρίσκονται έξω από το σκοπό αυτής της εργασίας. Εδώ, θα επικεντρωθούμε στη Μηχανική Μάθηση (ΜΜ), στον ευρύ υποτομέα της, την *Εξόρυξη Δεδομένων*, και πιο συγκεκριμένα στην *Κατηγοριοποίηση - Classification*. Ας πάρουμε όμως τα πράγματα από την αρχή.

Μηχανικη Μαθηση

Στόχος της ΜΜ είναι η σχεδίαση αλγορίθμων και προγραμμάτων υπολογιστών τα οποία είναι ικανά να μαθαίνουν, να προσαρμόζονται σε ένα πρόβλημα και να βελτιώνουν την επίδοσή τους αυτόματα από την εμπειρία. Πιο τυπικά, όπως ορίζεται στο [Mit97]:

Ένα πρόγραμμα υπολογιστή μαθαίνει, σε σχέση με ένα πρόβλημα T, μια μετρική επίδοσης P και μία μορφή εμπειρίας E, όταν η επίδοσή του πάνω στο T, όπως μετράται από την P, βελτιώνεται όσο αυξάνει η E.

Υπάρχει πληθώρα λόγων για την ανάπτυξη της ΜΜ. Κινητήριο μοχλό αποτέλεσε αρχικά η θέληση του ανθρώπου για ελαχιστοποίηση ζημιών και αύξηση της αποδοτικότητας, είτε αυτό μεταφράζεται σε νομισματικές μονάδες σε οποιοδήποτε οικονομικό σύστημα, ή σε καλύτερη ιατροφαρμακευτική περίθαλψη ασθενών μέσω προηγούμενων παραδειγμάτων διάγνωσης και θεραπείας. Αργότερα, η Εποχή της Πληροφορίας κατέκλυσε τον άνθρωπο με πλήθος δεδομένων τα οποία έπρεπε να φιλτραριστούν και να μετατραπούν σε (χρήσιμη) πληροφορία, κάτι που αποτελεί αντικείμενο της Εξόρυξης Δεδομένων (ΕΔ), ανιχνεύοντας πρότυπα "κρυμμένης πληροφορίας". Τέλος, κινητήριο μοχλό αποτέλεσε και θα αποτελεί η δυσκολία που συναντά ο Άνθρωπος τόσο στην περιγραφή όσο και στην απευθείας σχεδίαση διαδικασιών επίλυσής πολύπλοκων προβλημάτων, όπως για παράδειγμα η αναγνώριση εικόνων και προτύπων φωνής. Ανεξαρτήτως προβλήματος, κάθε προσπάθεια της ΜΜ για επιτυχή αντιμετώπισή του, πρέπει να χαρακτηρίζεται από προσαρμογή στο εξωτερικό περιβάλλον και ευελιξία στις μεταβολές του.

Όπως ξέρουμε, η έννοια της μάθησης και η προσπάθεια κατανόησης των διαδικασιών μάθησης δεν είναι κάτι καινοφανές επιστημονικά. Για αυτό και η ΜΜ δεν είναι ένας απομονωμένος τομέας αλλά θα μπορούσαμε να πούμε ότι είναι συγγενής με τις επιστήμες της Βιολογίας, της Ψυχολογίας, τη Στατιστική, και τομείς όπως οι Νευροεπιστήμες και η Θεωρία Πληροφοριών, πέρα φυσικά από την Επιστήμη των Υπολογιστών.

Οι αλγόριθμοι ΜΜ, ανάλογα με το επιθυμητό αποτέλεσμα ή τη μορφή της εισόδου κατά τη διαδικασία της εκπαίδευσής τους, μπορούν να οργανωθούν σε μία ταξονομία τριών θεμελιωδών μεθόδων μάθησης: την *Εποπτευόμενη Μάθηση*, την *Ενισχυτική Μάθηση* και τη *Μη-Εποπτευόμενη Μάθηση*.

Εποπτευόμενη Μάθηση

Η *εποπτευόμενη μάθηση* αναφέρεται στην εξαγωγή μίας συνάρτησης ή ενός μοντέλου από επισημασμένα δεδομένα εκπαίδευσης. Το μοντέλο αυτό χαρτογραφεί τη σχέση ανάμεσα σε ένα σύνολο γνωρισμάτων εισόδου και ενός (στην περίπτωση της μονοκατηγορικής ταξινόμησης) ή πολλών γνωρισμάτων εξόδου (στην περίπτωση που εξετάζουμε σε αυτή την εργασία - την πολυκατηγορική ταξινόμηση). Ένας αλγόριθμος εποπτευόμενης μάθησης δημιουργεί το εν λόγω μοντέλο έχοντας ως σκοπό την πρόβλεψη της τιμής εξόδου για κάθε τιμή εισόδου, γενικεύοντας παράλληλα από τα δεδομένα εκπαίδευσης σε άγνωστες καταστάσεις, με κάποια λογική διαδικασία [Tzi12].

Παράδειγμα εφαρμογής εποπτευόμενης μάθησης αποτελεί η απόφαση για την ύπαρξη εξωπλανητών που περιφέρονται γύρω από αστέρια, όπως είναι ο Ήλιος, μέσα στο γαλαξία μας ή ακόμα και σε άλλους. Η σύγχρονη Αστρονομία εστιάζει στον εντοπισμό πλανητικών συστημάτων, παίρνοντας μετρήσεις θέσης και ταχύτητας για να αποφασίσει εάν υπάρχουν πλανήτες στο εσωτερικό τους. Ένας αλγόριθμος εποπτευόμενης μάθησης θα είχε ως είσοδο ένα σύνολο διανυσμάτων μετρήσεων για διαφορετικούς αστέρες (ένα για τον καθένα), όπου το κάθε διάνυσμα συνοδεύεται από την πληροφορία παρατήρησης (ή μη) πλανητών σε δυαδική μορφή. Ο αλγόριθμος, εκπαιδευόμενος πάνω στο εν λόγω σύνολο δεδομένων, θα καλείτο να γενικεύσει και να δημιουργήσει ένα μοντέλο πρόβλεψης για μελλοντικές, μη επισημασμένες παρατηρήσεις.

Ενισχυτική Μάθηση

Εμπνευσμένη από τον συμπεριφορισμό, ένα από τα κύρια ρεύματα της Ψυχολογίας, η *ενισχυτική μάθηση* ασχολείται με τη μάθηση υπό το πρίσμα της ανταμοιβής και της τιμωρίας. Πιο συγκεκριμένα, ασχολείται με το πώς ένας πράκτορας (ή γενικότερα μία οντότητα λογισμικού) μπορεί να εκπαιδευτεί ώστε να επιλέγει ενέργειες σε ένα περιβάλλον, μεγιστοποιώντας κάποιας μορφής αθροιστική αριθμητική ανταμοιβή που λαμβάνει από αυτό. Το βασικό πλαίσιο ενισχυτικής μάθησης περιλαμβάνει ένα σύνολο καταστάσεων S του περιβάλλοντος, ένα σύνολο ενεργειών A του πράκτορα, κανόνες που διέπουν τη μετάβαση ανάμεσα στις καταστάσεις και κανόνες που καθορίζουν την άμεση βαθμωτή ανταμοιβή κατά τη μετάβαση σε μία κατάσταση. Ο πράκτορας αλληλεπιδρά με το περιβάλλον του σε διακριτές χρονικές

στιγμές. Σε δεδομένο χρόνο t, η κατάσταση του περιβάλλοντος είναι S_t και ο πράκτορας λαμβάνει μία παρατήρηση O_t, που συνήθως περιλαμβάνει και την ανταμοιβή r_t που προέκυψε κατά τη μετάβαση στην κατάσταση S_t. Ακολούθως, ο πράκτορας επιλέγει μία ενέργεια a_t από το σύνολο διαθεσίμων ενεργειών και αυτή αποστέλλεται στο περιβάλλον. Το περιβάλλον μεταβαίνει σε μία καινούρια κατάσταση S_{t+1}, στην οποία υπολογίζεται η ανταμοιβή r_{t+1} που συνδέεται με τη μετάβαση (S_t, a_t, S_{t+1}). Ο στόχος του πράκτορα είναι η μεγιστοποίηση όχι της άμεσης, αλλά της συνολικής ανταμοιβής που λαμβάνει. Όπως γίνεται κατανοητό, η ενισχυτική μάθηση διαφέρει από την εποπτευόμενη, καθώς το περιβάλλον δεν παρουσιάζει παραδείγματα στη μορφή ζευγαριών εισόδων/εξόδων στον υπό μαθήτευση αλγόριθμο, αλλά παίρνει έναν πιο "ενεργό" ρόλο.

Μη Εποπτευόμενη Μάθηση

Η *μη εποπτευόμενη μάθηση* διαφέρει ριζικά από τις δύο προηγούμενες τεχνικές μάθησης. Καμία είσοδος δεν συσχετίζεται με κάποιας μορφής έξοδο και δεν νοούνται αξιολογήσεις από το περιβάλλον. Τα παραδείγματα εκπαίδευσης διαθέτουν μόνο γνωρίσματα εισόδου. Σε αυτό το πλαίσιο, η μη εποπτευόμενη μάθηση αναφέρεται στην εξαγωγή μίας κρυμμένης δομής που αντανακλά τη στατιστική δομή των μη επισημασμένων δειγμάτων εισόδου. Η μη εποπτευόμενη μάθηση βρίσκει χρήσεις στη λήψη αποφάσεων, την πρόβλεψη μελλοντικών εισόδων ή την ομαδοποίηση παρομοίων εισόδων, ενώ βασικές της προσεγγίσεις αποτελούν η ομαδοποίηση (clustering) και ο διαχωρισμός σημάτων (blind signal separation) για τη μείωση της διαστατικότητας (dimensionality reduction).

Εξορυξη Δεδομενων

Προεκτείνοντας τα όσα αναφέραμε προηγουμένως, στόχος της ΕΔ είναι η εξαγωγή προηγουμένως αγνώστων, ορθών, κατανοητών και ενδιαφερόντων πληροφοριών από ένα αχανές και αταξινόμητο σύνολο δεδομένων. Ως Πληροφορία νοείται η επεξεργασία, οργάνωση, δόμηση και παρουσίαση σε πλαίσιο συμφραζομένων των ακατέργαστων κομματιών γνώσης που αποτελούν τα δεδομένα, ώστε να καταστούν χρήσιμα σε κάποια διαδικασία στην οποία εμπλέκεται ο άνθρωπος. Στη συνέχεια, εστιάζουμε στο γιγάντιο μέγεθος του εγχειρήματος, γιατί από αυτό εφορμάται η ΕΔ.

Η αλματώδης χρήση των υπολογιστών και του Διαδικτύου, που εν μέρει οφείλεται στην αυξανόμενη προσιτότητά τους στο καταναλωτικό κοινό, σε συνδυασμό με τη συνεχόμενη πτώση της τιμής των αποθηκευτικών μέσων, έχει αυξήσει εκθετικά τα παραγόμενα δεδομένα. Ανεξάρτητα από αυτό, ενδιαφέρον αποτελεί και το μέγεθος των παραγόμενων δεδομένων μη εμπορικών συστημάτων όπως ο ανιχνευτής σωματιδίων ATLAS στο CERN. Εκεί, όταν διενεργούνται πειράματα, κάθε δευτερόλεπτο συντελούνται 40 εκατομμύρια κρούσεις, με την κάθε μία να παράγει δεδομένα που καταλαμβάνουν 25 MB χώρου. Με άλλα λόγια, κάθε δευτερόλεπτο παράγεται ένα petabyte ή 1000 terabytes δεδομένων. Όπως είναι κατανοητό, κάποιου είδους λογισμικό ΕΔ πρέπει να μπει σε εφαρμογή, σε πρώτη φάση για

το φιλτράρισμα πληροφοριών οι οποίες είναι άχρηστες (ή καλύτερα γνωστές και τετριμμένες για τους επιστήμονες) και σε δεύτερη φάση για την ανίχνευση και απομόνωση της χρήσιμης πληροφορίας.

Γενικά, οι εργασίες ΕΔ χωρίζονται σε δυο κατηγορίες: τις *εργασίες πρόβλεψης* και τις *εργασίες περιγραφής*.

Οι **εργασίες πρόβλεψης** έχουν ως σκοπό την πρόβλεψη μίας ή περισσοτέρων εξαρτημένων μεταβλητών από ένα σύνολο ανεξάρτητων μεταβλητών που τις περιγράφουν. Πιο συγκεκριμένα [TSK05], ανάλογα με το αν η μεταβλητή προς πρόβλεψη είναι διακριτή ή συνεχής, αναφερόμαστε σε αλγορίθμους **ταξινόμησης ή κατηγοριοποίησης (classification)**, και **παλινδρόμησης (regression)**, αντίστοιχα.

Οι **εργασίες περιγραφής** έχουν ως στόχο την εξαγωγή προτύπων και την περιγραφή των υπό μελέτη προβλημάτων. Οι τρεις κατηγορίες στις οποίες ανήκουν οι εργασίες ανάλυσης δεδομένων είναι:

1. **Ανάλυση συσχετίσεων**: Εύρεση συσχετίσεων των γνωρισμάτων από τα δεδομένα.

2. **Ανίχνευση ανωμαλιών**: Ανίχνευση των δειγμάτων που εμφανίζουν σημαντικές διαφορές από το μεγαλύτερο μέρος των υπόλοιπων δειγμάτων.

3. **Ανάλυση ομαδοποίησης**: Σχηματισμός ομάδων (clusters) δειγμάτων με συναφή γνωρίσματα.

Η παρούσα εργασία εστιάζει σε εργασίες ταξινόμησης.

Κατηγοριοποιηση

Κατηγοριοποίηση δεδομένων *ονομάζεται η διαδικασία υπαγωγής των εγγράφων μίας βάσης δεδομένων σε ένα πεπερασμένο και προκαθορισμένο σύνολο κατηγοριών - ετικετών.*

Στην περίπτωση της μονοκατηγορικής ταξινόμησης, που εξετάζουμε σε αυτή την ενότητα, μία εγγραφή συσχετίζεται με μία μόνο κατηγορία, ενώ όπως θα δούμε αργότερα, στην πολυκατηγορική ταξινόμηση, μία εγγραφή συσχετίζεται με ένα σύνολο ετικετών L. Η τιμή της (μοναδικής) κατηγορίας στην περίπτωση της μονοκατηγορικής ταξινόμησης συχνά αναφέρεται ως *κλάση*.

Ένας αλγόριθμος ταξινόμησης λαμβάνει ως είσοδο ένα εκ των προτέρων κατηγοριοποιημένο σύνολο δεδομένων, το οποίο ονομάζουμε *σύνολο εκπαίδευσης*, και κατασκευάζει ένα μοντέλο πρόβλεψης, ανακαλύπτοντας συσχετίσεις των γνωρισμάτων των δειγμάτων[1] εκπαίδευσης με τις αντίστοιχές τους κλάσεις. Αφού δημιουργηθεί το μοντέλο πρόβλεψης, παρουσιάζεται σε αυτό ένα διαφορετικό εν γένει σύνολο δεδομένων, το οποίο ονομάζουμε *σύνολο ελέγχου ή σύνολο επικύρωσης*. Οι εγγραφές του συνόλου ελέγχου είναι *μη* επισημασμένες, δηλαδή δεν είναι γνωστή η κατηγορία στην οποία ανήκει η κάθε μία. Το μοντέλο, σε αυτή τη φάση της διαδικασίας, προσπαθεί να προβλέψει την τιμή της κατηγορίας στην οποία ανήκει το κάθε δείγμα. Αφού τελειώσει η επισήμανση κάθε δείγματος του συνόλου ελέγχου, ο

[1] Οι όροι *εγγραφή* και *δείγμα* χρησιμοποιούνται ισοδύναμα.

αλγόριθμος συγκρίνει την πραγματική κλάση με αυτήν που προέβλεψε το μοντέλο. Προφανώς, αν συμπίπτουν, η πρόβλεψη θεωρείται επιτυχημένη, ενώ στην αντίθετη περίπτωση αποτυχημένη.

Από την περιγραφή της γενικής φιλοσοφίας της διαδικασίας παραπάνω, είναι πρόδηλη η σημασία της ικανότητας γενίκευσης του μοντέλου: πρέπει να μπορεί όχι μόνο να κατηγοριοποιεί άγνωστα προς αυτό δείγματα, αλλά και να τα κατηγοριοποιεί με επιτυχία. Πρέπει δηλαδή να αποφύγει την υπερειδίκευση (overfitting-overtraining) πάνω στο σύνολο εκπαίδευσης, για να μπορεί να κατηγοριοποιεί τα δείγματα του συνόλου ελέγχου, και να έχει επαρκώς καλή ακρίβεια πρόβλεψης, για να ταξινομεί ορθά.

Το σύνολο ελέγχου είναι όμοιο στη δομή με αυτό που χρησιμοποιείται στην εκπαίδευση του αλγορίθμου. Για την εξαγωγή τους από ένα ενιαίο σύνολο δειγμάτων χρησιμοποιούνται οι εξής μέθοδοι:

Παρακράτηση (Holdout) Το αρχικό σύνολο δεδομένων D χωρίζεται σε δύο ανεξάρτητα υποσύνολα, D_{tr} και D_{te}. Το D_{tr} θα αποτελέσει το σύνολο εκπαίδευσης και το D_{te} το σύνολο ελέγχου.

Διασταυρωμένη Επικύρωση k-πτυχών (k-fold cross-validation) Το αρχικό σύνολο δεδομένων D χωρίζεται σε k υποσύνολα D_i. Στη συνέχεια, ο ταξινομητής[2] εκπαιδεύεται k φορές, χρησιμοποιώντας κάθε φορά όλα τα D_i για την εκπαίδευση εκτός από ένα. Έτσι, κάθε υποσύνολο D_i ανήκει $k-1$ φορές στο σύνολο εκπαίδευσης, ενώ μόνο μία φορά αποτελεί το ίδιο, εξ' ολοκλήρου, το σύνολο ελέγχου.

Δειγματοληψία με επανατοποθέτηση (Bootstrap) Το αρχικό σύνολο δεδομένων D δειγματοληπτείται τόσες φορές όσες είναι και τα δείγματά του. Κάθε φορά, επιλέγεται ένα δείγμα και τοποθετείται στο σύνολο εκπαίδευσης, χωρίς να αφαιρεθεί από το αρχικό σύνολο. Με αυτή τη διαδικασία προκύπτει το σύνολο εκπαίδευσης D_{tr}, με αριθμό δειγμάτων ίσο με το αρχικό σύνολο δεδομένων, που, εν γένει, περιέχει πολλαπλά αντίτυπα του ίδιου δείγματος. Το σύνολο ελέγχου προκύπτει ως η διαφορά των συνόλων $D - D_{tr}$.

Μέθοδοι κατηγοριοποίησης

Δέντρα αποφάσεων ταξινόμησης

Μορφολογικά, ένα δέντρο απόφασης είναι ένα διάγραμμα ροής που έχει τη μορφή δέντρου. Κάθε εσωτερικός κόμβος (internal node) αντιστοιχεί σε μία απόφαση για ένα γνώρισμα, ενώ κάθε φύλλο του δέντρου αναπαριστά την κλάση πρόβλεψης. Η ευκολία στην κατανόηση και ερμηνεία των αποτελεσμάτων, εν μέρει λόγω του μοντέλου λευκού κουτιού που χρησιμοποιούν, το σχετικά μικρό έργο προετοιμασίας των δεδομένων και η καλή τους επίδοση σε μεγάλα σύνολα δεδομένων, έχουν καταστήσει τα δέντρα απόφασης μία ευρέως χρησιμοποιούμενη μέθοδο ταξινόμησης [Mur98].

[2] Οι όροι *ταξινομητής* και *μοντέλο πρόβλεψης* χρησιμοποιούνται ισοδύναμα.

Πιθανοτικοί ταξινομητές

Αποτελούν μία πιθανοτική μέθοδο επίλυσης προβλημάτων, προβλέποντας την πιθανότητα ένα δείγμα να ανήκει σε μία κλάση [HSCB91]. Στηρίζονται στο θεώρημα του Bayes που συνδέει την εκ των προτέρων με την εκ των υστέρων πιθανότητα υπαγωγής ενός δείγματος σε μία συγκεκριμένη κλάση μέσω της σχέσης

$$P(H|\vec{X}) = \frac{P(\vec{X}|H)P(H)}{P(\vec{X})} \tag{2.1}$$

όπου $\vec{X}$ το σύνολο των παρατηρούμενων γνωρισμάτων και H η κλάση/κατηγορία. Οι *Απλοί Πιθανοτικοί Ταξινομητές* θεωρούν ότι η παρουσία ή απουσία ενός γνωρίσματος δεν σχετίζεται με την παρουσία ή απουσία ενός οποιουδήποτε άλλου γνωρίσματος, δεδομένης της κλάσης με την οποία συσχετίζεται το δείγμα. Στον αντίποδα, τα *Πιθανοτικά Δίκτυα Συσχέτισης* λαμβάνουν υπόψη τις αλληλεξαρτήσεις ανάμεσα στα γνωρίσματα, παρουσιάζοντας συνήθως καλύτερη ικανότητα ταξινόμησης από τους απλούς πιθανοτικούς ταξινομητές. Χρησιμοποιούν ένα κατευθυνόμενο ακυκλικό γράφο και ένα σύνολο πινάκων των δεσμευμένων πιθανοτήτων για κάθε κόμβο του γράφου.

Ταξινομητές k Πλησιέστερων Δειγμάτων

Με δεδομένο τον αριθμό k, οι ταξινομητές k πλησιέστερων δειγμάτων προβλέπουν την κλάση ενός νέου δείγματος με βάση τους k πλησιέστερους γείτονές του [AKA91]. Η μετρική απόστασης για τον εντοπισμό των k πλησιέστερων γειτόνων μπορεί να είναι η Ευκλείδεια απόσταση, η απόσταση Manhattan ή άλλες. Ο ταξινομητής υπολογίζει το μέσο όρο των τιμών της μετρικής πρόβλεψης για δείγματα με συνεχείς μεταβλητές προς πρόβλεψη ή λαμβάνει την απόφασή του από την ψηφοφορία των k γειτόνων του σημείου εάν η μεταβλητή προς πρόβλεψη ανήκει σε διακριτές κατηγορίες.

Τεχνητά Νευρωνικά Δίκτυα

Εμπνευσμένα από τη μελέτη του Κεντρικού Νευρικού Συστήματος των θηλαστικών, τα Τεχνητά Νευρωνικά Δίκτυα (ΤΝΔ), αποτελούν ένα προσαρμοστικό σύστημα που αποτελείται από ένα σύνολο τεχνητών κόμβων, ή νευρώνων, συνδεδεμένων μεταξύ τους σε μία δομή δικτύου που προσομοιάζει στα βιολογικά νευρωνικά δίκτυα [Zha00]. Κάθε νευρώνας έχει N εισόδους x_i και μία έξοδο y. Κάθε είσοδος φέρει ένα βάρος w_i με το οποίο πολλαπλασιάζεται. Η έξοδος του νευρώνα υπολογίζεται από τη σχέση

$$y = f\left(\sum_{i=1}^{N} w_i x_i\right) \tag{2.2}$$

όπου f μια συνάρτηση μεταφοράς, συχνά η σιγμοειδής $\frac{1}{1+e^{-x}}$ ή η υπερβολική εφαπτομένη $tanh(x)$. Η γενική αρχιτεκτονική ενός ΤΝΔ χρησιμοποιεί έναν αριθμό από επίπεδα νευρώνων, με το πρώτο να δέχεται τις εισόδους του ΤΝΔ και τα υπόλοιπα

εκτός από το τελευταίο να είναι κρυφά. Τα κρυφά επίπεδα δέχονται ως εισόδους τις εξόδους του προηγούμενου επιπέδου, ενώ το τελευταίο επίπεδο αποτελεί την έξοδο του ΤΝΔ.

Μηχανές Διανυσμάτων Υποστήριξης - SVMs

Σκοπός των *Μηχανών Διανυσμάτων Υποστήριξης (ΜΔΥ)* είναι η εύρεση ενός ή περισσοτέρων υπερεπιπέδων στον πολυδιάστατο χώρο των γνωρισμάτων που να διαχωρίζουν δύο κατηγορίες δειγμάτων με το μεγαλύτερο δυνατό περιθώριο (margin) μεταξύ τους, ελαχιστοποιώντας το σφάλμα γενίκευσης του ταξινομητή [BGV92]. Ανάλογα με το αν τα δεδομένα είναι γραμμικά διαχωρίσιμα μεταξύ τους, χρησιμοποιούνται γραμμικές ή μη γραμμικές ΜΔΥ. Στη μη γραμμική περίπτωση, χρησιμοποιούνται μη γραμμικοί πυρήνες (kernels) για το μετασχηματισμό του προβλήματος στο αντίστοιχο γραμμικό.

Ταξινομητές Βασισμένοι σε Κανόνες

Οι *Ταξινομητές Βασισμένοι σε Κανόνες* κατασκευάζουν ένα σύνολο *κανόνων πρόβλεψης* R, της μορφής *εάν - τότε (if - then)*. Ένας κανόνας πρόβλεψης $r_i \in R$ είναι μία λογική πρόταση που αποτελείται από δύο μέρη: το λογικό τμήμα συνθηκών πρόβλεψης (rule precondition/antecedant) και την κλάση πρόβλεψης (rule concequent) y_i, την οποία λέμε ότι *στηρίζει (advocates)* ο κανόνας.

$$r_i : (\text{Συνθήκη})_i \Rightarrow y_i \tag{2.3}$$

Το τμήμα συνθήκης έχει την εξής μορφή:

$$(\text{Συνθήκη})_i = (A_1 \; op \; u_1) \wedge (A_2 \; op \; u_2) \wedge \cdots \wedge (A_l \; op \; u_l) \tag{2.4}$$

όπου u_i μπορεί να είναι μία τιμή, ένα διάστημα, ή ένα σύνολο τιμών. Με άλλα λόγια, το τμήμα συνθήκης αποτελείται από τη σύζευξη ενός συνόλου ζευγών γνωρισμάτων A_i και των αντίστοιχων επιτρεπτών τιμών u_i, συνδυασμένων με έναν τελεστή op. Το σύνολο κανόνων R είναι η ένωση όλων των κανόνων r_i:

$$R = (r_1 \vee r_2 \vee \cdots \vee r_k) \tag{2.5}$$

Λέμε ότι ένας κανόνας r *καλύπτει* ένα δείγμα s όταν υπάρχει μία ένα-προς-ένα αντιστοιχία της διάταξης των τιμών των γνωρισμάτων του s με αυτές του τμήματος συνθήκης του r. Αν ο r καλύπτει το s, λέμε ότι ο r ενεργοποιείται για το δείγμα s, ενώ αν η κατηγορία που στηρίζει ο r είναι ίδια με αυτή του s, τότε ο κανόνας κάνει *ορθή* ταξινόμηση.

Ταξινομητές συνόλων

Οι **ταξινομητές συνόλων (ensemble classifiers)** εκπαιδεύουν ένα σύνολο ταξινομητών από τις παραπάνω κατηγορίες για το ίδιο σύνολο δεδομένων και στο τέλος συνδυάζουν τα αποτελέσματα του καθενός. Οι πιο διαδεδομένες μέθοδοι κατασκευής ταξινομητών συνόλων είναι οι εξής:

1. **Ομαδοποίηση (Bagging)** Το σύνολο εκπαίδευσης δειγματοληπτείται με επανατοποθέτηση ώστε να δημιουργηθεί ένα σύνολο μικρότερων συνόλων δεδομένων D_i και, στη συνέχεια, εκπαιδεύεται ένας ταξινομητής M_i με το κάθε ένα. Όταν χρειαστεί να ταξινομηθεί ένα άγνωστο δείγμα, κατηγοριοποιείται αρχικά από όλους τους ταξινομητές και η τελική απόφαση λαμβάνεται μέσω ψηφοφορίας μεταξύ τους.

2. **Ενίσχυση (Boosting)** Εδώ, εκπαιδεύεται και πάλι ένα σύνολο ταξινομητών M_i, με τη διαφορά ότι το κάθε δείγμα φέρει και ένα βάρος (ανάλογα με τη "σημαντικότητά" του στο σύνολο που ανήκει). Το βάρος αυτό ανανεώνεται επαναληπτικά: κάθε δείγμα που κατηγοριοποιείται λανθασμένα από τον M_i αυξάνει το βάρος του για τον επόμενο ταξινομητή, ενώ στην αντίθετη περίπτωση το μειώνει. Έτσι, κάθε ταξινομητής καλείται να δώσει μεγαλύτερη βαρύτητα στα δείγματα που έχουν διαφύγει ορθής κατηγοριοποίησης μέχρι στιγμής. Η τελική απόφαση λαμβάνεται μέσω ψηφοφορίας, πιθανόν με βάρος ως προς κάποια μετρική αξιολόγησης.

Μετρικές Αξιολόγησης Αλγορίθμων Κατηγοριοποίησης

Ως μέτρο αξιολόγησης της ικανότητας ταξινόμησης των διαφόρων ταξινομητών έχουν προταθεί πολλές μετρικές και εργαλεία. Παρακάτω αναφέρουμε τα σημαντικότερα.

Μετρικές με όρους αποδοχής και απόρριψης

Γενικά, λέμε ότι ένας ταξινομητής *αποδέχεται* ένα δείγμα όταν το κατατάσσει στην κατηγορία στόχο, ενώ το *απορρίπτει* σε αντίθετη περίπτωση.

Πίνακας 2.1: Τύποι Αποφάσεων Ταξινομητή

	Πραγματικότητα	
Απόφαση	**Κατηγορία Στόχος**	**Όχι Κατηγορία Στόχος**
Αποδοχή	Ορθή Αποδοχή (TP)	Εσφαλμένη Αποδοχή (FP)
Απόρριψη	Εσφαλμένη Απόρριψη (FN)	Ορθή Απόρριψη (TN)

Πιο συγκεκριμένα μπορούμε να διακρίνουμε τα ακόλουθα μεγέθη:

TP - true positives
Το πλήθος των *αληθώς αποδεκτών* δειγμάτων είναι ο αριθμός των δειγμάτων που αποδέχεται ορθά ο ταξινομητής.

FP - false positives
Το πλήθος των *εσφαλμένα αποδεκτών* δειγμάτων αντιπροσωπεύει τον αριθμό των δειγμάτων που αποδέχεται ο ταξινομητής, ενώ στην πραγματικότητα ανήκουν σε διαφορετική κλάση.

TN - true negatives
Το πλήθος των *ορθώς απορριφθέντων* δειγμάτων είναι ο αριθμός των δειγμάτων που ορθώς δεν ταξινομήθηκαν στην κλάση-στόχο.

FN - false negatives
Το πλήθος των *εσφαλμένα απορριφθέντων* δειγμάτων είναι ο αριθμός των δειγμάτων που απέρριψε ο ταξινομητής, ενώ στην πραγματικότητα ανήκουν στην κλάση-στόχο.

Χρησιμοποιώντας τα μεγέθη του πίνακα 2.1 μπορούμε να ορίσουμε τις παρακάτω μετρικές αξιολόγησης:

Πιστότητα (Precision)

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2.6}$$

Είναι το ποσοστό των ορθών κατηγοριοποιήσεων του ταξινομητή στο σύνολο των δειγμάτων που αποδέχεται.

Ανάκληση (Recall) ή Ευαισθησία (Sensitivity) ή True Positive Rate

$$\text{Recall} = \frac{TP}{TP + FN} \tag{2.7}$$

Είναι το ποσοστό των δειγμάτων που ο ταξινομητής κατηγοριοποιεί ορθά, από το σύνολο των δειγμάτων που στην πραγματικότητα ανήκουν στην κλάση-στόχο. Με άλλα λόγια, είναι η ικανότητα του ταξινομητή να εντοπίζει δείγματα που ανήκουν σε μία συγκεκριμένη κλάση.

Ειδικότητα (Specificity) ή True Negative Rate

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{2.8}$$

Αξιολογεί την ικανότητα ορθής απόρριψης του ταξινομητή. Πιο συγκεκριμένα, είναι το ποσοστό των δειγμάτων που απορρίφθηκαν ορθά στο σύνολο των δειγμάτων που έπρεπε να απορριφθούν.

Ακρίβεια (Accuracy)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{2.9}$$

Η ακρίβεια είναι η μετρική με τη συχνότερη χρήση στη βιβλιογραφία. Ορίζεται ως το ποσοστό των δειγμάτων που κατηγοριοποιεί ορθά ένας ταξινομητής από το σύνολο όλων των δειγμάτων που καλείται να κατηγοριοποιήσει.

Καμπύλες ROC

Οι καμπύλες ROC (Receiver Operating Characteristics) αποτελούν ένα εργαλείο για την οπτικοποίηση και απεικόνιση της μεταβολής της ευαισθησίας σε σχέση με την ειδικότητα. Για κάθε ταξινομητή είναι δεδομένο πως μπορεί να ταξινομεί ορθά τα δείγματα μίας κλάσης, δηλαδή να έχει υψηλή ευαισθησία, αλλά ταυτόχρονα να κατατάσσει στη συγκεκριμένη κλάση πολλά δείγματα που δεν ανήκουν σε αυτή, δηλαδή να έχει χαμηλή ειδικότητα. Με τη χρήση των καμπυλών ROC είναι δυνατόν να συγκριθούν διαφορετικοί ταξινομητές ως προς την ικανότητά τους να διαθέτουν υψηλή ευαισθησία και παράλληλα υψηλή ειδικότητα.

Μετρικές ειδικά για ταξινομητές βασισμένους σε κανόνες

Παρακάτω περιγράφονται μερικές κύριες μετρικές για την αξιολόγηση αυτή την φορά κανόνων και όχι ταξινομητών.

Η *κάλυψη* - *coverage* c_i ενός κανόνα είναι το ποσοστό των δειγμάτων $|A|$ για το οποίο ενεργοποιείται ο κανόνας, στο σύνολο όλων των δειγμάτων του συνόλου δεδομένων $|D|$.

$$c_i = \frac{|A|}{|D|} \tag{2.10}$$

Η *ακρίβεια* ενός κανόνα μπορεί να οριστεί ως ο αριθμός των δειγμάτων που ο κανόνας κατηγοριοποιεί σωστά, προς τον αριθμό των δειγμάτων που καλύπτει.

Οι δύο παραπάνω μετρικές δεν μπορούν να εκτιμήσουν με αξιοπιστία την πλήρη ποιότητα των κανόνων [HK06], και για αυτό τις χρησιμοποιούμε συχνά σε συνδυασμό με το *κέρδος πληροφορίας*, το *κέρδος FOIL* ή το *λόγο πιθανοφάνειας*. Ας τις κοιτάξουμε από κοντά.

Το *κέρδος πληροφορίας (information gain)* H ταυτίζεται με την έννοια της εντροπίας

$$H(A) = -\sum_{i=1}^{m} p_i \, log_2(p_i) \tag{2.11}$$

όπου A είναι το σύνολο των δειγμάτων που καλύπτει ο κανόνας ($m = |A|$), cl_i οι κλάσεις στο $|A|$ και p_i η πιθανότητα εμφάνισης της κλάσης cl_i στο A. Όσο μικρότερη η εντροπία ενός κανόνα, τόσο μεγαλύτερη αξιοπιστία έχει, καθώς πραγματοποιεί λιγότερα λάθη.

Το *κέρδος FOIL* χρησιμοποιείται για τη σύγκριση δύο κανόνων. Αν R και R' οι δύο κανόνες προς σύγκριση, p και p' ο αριθμός των δειγμάτων που ταξινομούν σωστά αντίστοιχα, και n και n' τα δείγματα που ταξινομούν λανθασμένα,τότε το $FOIL_{Gain}$ του R' ως προς το R δίνεται από τη σχέση

$$FOIL_{Gain} = p' \left(log_2 \left(\frac{p'}{p' + n'} \right) - log_2 \left(\frac{p}{p + n} \right) \right) \tag{2.12}$$

Ο *λόγος πιθανοφάνειας (likelihood ratio)* μπορεί να βοηθήσει στην εξέταση της στατιστικής σημαντικότητας, ή μη, των ταξινομήσεων ενός κανόνα, συνυπολογίζοντας την πιθανότητα ένα δείγμα να εμφανιστεί στο σύνολο δεδομένων.

$$LR = 2 \times \sum_{i=1}^{m} f_i \, log \left(\frac{f_i}{e_i} \right) \tag{2.13}$$

όπου f_i η παρατηρούμενη πιθανότητα της κατηγορίας i στο σύνολο δεδομένων και e_i η αναμενόμενη συχνότητα ενεργοποίησης του κανόνα εάν έκανε τυχαίες επιλογές.

Πολυκατηγορικη ταξινομηση

Στην προηγούμενη ενότητα εξετάσαμε μονοκατηγορικούς ταξινομητές, δηλαδή ταξινομητές των οποίων το έργο είναι η κατηγοριοποίηση δειγμάτων σε μία μόνο κλάση cl_i από ένα σύνολο αμοιβαίως αποκλειόμενων κλάσεων CL. Η πολυκατηγορική ταξινόμηση (ΠΤ) αποτελεί γενίκευση της μονοκατηγορικής ταξινόμησης. Οι πολυκατηγορικοί ταξινομητές δέχονται ως είσοδο και χειρίζονται δείγματα που σχετίζονται με ένα σύνολο ετικετών[3] $Y \subseteq L$, όπου L είναι το σύνολο των διαθέσιμων ετικετών. Στόχος τους αποτελεί η πρόβλεψη των ετικετών, δηλαδή η πρόβλεψη του αν ένα δείγμα ανήκει ή όχι σε μία ετικέτα, με βάση το χώρο γνωρισμάτων, εκμεταλλευόμενοι παράλληλα τις συσχετίσεις στο χώρο ετικετών. Ας υπογραμμίσουμε εδώ ότι οι κλάσεις, οι τιμές δηλαδή των ετικετών, στην ΠΤ, είναι δυαδικές για όλες τις ετικέτες.

Από την πλευρά της μηχανικής μάθησης, ο ορισμός του προβλήματος έχει ως εξής:

[3]Οι όροι *ετικέτα* και *κατηγορία* χρησιμοποιούνται ισοδύναμα.

Με βάση ένα σύνολο ετικετών L και ένα σύνολο δειγμάτων εκπαίδευσης D, καθένα από τα οποία συσχετίζεται με ένα υποσύνολο ετικετών $Y \subseteq L$, ένας (πολυκατηγορικός) ταξινομητής μαθαίνει να προβλέπει τις ετικέτες ενός συνόλου ελέγχου T, στο οποίο και αξιολογείται.

Η κατηγοριοποίηση μονής ετικέτας είναι σαφέστατα πιο διαδεδομένη από την ΠΤ και αποτέλεσε τη βάση για τις τεχνικές δημιουργίας και ανάπτυξης μοντέλων ΠΤ. Διαισθητικά, όμως, η ΠΤ δεν μπορεί να θεωρηθεί λιγότερο φυσική από τη μονοκατηγορική ταξινόμηση, καθώς ο ανθρώπινος εγκέφαλος είναι εγγενώς ικανός να συσχετίζει ένα αντικείμενο ή μία ιδέα με πολλαπλές χρήσεις ή έννοιες. Παρ' όλα αυτά, η ΠΤ εισάγει μία επιπλέον διάσταση, αυτή των πολλαπλών ετικετών με τις οποίες συσχετίζονται τα δείγματα, επηρεάζοντας έτσι τόσο τις διαδικασίες μάθησης, όσο και τις διαδικασίες αξιολόγησης των χρησιμοποιούμενων ταξινομητών.

Η βιβλιογραφία αναφέρει δύο βασικές μεθόδους για την κατασκευή πολυκατηγορικών ταξινομητών [TK07, Aly05]:

1. Το μετασχηματισμό ενός προβλήματος ΠΤ σε ένα ή περισσότερα προβλήματα κατηγοριοποίησης μίας κατηγορίας, χρησιμοποιώντας κλασικούς αλγορίθμους μηχανικής μάθησης των οποίων οι προβλέψεις μετασχηματίζονται στη ζητούμενη πολυκατηγορική πρόβλεψη, και

2. Την προσαρμογή γνωστών αλγορίθμων μονοκατηγορικής ταξινόμησης και τεχνικών μηχανικής μάθησης, ώστε να καταστούν απευθείας εφαρμόσιμοι στην πολυκατηγορική ταξινόμηση.

Μέθοδοι Μετασχηματισμού Προβλημάτων

BR - Δυαδικής συνάφειας (Binary Relevance)

Έστω ότι το πολυκατηγορικό πρόβλημα χρησιμοποιεί L ετικέτες. Τότε, ο μετασχηματισμός BR διασπά το πρόβλημα σε L δυαδικά προβλήματα απλής κατηγοριοποίησης, ένα για κάθε ετικέτα $l \in L$. Για κάθε ετικέτα l, εκπαιδεύεται ένας ταξινομητής, του οποίου καθήκον είναι να αποφανθεί εάν ένα δείγμα κατηγοριοποιείται στη συγκεκριμένη ετικέτα ή όχι. Η πολυκατηγορική ταξινόμηση ενός δείγματος συνίσταται στην επιλογή των ετικετών για τις οποίες οι αντίστοιχοι ταξινομητές αποφάνθηκαν θετικά. Μεγάλα μειονεκτήματα της μεθόδου BR είναι η αδυναμία της να λάβει υπόψη τις συσχετίσεις ανάμεσα στις διάφορες ετικέτες [YTS07, Rea10] (θα επεκταθούμε περισσότερο στο θέμα της συσχέτισης ετικετών στην Παρ. 2.5.3) και η έντονη επήρεια από την ύπαρξη ανισορροπίας κλάσεων [RLS04] – λόγω της τυπικής σποραδικότητας των ετικετών στα πολυκατηγορικά δεδομένα, κάθε δυαδικός ταξινομητής είναι πιθανό να πρέπει να εκπαιδευτεί με πολύ περισσότερα αρνητικά από ότι θετικά δείγματα.

PW - Ταξινόμηση ανά ζεύγη (Pairwise Classicification)

Σε αντίθεση με τον BR, ο μετασχηματισμός PW συσχετίζει τον κάθε ταξινομητή με ένα ζεύγος ετικετών και έτσι σχηματίζονται αντί για L, $|L|(|L|-1)/2$ δυαδικά προβλήματα, ένα για κάθε ζεύγος ετικετών. Μειονεκτήματα της μεθόδου αυτής είναι η μεγάλη χρονική πολυπλοκότητα και η αδυναμία να λάβει υπόψη της συσχετισμούς άνω των δύο ετικετών.

LC - Συνδυασμός Ετικετών (Label Combinations)

Ο μετασχηματισμός LC, γνωστός ως και ως μετασχηματισμός *Δυναμοσυνόλου Ετικετών* (Label Powerset), ορίζει ένα πρόβλημα κατηγοριοποίησης μονής ετικέτας, στο οποίο το πεδίο ορισμού των (μονών) ετικετών ταυτίζεται με το σύνολο των διακριτών (και υπαρκτών) συνδυασμών ετικετών στα πολυκατηγορικά δεδομένα εκπαίδευσης. Παρ' όλο που ο μετασχηματισμός LC λαμβάνει ευθέως υπόψη του τις συσχετίσεις μεταξύ των ετικετών, μπορεί να κατηγοριοποιήσει νέα δείγματα μόνο με βάση σύνολα ετικετών τα οποία έχει ήδη συναντήσει στο σύνολο εκπαίδευσης, γεγονός που τον καθιστά επιρρεπή στην υπερ-εκπαίδευση (overfitting). Έχει, επίσης μεγάλη υπολογιστική πολυπλοκότητα [TV07, RPH08], αφού απαιτεί στο μετασχηματισμένο πρόβλημα απλής κατηγοριοποίησης αριθμό κλάσεων ίσο με τους διακριτούς συνδυασμούς ετικετών στα δεδομένα εκπαίδευσης (δηλαδή $min(|Y|), 2^{|L|-1})$), ενώ συχνά πρέπει να αντιμετωπίσει και εξαιρετικά σπάνιους συνδυασμούς ετικετών, από την άποψη θετικών δειγμάτων εκπαίδευσης.

RT - Κατάταξης και Κατωφλίου (Ranking and Threshold)

Ο μετασχηματισμός RT ορίζει ένα πρόβλημα κατηγοριοποίησης πολλών κλάσεων (multi-class). Για κάθε πολυκατηγορικό δείγμα της μορφής (x, Y), όπου $Y \subseteq L$, δημιουργούνται $|Y|$ νέα δείγματα (x, l_i), ένα για κάθε ετικέτα $l_i \in Y$. Στη συνέχεια, εκπαιδεύεται ένας ταξινομητής για να εξάγει την πιθανότητα ένα δείγμα να ανήκει σε κάθε μία από τις ετικέτες του συνόλου L με βάση το μετασχηματισμένο σύνολο δειγμάτων και χρησιμοποιείται μία συνάρτηση κατωφλίου (Παρ. 2.5.6) για την παροχή πολυκατηγορικών προβλέψεων [TK07]. Πέρα από την αδυναμία μοντελοποίησης των συσχετίσεων των ετικετών, κατά την εφαρμογή του μετασχηματισμού RT μπορεί να παρουσιαστεί δυσκολία στον ορισμό κατάλληλης τιμής κατωφλίου.

CC - Αλυσίδων Ταξινομητών (Classifier Chains)

Το πολυκατηγορικό πρόβλημα διασπάται σε $|L|$ δυαδικά προβλήματα, για καθένα από τα οποία εκπαιδεύεται ένας ταξινομητής. Οι ταξινομητές συνδέονται σε σειρά, έτσι ώστε τα γνωρίσματα του δείγματος x του ταξινομητή cl_i μαζί με την απόφασή $\hat{l}_i$ να γίνονται είσοδοι για τον ταξινομητή cl_{i+1}, καταλήγοντας στη μορφή $[x, \hat{l}_1, \ldots, \hat{l}_{k-1}]$ για τον ταξινομητή k. Ο μετασχηματισμός CC καταφέρνει να λάβει υπόψη του τις συσχετίσεις ανάμεσα στις ετικέτες με το τίμημα, όμως, της αυξημένης υπολογιστικής πολυπλοκότητας [RPHF09].

Με βάση τις παραπάνω μεθόδους, έχουν αναπτυχθεί αρκετές παραλλαγές που προσπαθούν να αντιμετωπίσουν τα προβλήματα που εντοπίζονται σε κάθε περίπτωση, όπως η χρονική ή/και χωρική πολυπλοκότητα και η αδυναμία μοντελοποίησης των συσχετίσεων μεταξύ των ετικετών. Περισσότερα για τις παραλλαγές αυτές μπορεί να διαβάσει κανείς στα [TV07, RPH08].

Προσαρμογή Αλγορίθμων

Πέρα από τις μεθόδους μετασχηματισμού πολυκατηγορικών προβλημάτων, έχουν υπάρξει και αρκετές προσπάθειες στην κατεύθυνση της επέκτασης κλασικών μεθόδων μηχανικής μάθησης, ώστε να γίνουν απευθείας εφαρμόσιμες σε προβλήματα πολυκατηγορικής ταξινόμησης. Μεταξύ των προσπαθειών αυτών, ξεχωρίζουν οι ακόλουθες:

Πολυκατηγορικά Δέντρα Απόφασης

Τα *δέντρα απόφασης* έχουν τροποποιηθεί [CK01], ώστε στα φύλλα τους να έχουν διανύσματα ετικετών και, έτσι, να πραγματοποιούν απευθείας πολυκατηγορική ταξινόμηση. Είναι ιδιαίτερα δημοφιλή στον τομέα της Βιοπληροφορικής, λόγω της υψηλής ερμηνευσιμότητάς τους.

Πολυκατηγορικοί Ταξινομητές Συνόλων

Η *Ενίσχυση* (Boosting), και άλλες μέθοδοι ανάπτυξης ταξινομητών συνόλων, είναι δυνατό να χρησιμοποιηθούν σε συνδυασμό με κάποια από τις μεθόδους μετασχηματισμού προβλημάτων, ώστε να παρέχουν πολυκατηγορική ταξινόμηση χρησιμοποιώντας θετική και αρνητική ψηφοφορία [SS00].

Πολυκατηγορικοί Ταξινομητές k Πλησιέστερων Δειγμάτων

Όπως ακριβώς γίνεται και στην απλή κατηγοριοποίηση, επιλέγονται οι k γείτονες ενός δείγματος, αλλά τώρα ο ταξινομητής χρησιμοποιεί το σύνολο των ετικετών τους για την πρόβλεψη των ετικετών του αγνώστου δείγματος [ZZ07].

Πολυκατηγορικοί Ταξινομητές Πιθανοτήτων

Αποτελούν γενίκευση των *απλών πιθανοτικών ταξινομητών*. Υπολογίζουν ένα σύνολο εκ των υστέρων πιθανοτήτων για τους συνδυασμούς των ετικετών [ZZ10].

Πολυκατηγορικά Νευρωνικά Δίκτυα

Εκτός από τη χρήση τους σε συνδυασμό με μεθόδους μετασχηματισμού προβλημάτων, υπάρχουν αρκετές προσαρμογές των ίδιων των νευρωνικών δικτύων για την απευθείας αντιμετώπιση πολυκατηγορικών προβλημάτων [ZZ06].

Συσχετίσεις Ετικετών

Στο πλαίσιο της πολυκατηγορικής ταξινόμησης, η μάθηση εξαρτάται ευθέως από τις συσχετίσεις ανάμεσα στις ετικέτες του προβλήματος, οι οποίες μπορεί να εμφανίζονται με διαφορετική συχνότητα. Μία κινηματογραφική ταινία με θέμα

την πολιτική είναι πιθανότερο να σχετίζεται με την κοινωνία και το έγκλημα και λιγότερο πιθανό να σχετίζεται με την κωμωδία. Τυπικά, μία ετικέτα y_i συσχετίζεται με την ετικέτα y_j αν και μόνο εάν

$$|P(y_i \mid y_j) - P(y_i)| > \epsilon \tag{2.14}$$

όπου ϵ κάποια σταθερά ικανά μεγάλη, ώστε να εγγυάται πως στατιστικά ισχύει

$$P(y_i \mid y_j) \neq P(y_i) \tag{2.15}$$

Στην περίπτωση που σε ένα σύνολο δεδομένων δεν παρατηρούνται συσχετίσεις ανάμεσα στις ετικέτες των δειγμάτων του, το πολυκατηγορικό πρόβλημα μπορεί να διασπαστεί σε τόσα προβλήματα απλής κατηγοριοποίησης όσες είναι και οι ετικέτες.

Οι ερευνητές, για να οπτικοποιήσουν τις συσχετίσεις ανάμεσα στις ετικέτες, έχουν ακολουθήσει διάφορες προσεγγίσεις, ανάμεσα στις οποίες οι πιο δημοφιλείς είναι οι *Χάρτες θερμότητας* και οι *Γράφοι Συσχετίσεων*.

Χάρτες Θερμότητας (Heatmaps)

Για $|L|$ ετικέτες, χρησιμοποιείται ένας διδιάστατος πίνακας M, διαστάσεων $|L| \times |L|$. Το στοιχείο M_{ij} απεικονίζει κάποια απόχρωση του χρώματος γκρι, ανάλογα με το βαθμό συσχέτισης των ετικετών i και j και, συγκεκριμένα, την πιθανότητα εμφάνισης της ετικέτας i δεδομένης της εμφάνισης της j:

$$M_{ij} = P(y_i \mid y_j) = \frac{P(y_i \cap y_j)}{P(y_j)} \tag{2.16}$$

και

$$M_{ii} = P(y_i) \tag{2.17}$$

Παραδείγματα χαρτών θερμότητας, για τα σύνολα πολυκατηγορικών δεδομένων genbase και enron, φαίνονται στα Σχήματα 2.1 και 2.2, αντίστοιχα.

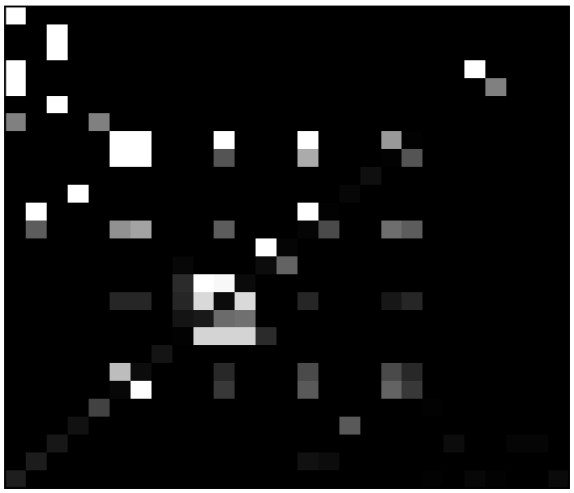
Γράφοι Συσχετίσεων

Αναπαριστούν δυαδικές συσχετίσεις μεταξύ των ετικετών. Κάθε ετικέτα αναπαρίσταται ως ένας κόμβος και το πάχος της ακμής που ενώνει δύο ετικέτες είναι ανάλογο της πιθανότητας $P(y_i \cap y_j)$.

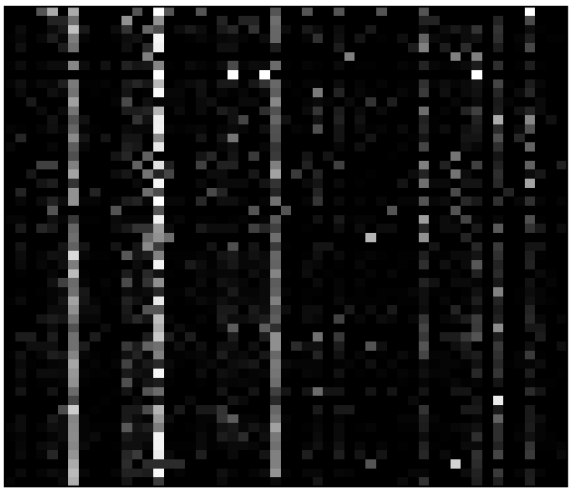
Ο γράφος συσχετίσεων των ετικετών των δειγμάτων του συνόλου scene παρουσιάζεται στο Σχήμα 2.3.

Ιδιότητες Πολυκατηγορικών Συνόλων Δεδομένων

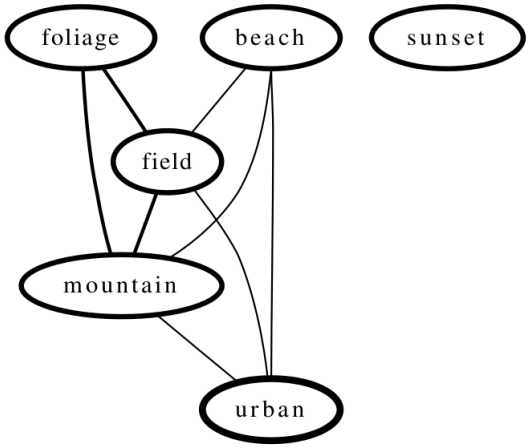
Η νέα διάσταση του χώρου των ετικετών καθιστά τα διάφορα πολυκατηγορικά σύνολα δεδομένων διαφορετικά ως προς τα εσωτερικά τους χαρακτηριστικά. Δεδομένου ότι μερικά από αυτά τα χαρακτηριστικά είναι μετρήσιμα, αναφέρουμε στη συνέχεια κάποιες σχετικές, βασικές μετρικές. Στους παρακάτω ορισμούς θεωρούμε ως D το σύνολο δεδομένων, L το σύνολο ετικετών, Y το σύνολο ετικετών με το οποίο συνδέεται ένα τυχαίο δείγμα $s \in D$ και X το σύνολο των γνωρισμάτων του D.



Σχήμα 2.1: Χάρτης Θερμότητας του συνόλου δεδομένων genbase.



Σχήμα 2.2: Χάρτης Θερμότητας του συνόλου δεδομένων enron.



Σχήμα 2.3: Γράφος Συσχέτισης του συνόλου δεδομένων scene.

Κατηγορική Πληθικότητα (Label Cardinality - LC)

Η *Κατηγορική Πληθικότητα* ενός συνόλου δεδομένων είναι ο μέσος όρος των ετικετών ανά δείγμα.

$$LC(D) = \frac{1}{|D|} \sum_{i=1}^{|D|} |Y_i| \tag{2.18}$$

Κατηγορική Πυκνότητα (Label Density - LD)

Η *Κατηγορική Πυκνότητα* είναι το μέσο ποσοστό ετικετών σε ένα σύνολο δεδομένων, σε σχέση με το πλήθος των ετικετών.

$$LD(D) = \frac{LC(D)}{|L|} = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i|}{|L|} \tag{2.19}$$

Ποσοστό Διακριτών Συνδυασμών Ετικετών

Είναι το ποσοστό των μοναδικών συνδυασμών ετικετών που υπάρχουν στο σύνολο των δειγμάτων:

$$P_{\text{Dist}}(D) = \frac{|S|\exists(x,S) \in D|}{|D|} \tag{2.20}$$

Ποσοστό Εμφάνισης Συχνότερου Συνδυασμού Ετικετών

Είναι ο λόγος του αριθμού των εμφανίσεων του συχνότερου συνδυασμού ετικετών προς το συνολικό αριθμό των δειγμάτων.

$$P_{\text{max}}(D) = \max_{Y|(x,Y)\in D} \frac{count(Y,D)}{|D|} \tag{2.21}$$

Υψηλή τιμή του P_{max} σε συνδυασμό με υψηλή τιμή του P_{Dist} είναι ενδεικτικές ασυμμετρίας στην κατανομή των ετικετών του D, αντίστοιχα προς το πρόβλημα ασυμμετρίας κατανομής κλάσεων στα προβλήματα απλής κατηγοριοποίησης.

Πολυπλοκότητα Συνόλου Δεδομένων (Dataset Complexity)

ορίζεται ως το γινόμενο του αριθμού των δειγμάτων του D, επί το σύνολο του αριθμού των ετικετών $|L|$, επί τον αριθμό των γνωρισμάτων $|X|$:

$$\text{Complexity} = |D| \times |L| \times |X| \tag{2.22}$$

Αξιολόγηση Πολυκατηγορικών Ταξινομητών

Σε αντιδιαστολή με την απλή ταξινόμηση, η ύπαρξη της επιπλέον διάστασης των πολλαπλών ετικετών σημαίνει ότι η προσέγγιση σωστής/λανθασμένης κατηγοριοποίησης δεν είναι δυνατόν να αναπαραστήσει πλήρως την ποιότητα των ταξινομητών. Ένα δείγμα μπορεί να ταξινομείται ορθά σε μία κατηγορία, αλλά λανθασμένα σε κάποια άλλη. Υπάρχουν δύο δρόμοι που μπορούμε να ακολουθήσουμε: η *αξιολόγηση βάσει ετικετών* και η *αξιολόγηση βάσει συνόλου ετικετών*.

1. **Αξιολόγηση βάσει ετικετών (label-based)**

   Εξετάζει την κάθε ετικέτα ξεχωριστά, κάνοντας χρήση στην ουσία των αξιολογήσεων απλής ταξινόμησης. Αποτυγχάνει, όπως είναι κατανοητό, να λάβει υπόψη τις συσχετίσεις ανάμεσα στις ετικέτες και την πολυπλοκότητα του συνόλου δεδομένων, με αποτέλεσμα να είναι αρκετά επιεικής, ιδιαίτερα σε περιπτώσεις μικρού αριθμού ενεργοποιημένων ετικετών ανά δείγμα (χαμηλό label cardinality).

2. **Αξιολόγηση βάσει συνόλου ετικετών (labelset-based)**

   Σε αντίθεση με την αξιολόγηση βάσει ετικετών, εκτιμά καλύτερα την ικανότητα του ταξινομητή σε ένα πολυκατηγορικό πρόβλημα, λαμβάνοντας υπόψη τις συσχετίσεις μεταξύ των ετικετών, αλλά μπορεί να αποδειχθεί υπερβολικά αυστηρή, ειδικά σε περιπτώσεις υψηλού αριθμού ενεργοποιημένων ετικετών ανά δείγμα (υψηλό label cardinality).

Για την αξιολόγηση των πολυκατηγορικών ταξινομητών, μπορούμε να χρησιμοποιήσουμε τις, κατάλληλα τροποποιημένες [LZZ06], μετρικές που έχουμε ήδη αναφέρει (παρ. 2.4.2). Έστω L το σύνολο των ετικετών του συνόλου δεδομένων D, με δείγματα της μορφής (x_i, Y_i), όπου $i = 1 \dots |D|$, $Y_i \subseteq L$, και $\hat{Y}_i = H(x_i)$ η συνάρτηση πρόβλεψης του ταξινομητή. Τότε, ορίζουμε τις εξής μετρικές:

Ακριβής Ορθότητα (Exact Match)

Είναι η απλούστερη και αυστηρότερη μετρική. Υπολογίζεται ως το ποσοστό των δειγμάτων που ταξινομήθηκαν ορθά για όλες τους τις κατηγορίες προς όλες τις ταξινομήσεις που έγιναν:

$$\textsc{ExactMatch} = \frac{|C|}{|D|} \tag{2.23}$$

όπου C το σύνολο των δειγμάτων που ταξινομήθηκαν απολύτως ορθά, σε πλήρη αντιστοιχία με τις πραγματικές τους ετικέτες, δηλαδή $Y_i \equiv \hat{Y}_i$

Ακρίβεια (Accuracy)

Είναι ο μέσος όρος, υπολογισμένος με βάση όλα τα δείγματα, των λόγων του μεγέθους του συνόλου τομής προς αυτό της ένωσης των προβλεπόμενων και πραγματικών ετικετών:

$$\textsc{Accuracy}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap \hat{Y}_i|}{|Y_i \bigcup \hat{Y}_i|} \tag{2.24}$$

Πιστότητα (Precision)

Είναι ο μέσος όρος, υπολογισμένος με βάση όλα τα δείγματα, των λόγων του μεγέθους του συνόλου τομής των προβλεπόμενων και πραγματικών ετικετών προς αυτό των προβλεπόμενων ετικετών:

$$\textsc{AveragePrecision}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap \hat{Y}_i|}{|\hat{Y}_i|} \tag{2.25}$$

Ανάκληση (Recall)

Είναι ο μέσος όρος, υπολογισμένος με βάση όλα τα δείγματα, των λόγων του μεγέθους του συνόλου τομής των προβλεπόμενων και πραγματικών ετικετών προς αυτό των πραγματικών ετικετών:

$$\textsc{AverageRecall}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \bigcap \hat{Y}_i|}{|Y_i|} \tag{2.26}$$

Απώλεια Hamming (Hamming Loss) Λαμβάνει υπόψη τις εσφαλμένα θετικές (FP) και αρνητικές (FN) προβλέψεις ετικετών του ταξινομητή:

$$\textsc{HammingLoss}(H, D) = \frac{1}{|D|} \sum_{i=1}^{|D|} \frac{|Y_i \Delta \hat{Y}_i|}{|L|} \tag{2.27}$$

όπου Δ ο τελεστής της συμμετρικής διαφοράς (XOR-symmetric difference) δύο συνόλων.

Από τις παραπάνω μετρικές, η Ακριβής Ορθότητα, η Ακρίβεια, η Πιστότητα και η Ανάκληση είναι βασισμένες σε σύνολα ετικετών (labelset-based), ενώ η Απώλεια Hamming είναι βασισμένη σε ετικέτες (label-based).

Στρατηγικές Συμπερασμού Πολυκατηγορικών Ταξινομητών

Ονομάζουμε *στρατηγική συμπερασμού* τη μέθοδο ταξινόμησης νέων, αγνώστων, δειγμάτων, δεδομένου ενός μοντέλου ταξινόμησης. Στη συνέχεια, χωρίς βλάβη της γενικότητας, υποθέτουμε ότι το χρησιμοποιούμενο μοντέλο ταξινόμησης βασίζεται σε κανόνες.

Στα πλαίσια της μονοκατηγορικής ταξινόμησης, η διαδικασία επιλογής κλάσης είναι αρκετά απλή και άμεση. Μπορεί να περιλαμβάνει είτε επιλογή της απόφασης του βέλτιστου (ως προς κάποια μετρική αξιολόγησης) κανόνα που καλύπτει το δεδομένο δείγμα, ή ψηφοφορία μεταξύ όλων των κανόνων που το καλύπτουν. Στα πλαίσια, όμως, της πολυκατηγορικής ταξινόμησης, όπου οι ετικέτες δεν είναι αμοιβαίως αποκλειόμενες και ο αριθμός των ενεργοποιημένων ετικετών δεν είναι εκ των προτέρων γνωστός, η στρατηγική συμπερασμού καθίσταται πολυπλοκότερη, ενώ υπάρχουν εναλλακτικές στρατηγικές που μπορούν να ακολουθηθούν. Στη συνέχεια εξετάζουμε από πιο κοντά τις δύο στρατηγικές που χρησιμοποιούνται σε αυτή την εργασία.

1. **Επιλογή βέλτιστου κανόνα**. Με την είσοδο ενός μη επισημασμένου δείγματος, συγκεντρώνονται οι κανόνες που το καλύπτουν. Κάθε κανόνας χαρακτηρίζεται από την καταλληλότητά του (την τιμή μίας συγκεκριμένης μετρικής ή κάποιας συνάρτησης ενός ή περισσοτέρων μετρικών), πάνω στην οποία βασίζεται αυτή η μέθοδος. Απόφαση για την κάθε ετικέτα του δείγματος λαμβάνει ο πλέον κατάλληλος κανόνας από το σύνολο των κανόνων που το καλύπτουν. Στην περίπτωση που ο εν λόγω κανόνας αδιαφορεί για κάποια ετικέτα, δηλαδή δεν είναι σε θέση να παρέχει κάποια συγκεκριμένη απόφαση για αυτή, η απόφαση λαμβάνεται από τον κανόνα με την αμέσως μικρότερη καταλληλότητα κ.ο.κ.

2. **Ψηφοφορία Μέσου όρου**. Μετά τη συγκέντρωση των κανόνων που καλύπτουν το μη επισημασμένο δείγμα, ο καθένας ψηφίζει θετικά ή αρνητικά, σταθμισμένα ανάλογα με την καταλληλότητά του, για την κάθε ετικέτα, (απέχοντας από την ψηφοφορία στην περίπτωση που αδιαφορεί), αναλόγως με το τί προβλέπεται στο τμήμα της απόφασής του. Η τελική απόφαση λαμβάνεται με βάση το μέσο όρο των ψήφων για την κάθε ετικέτα. Επιπλέον, μπορεί να χρησιμοποιείται μία συνάρτηση κατωφλίου για το διαχωρισμό των ετικετών σε αυτές που είναι σχετικές και άσχετες με το δείγμα, όπως π.χ. στον αλγόριθμο RAkEL [TV07].

Όπως έχει ήδη αναφερθεί, η επιλογή του κατωφλίου είναι σε κάποιες περιπτώσεις πολύ σημαντική, καθώς επηρεάζει άμεσα την ικανότητα κατηγοριοποίησης του ταξινομητή. Μία συνάρτηση κατωφλίου έχει τη δομή

$$\hat{y}_i = f_{L,t}(w) = \begin{cases} 1 & w_i \geq t_i \\ 0 & w_i < t_i \end{cases} \tag{2.28}$$

όπου $\hat{y}_i$ η πρόβλεψη του ταξινομητή για την ύπαρξη της ετικέτας i και t_i το κατώφλι για αυτή την ετικέτα. Για τη διευκόλυνση επιλογής μία συγκεκριμένης τιμής κατωφλίου, το διάνυσμα βεβαιοτήτων w κανονικοποιείται με αποτέλεσμα τον περιορισμό των τιμών του στο διάστημα $[0, 0.5]$. Τέλος, οι περισσότερες προσεγγίσεις περιλαμβάνουν μία διαδικασία ρύθμισης του κατωφλίου, είτε εσωτερικά είτε εξωτερικά.

Η *Επιλογή με Εσωτερική Αξιολόγηση (Internal Validation - IVAL)* προσπαθεί να ορίσει μία τιμή κατωφλίου τέτοια ώστε να μεγιστοποιεί μία συγκεκριμένη μετρική, με διαδοχικούς εσωτερικούς ελέγχους ως προς τη μετρική αυτή. Η διαδικασία αυτή είναι χρονικά απαιτητική λόγω των επαναλαμβανόμενων αξιολογήσεων, ωστόσο μπορεί να βελτιωθεί από άποψη υπολογιστικής πολυπλοκότητας αξιοποιώντας το γεγονός ότι οι περισσότερες συναρτήσεις μετρικών με μεταβλητή το ίδιο το κατώφλι είναι κυρτές. Αυτό συμβαίνει αφού για μηδενικό κατώφλι, επιλέγονται πολλές ετικέτες, ενώ, καθώς αυτό αυξάνεται, αυξάνει και η ειδικότητα, αφού επιλέγονται λιγότερες ετικέτες. Προφανώς, οι μετρικές της Ακρίβειας και της Ακριβούς Ορθότητας, αρχικά αυξάνονται και στη συνέχεια μειώνονται ως κυρτές συναρτήσεις.

Η *Επιλογή με Ποσοστιαία Αποκοπή (Proportional Cut - PCUT)* ρυθμίζει εξωτερικά το κατώφλι, και προσεγγίζει την τιμή του επαναληπτικά, στοχεύοντας στην ελαχιστοποίηση της διαφοράς της πολυκατηγορικής πληθικότητας LC σε ένα σύνολο δεδομένων. Λεπτομέρειες για τη μέθοδο με την οποία γίνεται η ψηφοφορία και για τις μεθόδους ρύθμισης του κατωφλίου αναφέρουμε στην Παρ. 5.4.

3
Γενετικοί Αλγόριθμοι

Οι Γενετικοί Αλγόριθμοι (ΓΑ) είναι μέθοδοι εξερεύνησης, βελτιστοποίησης και μηχανικής μάθησης. Εμπνευσμένοι από τη Θεωρία της Εξέλιξης των Ειδών που θεμελίωσε ο Κ. Δαρβίνος, κάνουν χρήση της εξελικτικής διαδικασίας για την επίλυση υπολογιστικών προβλημάτων και την αναζήτηση ολικά βέλτιστων λύσεων, όπως ακριβώς και τα υπόλοιπα είδη αλγορίθμων του τομέα της Εξελικτικής Υπολογιστικής: ο Γενετικός Προγραμματισμός και οι Εξελικτικοί Αλγόριθμοι.

Βιολογια και Γενετικοι Αλγοριθμοι

Οι πρώτες εξελικτικές θεωρίες αναδύθηκαν στις αρχές του 19ου αιώνα, υποστηρίζοντας ότι τα είδη του φυσικού κόσμου συμμετέχουν και έχουν προκύψει από μία διαδικασία συνεχούς και αέναης εξέλιξης. Βασιζόμενοι στο έργο του Jean Batista Lamark που απέρριπτε την ιδέα ότι οι μορφές ζωής είναι αμετάβλητες, οι Darwin και Wallace, ανεξάρτητα ο ένας από τον άλλον, ανέπτυξαν την επαναστατική ιδέα της φυσικής επιλογής, όπως αυτή εκφράζεται στο βιβλίο του πρώτου "*The origin of species*".

Η εξελικτική θεωρία υποστηρίζει την ύπαρξη ενός *γενοτύπου* στον κάθε οργανισμό, δηλαδή μίας σαφώς ορισμένης κατάστασης ή διάταξης των τιμών του γονιδιώματος στο εσωτερικό του κάθε οργανισμού. Υπενθυμίζουμε πως γονιδίωμα ονομάζουμε το σύνολο των χρωμοσωμάτων ενός οργανισμού σε ένα κύτταρο του. Η μετάφραση του γενοτύπου σε επίπεδο φυσικού κόσμου, δηλαδή τα φυσικά χαρακτηριστικά γνωρίσματα, οι ικανότητες, οι αδυναμίες και τα προτερήματα ενός οργανισμού, συνιστούν το *φαινότυπο* του. Στη Φύση, οι γενότυποι ενός είδους διαφέρουν σημαντικά ανάμεσα στα άτομα που το συνιστούν[1], ως αποτέλεσμα της

[1]Εγκυκλοπαιδικά, ο άνθρωπος διαφέρει σημαντικά από την πλειονότητα των άλλων ειδών στη Γη, καθώς κάθε άτομο διαφέρει λιγότερο από ένα οποιοδήποτε άλλο σε σχέση με άλλα είδη. Αυτό είναι το αποτέλεσμα μίας πληθυσμιακής συμφόρησης (bottleneck) στην Ιστορία εξέλιξης του ανθρώ-

εξελικτικής διαδικασίας, μέσω της οποίας τα άτομα τα οποία είναι καταλληλότερα για επιβίωση στο περιβάλλον τους, θα αναπαράγονται με μεγαλύτερη πιθανότητα[2], κληροδοτώντας στα παιδιά τους μέρος των γονιδίων τους και, έτσι, πολλαπλασιάζοντάς τα (survival of the fittest). Παράλληλα με τη διαδικασία *διασταύρωσης* των γονιδίων ανάμεσα στους αρσενικούς και θηλυκούς προγόνους, εισάγονται, τυχαία, νέα χαρακτηριστικά στον πληθυσμό του είδους με το μηχανισμό της *μετάλλαξης*.

Η παραπάνω περιγραφόμενη διαδικασία αποτελεί και το πλαίσιο λειτουργίας των Γενετικών Αλγορίθμων. Ένας Γενετικός Αλγόριθμος είναι μία συνεχής διαδικασία που μεταχειρίζεται έναν *πληθυσμό ατόμων* και στα πλαίσια της οποίας τα άτομα:

1. *αναπαράγονται*, μεταφέροντας το γενότυπό τους στους απογόνους τους,
2. *μεταλλάσσονται*, με καθαρά τυχαίο τρόπο,
3. *ανταγωνίζονται*, καθώς ζουν σε περιβάλλοντα με περιορισμένους πόρους, και
4. *επιλέγονται* (για αναπαραγωγή) ως αναπόφευκτο αποτέλεσμα του ανταγωνισμού τους για τους παραπάνω πόρους.

Τα άτομα του πληθυσμού σε έναν ΓΑ αναπαριστούν μία πιθανή λύση στο υποκείμενο πρόβλημα. Κάθε άτομο αποτελείται από μία διάταξη *γονιδίων*, το καθένα από τα οποία εκφράζει και ένα χαρακτηριστικό του ατόμου. Η σύνθεση των γονιδίων σε μία ενότητα μέσα στο άτομο είναι το *χρωμόσωμά* του. Για την ευκολότερη αναπαράσταση και επεξεργασία από τους υπολογιστές, τα γονίδια είναι κωδικοποιημένα δυαδικά, με σημασία ανάλογη προς το πρόβλημα προς επίλυση. Ωστόσο, οι γενικές αρχές των Γενετικών Αλγορίθμων είναι ανεξάρτητες από τον τρόπο αναπαράστασης των γονιδίων και του χρωμοσώματος, αλλά και του ίδιου το προβλήματος. Εδώ ακριβώς φανερώνεται και η αρετή των Γενετικών Αλγορίθμων στην επίλυση δυσνόητων (ή ακόμα και μη κατανοητών) προβλημάτων. Ο προγραμματιστής δεν χρειάζεται να γνωρίζει τις λεπτομέρειες των εσωτερικών χαρακτηριστικών του προβλήματος. Αρκεί να μπορεί να αναπαραστήσει τη δομή του *χώρου αναζήτησης*[3] και να μπορεί να αξιολογήσει τα άτομα - λύσεις του προβλήματος.

Για την αξιολόγηση της ποιότητας κάθε ατόμου - λύσης, πάνω στην οποία βασίζεται η υλοποίηση της φυσικής επιλογής, χρησιμοποιείται μία συνάρτηση αξιολόγησης της ποιότητας, ή καταλληλότητας, κάθε ατόμου. Με άλλα λόγια, η *καταλληλότητα* (fitness) είναι ένα μέτρο της ικανότητας του ατόμου για επίλυση του

πινου είδους και οφείλεται στην υπερηφαιστιακή έκρηξη στη σημερινή λίμνη Toba στη Σουμάτρα Ινδονησίας, στα 73000 ± 4000 Π.Χ. Ο δεκαετής χειμώνας που προέκυψε από την κάλυψη του ουρανού από ηφαιστειακή τέφρα έφερε το ανθρώπινο είδος, που μέχρι εκείνη την περίοδο πιστεύεται ότι εδραζόταν στην Ανατολική Αφρική, στα πρόθυρα της ολοκληρωτικής εξάλειψης, μειώνοντας τον πληθυσμό του σε μερικές χιλιάδες άτομα (3.000 - 10.000). Η γονιδιακή δεξαμενή (genetic pool) συρρικνώθηκε σε τεράστιο βαθμό, στα άτομα που ήταν πλέον κατάλληλα για επιβίωση στις ακραίες εκείνες συνθήκες.

[2]Τα άτομα που είναι καταλληλότερα για επιβίωση στο περιβάλλον τους, αναπαράγονται με μεγαλύτερη πιθανότητα καθώς, ενδογενώς, σκοπός του κάθε είδους είναι η διαιώνισή του. Κάθε άτομο θα επιλέξει ένα ταίρι του αντίθετου φύλλου για αναπαραγωγή, τέτοιο που να του εξασφαλίζει πρωτίστως την επιβίωση των απογόνων του και δευτερευόντως την ευημερία του.

[3]Ορίζουμε ως *Χώρο Αναζήτησης* ενός προβλήματος το σύνολο όλων των δυνατών και έγκυρων λύσεων, μέσα στο οποίο ανήκουν και οι λύσεις του προβλήματος.

δεδομένου προβλήματος ή ενός μέρους αυτού. Καθώς η καταλληλότητα είναι η μοναδική μετρική με την οποία συγκρίνονται τα άτομα, η μέθοδος υπολογισμού της είναι κρίσιμης σημασίας. Η συνάρτηση αξιολόγησης, επομένως, θα πρέπει να διαχωρίζει με επιτυχία καλές από κακές λύσεις, αλλά και να δημιουργεί έμμεσα αποτελεσματική πίεση (fitness pressure) προς την αναπαραγωγή και επιβίωση των ατόμων που αποτελούν τις καταλληλότερες λύσεις για το πρόβλημα.

Τέλος, αξίζει να παρατηρήσουμε τη στοχαστική φύση των εξελικτικής διαδικασίας. Πρώτον, όπως αναφέραμε, η μετάλλαξη είναι ένα καθαρά στοχαστικό γενετικό συμβάν, καθώς, στην πράξη, σφάλματα στη μεταφορά της γενετικής πληροφορίας είναι αναπόδραστα και απρόβλεπτα. Δεύτερον, και η ίδια η επιλογή δεν είναι αιτιοκρατική. Η ποιότητα, ή καταλληλότητα, του ατόμου είναι μεν η σημαντικότερη του παράμετρος, γιατί είναι μέτρο της ικανότητάς του για επιβίωση, υπάρχουν όμως πολλοί εξωτερικοί παράγοντες που μπορούν να μεταβάλλουν τη διαδικασία επιλογής. Μία ενδιαφέρουσα παράμετρος, που όμως ξεφεύγει από τα πλαίσια αυτής της εργασίας, είναι οι μετατοπίσεις του ορισμού της *καταλληλότητας* από κάθε κοινωνία στο πέρασμα των αιώνων. Κάθε περίοδος επιτάσσει και τα δικά της κριτήρια για το τί συνιστά την καταλληλότητα, ανάλογα και με την ανθρώπινη πρόοδο και τις πτυχές της. Εβδομήντα χιλιάδες χρόνια πριν, κατάλληλος θεωρείτο αυτός που μπορούσε να αψηφίσει το εχθρικό περιβάλλον με τις φυσικές και πνευματικές του δυνάμεις και να υποτάξει τη φύση για το καλό της "αγέλης"[4]. Στην Αρχαία Αθήνα η καταλληλότητα είχε διαφορετικό ορισμό. Το ίδιο και μετά από κάθε κοινωνικό, οικονομικό ή πολιτισμικό paradigm shift.

Αλγοριθμος Εξελιξης

Μια γενική μορφή της διαδικασίας εξέλιξης για έναν απλό ΓΑ φαίνεται στον Αλγόριθμο 3.1.

Αλγόριθμος 3.1 Γενική Μορφή Γενετικού Αλγορίθμου.

```
1: geneticAlgorithm()
2: t ← 0
3: P(t) ← initializePopulation()
4: while terminationConditionNotMet do
5:     t ← t + 1
6:     P(t) ← Evaluate(P(t))
7:     P'(t) ← SelectParentsFrom(P(t − 1))
8:     P'(t) ← Crossover(P'(t))
9:     P'(t) ← Mutate(P'(t))
10:    P(t) ← P'(t)
11: end while
```

[4]Παλαιότερα, και μέχρι την ανάπτυξη του εμπορίου μέσω μη χρηματιστικών συναλλαγών, απευθείας ανταλλαγών δηλαδή, ο άνθρωπος συγκροτούσε κοινωνίες μέχρι 150 άτομα.

Ο πληθυσμός των ατόμων ενός Γενετικού Αλγορίθμου εξελίσσεται μέσα από μία επαναληπτική διαδικασία *επιλογής - διασταύρωσης - μετάλλαξης - αντικατάστασης* των ατόμων.

Σε πρώτο στάδιο, ο πληθυσμός αρχικοποιείται, είτε τυχαία, είτε με βάση υπάρχουσα γνώση πάνω στο πρόβλημα που μεταφράζεται (στο πεδίο λύσεων) σε άτομα. Σε δεύτερο στάδιο, στην αρχή κάθε επανάληψης, αποτιμάται η καταλληλότητα των ατόμων του πληθυσμού, μέσω της συνάρτησης αξιολόγησης. Στη συνέχεια, ακολουθεί η εφαρμογή των τελεστών διασταύρωσης και μετάλλαξης, με αποτέλεσμα έναν πληθυσμό απογόνων που αντικαθιστά τον τρέχοντα πληθυσμό πριν την επόμενη επανάληψη.

Αναλύουμε παρακάτω τις διαφορετικές μεθόδους που μπορούν να ακολουθηθούν σε σχέση με τη φυσική επιλογή, και τους τελεστές διασταύρωσης και μετάλλαξης.

Διαδικασια Φυσικης Επιλογης

Η διαδικασία της Φυσικής Επιλογής προσομοιώνει το φυσικό μηχανισμό της επιβίωσης του ικανοτέρου: είναι ο μηχανισμός που καθορίζει τον τρόπο με τον οποίο επιλέγονται τα καταλληλότερα άτομα του πληθυσμού για αναπαραγωγή, ώστε να αποτελέσουν τη γενετική βάση του πληθυσμού της επόμενης γενιάς.

Για τη δημιουργία του νέου πληθυσμού υπάρχουν δύο βασικές μέθοδοι [Buc02]:

1. **Ελιτισμός (Elitism)** είναι η διαδικασία κατά την οποία ένα κομμάτι του πληθυσμού, η επονομαζόμενη ελίτ, θα διοχετευθεί αυτούσιο στο νέο πληθυσμό, λόγω των καλών χαρακτηριστικών των ατόμων της, ενώ το υπόλοιπο κομμάτι του νέου πληθυσμού θα προκύψει μέσα από αναπαραγωγή των κατάλληλων γονέων του προηγούμενου πληθυσμού.

2. **Επιλογή Σταθερής Κατάστασης (Steady-State Selection)** είναι μία διαδικασία αντίθετη προς τον ελιτισμό. Το μεγαλύτερο κομμάτι του πληθυσμού μεταφέρεται αυτούσιο στο νέο πληθυσμό, ενώ κάποια άτομα (σχετικά λίγα), με χαρακτηριστικά χαμηλότερη καταλληλότητα, ή που πληρούν άλλα κριτήρια, αντικαθίστανται από νέα άτομα, προκύπτοντα από την αναπαραγωγή κατάλληλων γονέων.

Η επιλογή των ατόμων του πληθυσμού που θα αποτελέσουν τους γονείς, μέσω της αναπαραγωγής των οποίων θα προκύψουν τα άτομα - απόγονοι, μπορεί να συντελεστεί με μία πληθώρα τρόπων. Όλοι όμως οι τρόποι αυτοί βασίζονται στη γενική ιδέα της πόλωσης της διαδικασίας επιλογής προς άτομα με μεγαλύτερη καταλληλότητα. Μέθοδοι όπως η *επιλογή ρουλέτας* ή η *στοχαστική καθολική επιλογή* αναθέτουν σε κάθε άτομο του πληθυσμού μία πιθανότητα επιλογής ανάλογη προς την καταλληλότητά του. Άλλα σχήματα επιλογής, όπως η *επιλογή τουρνουά* ακολουθούν μία αιτιοκρατική προσέγγιση, κατατάσσοντας σε φθίνουσα σειρά καταλληλότητας τα άτομα του πληθυσμού και επιλέγοντας αυτά που βρίσκονται στις υψηλότερες θέσεις.

Καθώς αυτή η εργασία χρησιμοποιεί αποκλειστικά την επιλογή ρουλέτας, την εξετάζουμε πιο αναλυτικά στην επόμενη παράγραφο.

Επιλογή Ρουλέτας

Η επιλογή ρουλέτας υλοποιείται ως μια εξομοίωση πραγματικής ρουλέτας. Κάθε άτομο του πληθυσμού τοποθετείται σε ένα χωρίο της ρουλέτας, του οποίου το μέγεθος είναι ανάλογο της καταλληλότητας του ατόμου. Όσο περισσότερο ικανό είναι ένα άτομο i, τόσο μεγαλύτερο χωρίο θα πληρώσει, άρα και τόσο μεγαλύτερη πιθανότητα θα έχει να επιλεγεί. Για πληθυσμό n ατόμων, αυτή η πιθανότητα είναι ίση με

$$P(i) = \frac{fitness(i)}{\sum_{j=1}^{n} fitness(j)} \tag{3.1}$$

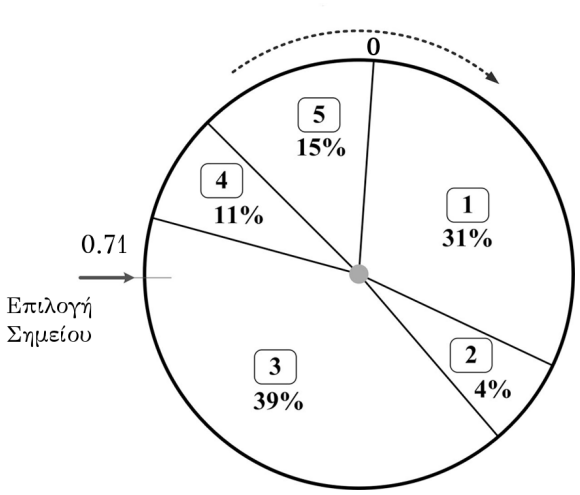
Στον Αλγόριθμο 3.2 παραθέτουμε τον αλγόριθμο επιλογής ρουλέτας σε μορφή ψευδοκώδικα, δεδομένου του πληθυσμού $[P]$.

Αλγόριθμος 3.2 Επιλογής Γονέα μέσω Επιλογής Ρουλέτας.

```
1: rouletteWheelSelection(P)
2: F ← sumOfIndividualsFitnessIn(P)
3: f ← random[0, F]
4: s ← 0
5: index ← 0
6: while s < f do
7:     index ← index + 1
8:     f(i) ← fitnessOfIndividualInSegment(i)
9:     s ← s + f(i)
10: end while
11: parent ← individualInSection(index)
```

Στην αρχή υπολογίζεται το άθροισμα των καταλληλοτήτων F των ατόμων που απαρτίζουν τον πληθυσμό, και στη συνέχεια, πολλαπλασιάζεται με έναν τυχαίο αριθμό στο διάστημα $[0, 1]$, ώστε να προκύψει ένας τυχαίος αριθμός f στο διάστημα $[0, F]$. Με αφετηρία το πρώτο άτομο, αθροίζουμε τις καταλληλότητες των ατόμων, σειριακά, έως ότου το άθροισμα ξεπεράσει τον αριθμό f. Το τρέχον άτομο θα αποτελέσει το ζητούμενο γονέα. Στο Σχήμα 3.1 παριστάνεται η παραπάνω διαδικασία για $f = 0.71 \cdot F$.

Μία πιθανή διαφορά επίδοσης του μηχανισμού της επιλογής θα υπήρχε εάν το σημείο της αφετηρίας για την άθροιση των καταλληλοτήτων επιλεγόταν τυχαία κάθε φορά ή αν ο πληθυσμός ταξινομούταν πριν από κάθε γενετική επιλογή γονέα, με φθίνουσα σειρά καταλληλότητας των ατόμων του. Η διαφορά οφείλεται στη διάταξη των ατόμων στον πληθυσμό, καθώς, θεωρητικά, τα μέλη του πληθυσμού που βρίσκονται στο τέλος του, είναι αυτά που έχουν προστεθεί τελευταία και είναι, εν γένει, περισσότερο κατάλληλα, λόγω της πίεσης προς αυξανόμενη καταλληλότητα (fitness pressure). Αυτή η προσέγγιση είναι σημαντικά περισσότερο κοστοβόρα υπο-



Σχήμα 3.1: Κατανομή ατόμων σε μία εικονική ρουλέτα.

λογιστικά, αφού ο πληθυσμός αποτελείται από αριθμό ατόμων ανάλογο προς την πολυπλοκότητα του προβλήματος και τα σύνολα πραγματικών δεδομένων απαιτούν μερικές χιλιάδες έως δεκάδες χιλιάδες άτομα.

Τελεστες Γενετικης Διασταυρωσης

Η γενετική διασταύρωση στοχεύει στην ανάδειξη ατόμων που αποτελούν καλύτερες λύσεις του προβλήματος από τους προγόνους τους, με τον εντοπισμό αρκούντως κατάλληλων γονέων και την ανταλλαγή γενετικού υλικού (γονιδίων, όπως τα ορίσαμε παραπάνω) ανάμεσα στους δύο γονείς - χρωμοσώματα. Στις περισσότερες εφαρμογές ο αριθμός των απογόνων είναι δύο.

Κάποιες συχνά χρησιμοποιούμενες μέθοδοι διασταύρωσης παρουσιάζονται παρακάτω:

Διασταύρωση ενός σημείου
Επιλέγεται τυχαία ένα σημείο, κατά το μήκος των χρωμοσωμάτων. Αριστερά του σημείου διασταύρωσης, τα γονίδια του πρώτου απογόνου θα ταυτίζονται με αυτά του πρώτου γονέα, ενώ δεξιά του σημείου διασταύρωσης θα ταυτίζονται με τα γονίδια του δεύτερου. Για το δεύτερο απόγονο ισχύει το αντίθετο.

Διασταύρωση δύο σημείων
Σε αυτή την περίπτωση επιλέγονται τυχαία δύο σημεία. Τα γονίδια δεξιά και αριστερά των σημείων διασταύρωσης μεταφέρονται αυτούσια από τους γονείς στους απογόνους, και αυτά μεταξύ των δύο σημείων εναλλάσσονται.

Ομοιόμορφη Διασταύρωση
Χρησιμοποιεί τόσα σημεία διασταύρωσης όσα γονίδια διαθέτει το χρωμόσωμα. Κάθε γονίδιο ενός απογόνου αντιγράφεται από τον έναν γονέα ή τον

άλλο, βάσει μιας δυαδικής μάσκας διασταύρωσης, η οποία δημιουργείται τυχαία για κάθε ζευγάρι γονέων, πριν ξεκινήσει η διαδικασία. Η μάσκα διασταύρωσης αποκτά την τιμή 1 σε μία θέση σύμφωνα με μία δοσμένη πιθανότητα ομοιόμορφης διασταύρωσης.

Εμπειρικά, η πιθανότητα διασταύρωσης τοποθετείται στο διάστημα $[0.7, 0, 9]$. Στην περίπτωση που δεν επιλεχθεί διασταύρωση, οι απόγονοι προκύπτουν ως αντίγραφα των γονέων. Στην Παρ. 6.2.1 εξετάζουμε μία τέταρτη μέθοδο, ειδικά για τη διασταύρωση κανόνων πολυκατηγορικής ταξινόμησης.

Τελεστες Γενετικης Μεταλλαξης

Μετά τη διασταύρωση εφαρμόζεται η γενετική μετάλλαξη. Παρά την πληθώρα υλοποιήσεων τελεστών μετάλλαξης, όλοι λειτουργούν με τον ίδιο περίπου τρόπο. Σε αντίθεση με τη γενετική διασταύρωση, δέχονται ως είσοδο ένα μόνο χρωμόσωμα, ενώ εισάγουν νέες, τυχαίες τιμές στα γονίδια των απογόνων (τιμές που διαφορετικά μπορεί να μην είχαν εμφανιστεί), συμβάλλοντας σημαντικά στη γενετική ποικιλομορφία του πληθυσμού. Επιπρόσθετα, η έρευνα του χώρου αναζήτησης γίνεται πιο αποτελεσματική, καθώς, μειώνεται η πιθανότητα ο πληθυσμός να μείνει στάσιμος σε μία υπό-βέλτιστη λύση. Η πιθανότητα μετάλλαξης είναι σημαντικά μικρότερη από την πιθανότητα διασταύρωσης στους γενετικούς αλγορίθμους που χρησιμοποιούν δυαδική κωδικοποίηση και τοποθετείται στο διάστημα $[0.01, 0.1]$.

Οι παραπάνω λειτουργίες της επιλογής, της διασταύρωσης και της μετάλλαξης, όταν εφαρμοστούν ξεχωριστά η καθεμία, έχουν αποδειχθεί αναποτελεσματικές [Gol02]. Όταν όμως συνδυαστούν, παράγουν χρήσιμα αποτελέσματα: ο συνδυασμός της επιλογής με τη διασταύρωση εισάγει μία λειτουργία καινοτομίας, ενώ ο συνδυασμός της μετάλλαξης με την επιλογή δημιουργεί γόνιμο έδαφος για συνεχή βελτίωση μέσω της τοπικής αναζήτησης.

Γενετικοι Αλγοριθμοι και Εξορυξη Δεδομενων

Οι Γενετικοί Αλγόριθμοι έχουν τρεις βασικές εφαρμογές στην Εξόρυξη Δεδομένων [Fre05]:

1. **Εύρεση Κανόνων Ταξινόμησης**
   Η αναζήτηση και εξέλιξη κανόνων ταξινόμησης για τη δημιουργία προβλεπτικών μοντέλων μπορεί να υλοποιηθεί με τη χρήση Γενετικών Αλγορίθμων.

2. **Ομαδοποίηση**
   Η ομαδοποίηση αναφέρεται στη δημιουργία ομάδων από τα δεδομένα και χρησιμοποιείται σε εργασίες περιγραφής δεδομένων.

3. **Προεπεξεργασία Δεδομένων**
   Οι Γενετικοί Αλγόριθμοι μπορούν να χρησιμοποιηθούν για την επιλογή ή σύνθεση γνωρισμάτων ενός συνόλου δεδομένων.

Καθώς η παρούσα εργασία χρησιμοποιεί τους Γενετικούς Αλγορίθμους αποκλειστικά για την εύρεση κανόνων ταξινόμησης, επεκτείνουμε την ανάλυση της εφαρμογής αυτής στην επόμενη ενότητα.

Αναπαραστάσεις Κανόνων ως Χρωμοσώματα

Για την αναπαράσταση των κανόνων σε μορφή χρωμοσώματος υπάρχουν δύο βασικές προσεγγίσεις.

1. **Αναπαράσταση Συνόλου Κανόνων ως Άτομο** (Pittsburg Approach)
   Ένα χρωμόσωμα αναπαριστά ένα σύνολο κανόνων, μη σταθερού μεγέθους, το οποίο αποτελεί και μία πλήρη λύση του προβλήματος. Με αυτή την προσέγγιση, λαμβάνονται υπόψη οι πιθανές αλληλεπιδράσεις των κανόνων, όμως δημιουργούνται χρωμοσώματα μεγάλου μεγέθους, καθιστώντας δυσχερή την κατασκευή αποτελεσματικών γενετικών τελεστών.

2. **Αναπαράσταση ενός Κανόνα ως Άτομο** (Michigan Approach)
   Κάθε κανόνας είναι και ένα χρωμόσωμα. Σε αντίθεση με την προσέγγιση Pittsburg, τα άτομα έχουν σταθερό μήκος, είναι μικρότερα σε μέγεθος και η κατασκευή των γενετικών τελεστών είναι σημαντικά ευκολότερη. Παρ' όλα αυτά, αυτή η μέθοδος αναπαράστασης δεν λαμβάνει ρητά υπόψη της τις πιθανές αλληλεπιδράσεις των κανόνων. Ο Γενετικός Αλγόριθμος, στην περίπτωση της προσέγγισης Michigan, προσπαθεί να κατασκευάσει ένα σύνολο κανόνων που "συνεργάζονται" ώστε όλοι μαζί να αποτελέσουν τη λύση του προβλήματος. Στην κατεύθυνση, λοιπόν, της διατήρησης της ποικιλότητας των κανόνων του πληθυσμού, χρησιμοποιούνται μέθοδοι όπως ο διαμοιρασμός καταλληλότητας, που τιμωρεί άτομα που ικανοποιούν μία ορισμένη συνθήκη ομοιότητας μεταξύ τους.

Η παρούσα εργασία χρησιμοποιεί αποκλειστικά την προσέγγιση Michigan και, συνεπώς, στα επόμενα κεφάλαια θα προϋποθέτουμε αυτή τη γνώση.

Περιορισμοί και Πλεονεκτήματα των Γενετικών Αλγορίθμων

Σε γενικές γραμμές, οι Γενετικοί Αλγόριθμοι εμφανίζουν χαρακτηριστικά που δεν απαντώνται στους υπόλοιπους αλγορίθμους αναζήτησης και βελτιστοποίησης, ειδικά σε αυτούς που χρησιμοποιούν άπληστες ή αιτιοκρατικές προσεγγίσεις.

Οι Γενετικοί Αλγόριθμοι, αν και έχουν σαφέστατα μεγαλύτερο χρόνο εκτέλεσης από τις υπόλοιπες τεχνικές εξόρυξης δεδομένων, διαθέτουν λειτουργίες που μπορούν εύκολα να παραλληλοποιηθούν. Λόγω της μη αιτιοκρατικής τους φύσης, μπορούν να ανακαλύψουν λύσεις δύσκολα εντοπίσιμες, να ανακαλύψουν σχετικά εύκολα βέλτιστες λύσεις, αλλά και να αποφύγουν μη βέλτιστες λύσεις. Δυστυχώς, όμως, δυσκολεύονται, να βρουν το ακριβές σημείο της βέλτιστης λύσης χωρίς χρήση μεθόδων τοπικής έρευνας. Είναι εύρωστοι, καθώς λειτουργούν καλά παρουσία θορύβου στα δεδομένα, δεν είναι ευαίσθητοι σε (μικρές) αλλαγές των παραμέτρων τους, η λειτουργία τους δεν επηρεάζεται από ασυνέχειες του χώρου αναζήτησης

και μπορούν να αποδώσουν το ίδιο καλά με άλλες τεχνικές σε προβλήματα ευρείας κλίμακας. Τέλος, μπορούν να εφαρμοσθούν για την επίλυση ποικίλων προβλημάτων αναζήτησης και βελτιστοποίησης, καθώς δεν απαιτούν την apriori κατανόηση της εσωτερικής δομής του προβλήματος-στόχου, και το μόνο που χρήζει αλλαγής είναι η αναπαράσταση των χρωμοσωμάτων και ο τρόπος αξιολόγησής τους. Υπάρχουν, βέβαια, και προβλήματα για τα οποία η κατασκευή μίας αποτελεσματικής αναπαράστασης χρωμοσωμάτων ή/και συνάρτησης αξιολόγησης είναι έργα από πολύπλοκα έως αδύνατα. Ο μεγάλος αριθμός των παραμέτρων τους και ο ακριβής χειρισμός τους, τέλος, είναι ένα δύσκολο καθήκον, ιδιαίτερα σε περιπτώσεις απουσίας εκ των προτέρων γνώσης του προβλήματος.

4

Μανθάνοντα Συστήματα Ταξινομητών για Μονοκατηγορική Ταξινόμηση

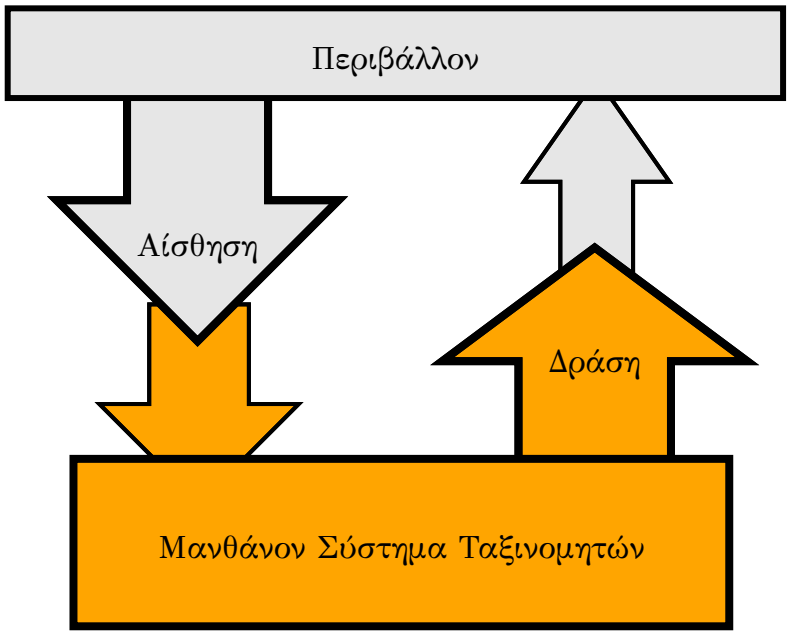
Επινοηθέντα το 1975 από τον John Holland [Hol75], τα Μανθάνοντα Συστήματα Ταξινομητών (ΜαΣΤ), αποτελούν μια προσέγγιση Μηχανικής Μάθησης Βασισμένη στη Γενετική. Χρησιμοποιούν κανόνες ταξινόμησης για την επίλυση προβλημάτων Μαρκοβιανής Απόφασης και Ταξινόμησης, παρέχοντας μία σειρά αρχών για online μηχανική μάθηση μέσω της προσαρμογής [HR78].

Γενικο Μοντελο

Η μορφή ενός ΜαΣΤ προσομοιάζει με αυτή ενός ελεγκτή, όπως φαίνεται στο Σχήμα 4.1.

Στα ΜαΣΤ, η μάθηση μοντελοποιείται ως μία διαδικασία online προσαρμογής σε ένα άγνωστο περιβάλλον, το οποίο αντιπροσωπεύει το πρόβλημα, και το οποίο τροφοδοτεί το σύστημα με, συνήθως αριθμητικές, ανταμοιβές. Ένα ΜαΣΤ αντιλαμβάνεται το περιβάλλον του μέσω των ανιχνευτών (detectors) του, και, με βάση τις αισθήσεις του, επιλέγει μία ενέργεια που εφαρμόζεται στο περιβάλλον του μέσω των επενεργητών (effectors) του. Πρακτικά, δέχεται ένα διάνυσμα εισόδου (vision vector) από το περιβάλλον και εξάγει μία απόφαση - δράση προς αξιολόγηση από το περιβάλλον. Ανάλογα με την αποτελεσματικότητα των ενεργειών του, το περιβάλλον μπορεί να αποδώσει κάποιου είδους ανταμοιβή στο σύστημα, και, συνεπώς, στο πλαίσιο αυτό, ένα ΜαΣΤ προσπαθεί να μάθει μεγιστοποιώντας το ποσό της λαμβανόμενης ανταμοιβής.

Η εσωτερική δομή ενός ΜαΣΤ αποτελείται από ένα σύνολο τμημάτων και συνιστωσών, μέρος του οποίου φαίνεται στο Σχήμα 4.2. Πιο συγκεκριμένα, ένα ΜαΣΤ αποτελείται από:



Σχήμα 4.1: Εξωτερική Μορφή Μανθάνοντος Συστήματος Ταξινομητών.

1. Ένα *σύνολο κανόνων ταξινόμησης* $[P]$ που ονομάζεται πληθυσμός. Καθώς ένας κανόνας μπορεί να εμφανίζεται περισσότερο από μία φορά μέσα στον πληθυσμό, μακροσκοπικά, αντιλαμβανόμαστε τη συγχώνευσή τους σε μακροκανόνες (macroclassifiers), δηλαδή κανόνες με μία συγκεκριμένη πληθικότητα, ίση με τον αριθμό των επαναλήψεων του δεδομένου κανόνα στον πληθυσμό.

   Όπως αναφέραμε και στην Παρ. 2.4.1, ένας κανόνας ταξινόμησης αποτελείται από δύο τμήματα: το τμήμα συνθήκης και το τμήμα απόφασης. Η *συνθήκη* προσδιορίζει ένα τμήμα του χώρου γνωρισμάτων του προβλήματος, ενώ η *απόφαση* είναι η ενέργεια που σχετίζεται με το υποπρόβλημα που προσδιορίζεται από τη συνθήκη του ίδιου κανόνα. Στα πλαίσια των ΜαΣΤ, εισάγονται διάφορα μεγέθη με τα οποία σχετίζονται οι κανόνες, ανάμεσά τους, η *πρόβλεψη* και η *καταλληλότητα*. Η πρόβλεψη (ή δύναμη) ενός κανόνα είναι υπεύθυνη για την εκτίμηση της αξίας της ενέργειας του κανόνα, από την άποψη των μελλοντικών ανταμοιβών του ΜαΣΤ. Η καταλληλότητα προσφέρει μία εκτίμηση της ποιότητάς του.

2. Ένα *σύνολο ενεργοποίησης* $[M]$ (*match set*) που δημιουργείται από τους κανόνες του $[P]$ που ενεργοποιούνται από (καλύπτουν) το διάνυσμα εισόδου.

3. Ένα σύνολο $[A]$ (*σύνολο ενεργειών - action set*) που αποτελείται από τους κανόνες του $[M]$, των οποίων η απόφαση συνάδει με αυτή που επιλέγει το ΜαΣΤ.

4. Το *Γενετικό Αλγόριθμο*, ως το ένα από τα δύο μέρη της Συνιστώσας Ανακάλυψης, που κατά τη φάση της εξερεύνησης είναι υπεύθυνος για την επιλογή κανόνων με βάση την καταλληλότητά τους, την αντιγραφή τους, την εφαρμογή

των γενετικών τελεστών της διασταύρωσης και μετάλλαξης στα προαναφερθέντα αντίγραφα και την υπό συνθήκες αφομοίωση των απογόνων από ήδη υπάρχοντες κανόνες του πληθυσμού. Για την εξέλιξη κανόνων στα ΜαΣΤ έχει επικρατήσει η χρήση Γενετικών Αλγορίθμων Σταθερής Κατάστασης (Εν. 3.3), σε συνδυασμό, όπως προείπαμε, με την προσέγγιση Michigan για την αναπαράσταση των χρωμοσωμάτων. Ο συνδυασμός των δύο είναι φυσικός, καθώς επιχειρείται η εύρεση ενός συνόλου συνεργαζόμενων κανόνων.

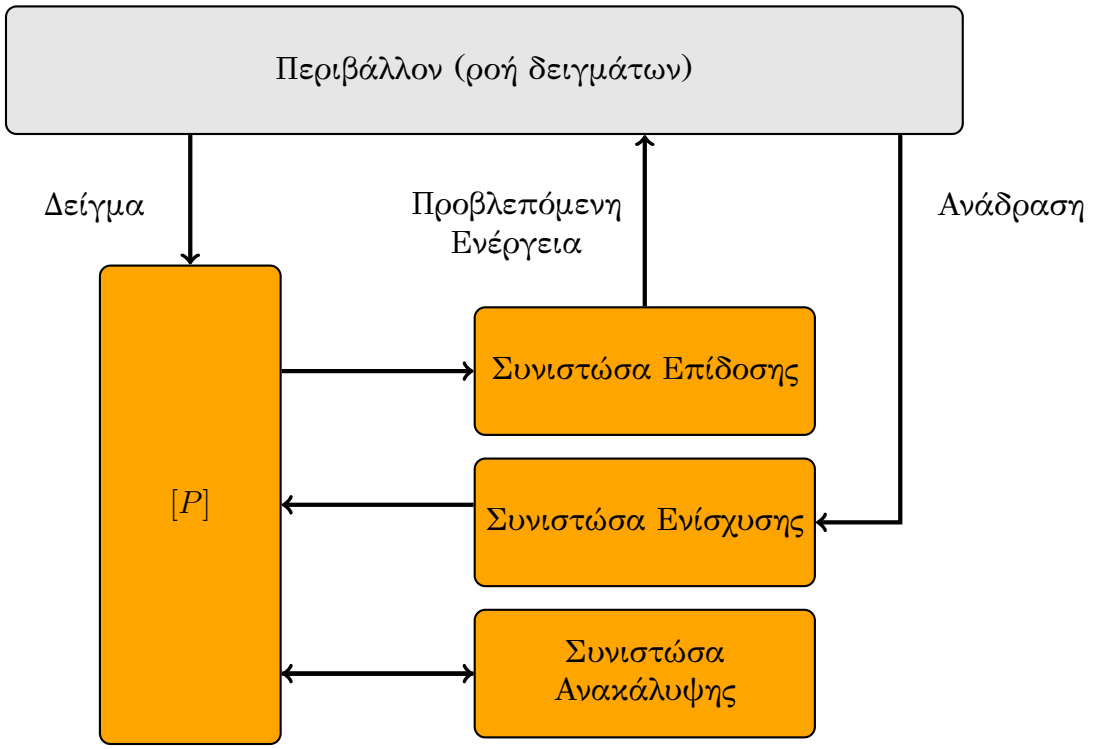
5. Το *τμήμα Κάλυψης* ως το δεύτερο μέρος της Συνιστώσας Ανακάλυψης. Το τμήμα Κάλυψης παράγει κανόνες όταν δεν υπάρχουν κανόνες στον πληθυσμό που να ενεργοποιούνται για ένα δεδομένο διάνυσμα εισόδου, συνήθως στις πρώτες επαναλήψεις όπου ο πληθυσμός είναι κενός κανόνων.

6. Τη *λειτουργία ομαδοποίησης - αφομοίωσης* (subsumption). Αναλαμβάνει την αφομοίωση απογόνων, μετά τη δημιουργία τους, από κανόνες αρκούντως κατάλληλους και έμπειρους, με τμήμα συνθήκης γενικότερο και τμήμα απόφασης ειδικότερο από τους απογόνους.

7. Τη *λειτουργία διαγραφής κανόνων* που αναλαμβάνει να διατηρήσει κάτω από ένα όριο τον αριθμό των κανόνων του πληθυσμού. Ενεργοποιείται όταν, λόγω συνεχούς παραγωγής κανόνων από το Γενετικό Αλγόριθμο, ο αριθμός των κανόνων του πληθυσμού ξεπεράσει το προαναφερθέν όριο, διαγράφοντας κανόνες με πιθανότητα αντίστροφη της καταλληλότητάς τους.

8. Τη *συνιστώσα ενίσχυσης ή απόδοσης ανταμοιβής*. Κατανέμει την εισερχόμενη από το περιβάλλον ανταμοιβή στους κανόνες που είναι υπεύθυνοι για αυτήν και ενημερώνει τις σχετικές με την ποιότητα παραμέτρους τους.

9. Το *τμήμα Επίδοσης*. Μετά το σχηματισμό του $[M]$, το σύστημα καλείται να λάβει μία απόφαση από αυτές που υποστηρίζουν οι κανόνες συνόλου αυτού. Εν γένει, οι ενέργειες που υποστηρίζονται είναι διαφορετικές και, συνεπώς, αντικρουόμενες. Όπως αναφέρθηκε παραπάνω, για κάθε νέα είσοδο, το ΜαΣΤ επιλέγει μία δράση βάσει των αποφάσεων των κανόνων του συνόλου $[M]$. Η επιλογή της δράσης εξαρτάται από τον τρόπο λειτουργίας του συστήματος σε εκείνη τη στιγμή. Εάν βρίσκεται σε φάση *Αξιοποίησης* (exploitation), επιλέγεται η δράση εκείνη που είναι βέλτιστη ως προς κάποια μετρική ή γίνεται ψηφοφορία μεταξύ των κανόνων του $[M]$. Στην περίπτωση που το σύστημα βρίσκεται σε φάση *Εξερεύνησης* (exploration), η επιλογή της δράσης μπορεί να βασίζεται σε κάποια πιθανοτική κατανομή, να είναι τυχαία ή να εναλλάσσεται ανάμεσα σε τυχαία και βέλτιστη. Το ΜαΣΤ, λοιπόν, αξιολογεί κάθε προτεινόμενη ενέργεια από το σύνολο των κανόνων του $[M]$ και με κάποιο προκαθορισμένο τρόπο επιλέγει μία προς εφαρμογή, αποστέλλοντας την στο περιβάλλον, από το οποίο στη συνέχεια περιμένει να λάβει τη βαθμωτή ανταμοιβή.



Σχήμα 4.2: Εσωτερική Μορφή ΜαΣΤ τύπου Michigan.

Είδη Προβλημάτων

Τα προβλήματα που καλούνται να λύσουν τα ΜαΣΤ χωρίζονται σε δύο κύριες κατηγορίες: τα προβλήματα *διαδοχικών αποφάσεων* και τα προβλήματα *ενός βήματος*. Για την περιγραφή των προβλημάτων πολλών βημάτων αξιοποιείται η έννοια της Μαρκοβιανής Διαδικασίας Απόφασης και η τεχνική της ενισχυτικής μάθησης με τον αλγόριθμο Q-learning [Wat89]. Αντίθετα, για τα προβλήματα ενός βήματος, όπως είναι η εξόρυξη δεδομένων, η χρήση μεθόδων ενισχυτικής μάθησης δεν είναι απαραίτητη καθώς έχουν αναπτυχθεί ΜαΣΤ που αξιοποιούν μεθόδους επιβλεπόμενης μάθησης [BBMH08].

ZCS: ΜαΣΤ Βασισμενο στη Δυναμη

Το πρώτο ολοκληρωμένο ΜαΣΤ ήταν ο ZCS (Zeroth level Classifier System), που δημιουργήθηκε από τον Wilson το 1994 [Wil94]. Κάθε κανόνας του πληθυσμού στον ZCS χαρακτηρίζεται (κυρίως) από μία παράμετρο, τη δύναμή (strength) του, η οποία εκτιμά την ανταμοιβή του κανόνα, και, ταυτόχρονα, ορίζει και την καταλληλότητά του. Ένας κανόνας αναπαρίσταται από την τριάδα (c, a, s) όπου c είναι η συνθήκη του κανόνα, a η προτεινόμενη ενέργεια και s η δύναμή του. Όπως είναι φανερό, για την προσέγγιση της τιμής της δύναμης κάθε κανόνα, χρησιμοποιούνται αλγόριθμοι ενισχυτικής μάθησης. Με την είσοδο του διανύσματος εισόδου, σχηματίζεται το $[M]$ και επιλέγεται κάποια δράση, με τον τρόπο που αναφέρθηκε παραπάνω (Εν. 4.1).

Μετά την εκτέλεση της δράσης, η συνιστώσα ενίσχυσης αναλαμβάνει να μοιράσει την ανταμοιβή που λήφθηκε από το σύστημα, η οποία μπορεί να είναι και μηδενική, στους κανόνες του $[M]$ των οποίων η απόφαση συμφωνεί με αυτήν που επέλεξε το σύστημα (στους κανόνες του $[A]$ δηλαδή). Ο Γενετικός Αλγόριθμος εκτελείται στο σύνολο των κανόνων $[P]$, επιλέγοντάς τους πιθανοτικά, με πιθανότητα επιλογής ανάλογη της καταλληλότητας (δύναμής) τους.

Η προσέγγιση του ZCS εξελίσσει κανόνες που είναι *συνεπώς σωστοί* στην πρόβλεψη της προσβλεπόμενης ανταμοιβής, χαράσσοντας ένα *Χάρτη Βέλτιστων Αποφάσεων (ΧΒΑ)*, καθώς η καταλληλότητα ενός ατόμου είναι ευθέως ανάλογη της εκτιμώμενης του ανταμοιβής. Εν γένει, στην ουσία, ένας ΧΒΑ περιέχει τους κανόνες εκείνους οι οποίοι προβλέπουν τη σωστή κλάση για κάθε δείγμα που καλύπτουν.

Για πολλές εφαρμογές, η εξέλιξη ενός ΧΒΑ είναι πλήρως αποδεκτή και, επιπρόσθετα, έχει το πλεονέκτημα της εξέλιξης συνόλων κανόνων πολύ μικρότερων από έναν *Πλήρη Χάρτη Αποφάσεων (ΠΧΑ)* και, συνεπώς, της εξαγωγής λύσεων πιο εύκολων στην ερμηνεία. Ένας ΠΧΑ περιλαμβάνει όλους τους *συνεπώς ακριβείς κανόνες*, ανεξάρτητα από το αν παρέχουν σωστές προβλέψεις ή όχι και, επομένως, ένας ΠΧΑ δεν περιλαμβάνει μόνο όλους τους συνεπώς σωστούς κανόνες, αλλά και όλους τους *συνεπώς λανθασμένους κανόνες*, οι οποίοι παρέχουν λανθασμένες προβλέψεις για κάθε δείγμα που καλύπτουν.

Παρ' όλα αυτά, όπως σημειώνει ο Wilson [Wil95], η επιλογή κανόνων βάσει της δύναμής τους μπορεί να οδηγήσει τον ZCS σε πρόωρη σύγκλιση σε υπό-βέλτιστες λύσεις, προτού καταφέρει να εξερευνήσει πλήρως το χώρο αναζήτησης. Επιπρόσθετα, η βασισμένη-στη-δύναμη επιλογή κανόνων ενδέχεται να δημιουργήσει προβλήματα στην εξερεύνηση, παρουσία αρχικών λύσεων υψηλής καταλληλότητας. Ένα ακόμα ζήτημα που απορρέει από αυτή την αρχιτεκτονική είναι η παραγωγή κανόνων μηδενικού οφέλους για το σύστημα. Αυτό οφείλεται στο γεγονός ότι ο Γενετικός Αλγόριθμος επιλέγει άτομα προς αναπαραγωγή από το σύνολο των διαθέσιμων κανόνων (panmictic selection), με αποτέλεσμα οι γονείς πιθανώς να διαφέρουν ριζικά στις αποφάσεις τους. Τα παραπάνω μειονεκτήματα, σε συνδυασμό με ορισμένες συνθήκες (π.χ δυναμικά περιβάλλοντα ή περιβάλλοντα με ισχυρή παρουσία θορύβου) και η ανάγκη για εξυπηρέτηση των ιδιαίτερων περιορισμών που επιβάλλουν ορισμένες εφαρμογές (η Εξόρυξη Δεδομένων είναι η κυριότερη), οδήγησαν τον Wilson στο συμπέρασμα ότι η εξέλιξη ενός ΠΧΑ ίσως είναι προτιμότερη, σχεδιάζοντας τον XCS.

XCS: ΜΑΣΤ Βασισμενο στην Ακριβεια Προβλεψης

Ο XCS[1] ήταν το πρώτο ΜΑΣΤ [Wil95] για το οποίο αναφέρθηκαν ακριβείς και βέλτιστες (maximal) γενικεύσεις. Η ικανότητά του αυτή οφείλεται, αφενός, στην αντικατάσταση της βάσης της καταλληλότητας, από τη δύναμη στην ακρίβεια πρόβλεψης, αφετέρου στη μεταβολή του συνόλου από όπου γίνεται επιλογή γονέων από το Γενετικό Αλγόριθμο, από τον πληθυσμό $[P]$, στο σύνολο ενεργειών $[A]$. Με άλλα λόγια, το σύστημα, πλέον, εξελίσσει εκείνους τους κανόνες που είναι οι πλέον ακριβείς στην πρόβλεψη της αναμενόμενης ανταμοιβής και που είναι υπεύθυνοι για

[1] Μια πληρέστερη αλγοριθμική περιγραφή του XCS περιέχεται στο [BW01].

τις προσφάτως εφαρμοσθείσες ενέργειες. Κάθε κανόνας αντιπροσωπεύεται πλήρως από την πεντάδα (c, a, p, ϵ, F) όπου c είναι η συνθήκη του κανόνα, a η ενέργεια που υποστηρίζει, p το ποσό της προβλεπόμενης ανταμοιβής, ϵ το σφάλμα πρόβλεψης, και F η καταλληλότητά του, η οποία πρέπει να ξανατονίσουμε ότι υπολογίζεται ως συνάρτηση της *ακρίβειας πρόβλεψης* και όχι του μεγέθους της αριθμητικής πρόβλεψης.

Η εξέλιξη ακριβών γενικεύσεων είναι ένα καίριο ζήτημα στα προβλήματα κατηγοριοποίησης, το οποίο θα μας απασχολήσει αργότερα, καθώς το εξαγόμενο μοντέλο θα πρέπει να είναι σε θέση να κατηγοριοποιήσει, όσο το δυνατόν ακριβέστερα, σύνολα δεδομένων με τα οποία δεν έχει εκπαιδευτεί, και, συνεπώς, του είναι πλήρως άγνωστα. Όσο πιο γενικό είναι το μοντέλο, τόσο περισσότερο εξασφαλίζεται ότι θα υπάρχουν κανόνες που θα ενεργοποιούνται για κάποιο τυχαίο δείγμα s. Από την άλλη, όσο περισσότερο ακριβείς είναι οι κανόνες, τόσο αποτελεσματικότερα θα κατηγοριοποιήσουν το δείγμα, καθιστώντας το μοντέλο αξιόπιστο.

Λόγω ακριβώς της μεταστροφής του υπολογισμού της καταλληλότητας από τη δύναμη στην ακρίβεια πρόβλεψης, ο XCS εξελίσσει ΠΧΑ, δηλαδή *συνεπώς ακριβείς* κανόνες, ανεξάρτητα της ορθότητας πρόβλεψης. Ένας ΠΧΑ, δηλαδή, περιέχει όχι μόνο τους *συνεπώς ορθούς* κανόνες, αλλά και τους *συνεπώς λανθασμένους*, οι οποίοι προβλέπουν άριστα μία μηδενική ανταμοιβή από το σύστημα. Οι ΠΧΑ απαιτούν εκτενέστερη εξερεύνηση του χώρου αναζήτησης καθώς συμπεριλαμβάνουν περιοχές οι οποίες δεν είναι σημαντικές για το πρόβλημα - στόχο, δηλαδή τις περιοχές όπου βρίσκονται οι συνεπώς λανθασμένοι κανόνες. Το γεγονός αυτό αποδεικνύεται επιβαρυντικό σε προβλήματα κατηγοριοποίησης μεγάλης διαστατικότητας ή/και πολλών κλάσεων. Επιπρόσθετα, το μέγεθος ενός ΠΧΑ μπορεί να είναι έως και n φορές μεγαλύτερο από το μέγεθος ενός ΧΒΑ, για δεδομένο πρόβλημα n κλάσεων, απαιτώντας πολύ μεγαλύτερο μέγεθος πληθυσμού και, συνεπώς, περισσότερους υπολογιστικούς πόρους από την εξέλιξη ενός ΧΒΑ [Kov00]. Επιπρόσθετα, επειδή το μέγεθος του χάρτη κάλυψης έχει αναγνωριστεί ως παράγοντας πολυπλοκότητας για τα ΜαΣΤ [KK01], η εξέλιξη ενός ΠΧΑ απαιτεί περισσότερες επαναλήψεις εκπαίδευσης και, επομένως, μεγαλύτερους χρόνους εκπαίδευσης. Παρ' όλα αυτά, ο Kovacs [Kov00] σημειώνει ότι μπορεί να υπάρχουν και πλεονεκτήματα στη διατήρηση των συνεπώς λανθασμένων κανόνων σε ένα χάρτη, καθώς, η ύπαρξη τους μπορεί να αποτρέψει το σύστημα από το να τους ανακαλύψει εκ νέου, βελτιώνοντας έτσι τη διαδικασία εξερεύνησης, ενώ, στη φάση της Αξιοποίησης, αυτοί οι κανόνες μπορούν να αποτελέσουν μία λίστα επιβλαβών επιλογών τις οποίες το σύστημα θα αποφύγει να κάνει.

Συνολικά, δεδομένης της φύσης των προβλημάτων κατηγοριοποίησης ενός βήματος, είναι φανερό ότι, για την παραγωγή μίας αποτελεσματικής γνωστικής αναπαράστασης ενός προβλήματος, μόνο οι συνεπώς ορθοί κανόνες είναι απαραίτητοι. Σε αυτή την κατεύθυνση κινείται η επέκταση του XCS, ο UCS.

UCS: ΜαΣΤ για προβληματα Επιβλεπομενης Μαθησης

Ο UCS είναι το πρώτο ΜαΣΤ που χρησιμοποιεί αποκλειστικά επιβλεπόμενη μάθηση. Κληρονομεί τις κύριες συνιστώσες του XCS και τις προσαρμόζει στα πλαίσια της επιβλεπόμενης μάθησης. Οι διαφορές με τον XCS είναι ότι i) χρησιμοποιεί ελαφρώς διαφορετικές παραμέτρους και εναλλακτικούς τρόπους για την ενημέρωσή τους, και, ii) δεν χρησιμοποιεί μάθηση χρονικών διαφορών για την ενημέρωση της ακρίβειας των κανόνων, αλλά την εκτιμά απευθείας από το ποσοστό ορθών κατηγοριοποιήσεων τους. Επειδή ακριβώς οι κλάσεις στις οποίες ανήκει το κάθε δείγμα $s \in D$ του συνόλου δεδομένων εκπαίδευσης για ένα πρόβλημα κατηγοριοποίησης είναι εκ των προτέρων γνωστές, ο UCS, αντί για το σύνολο $[A]$, δημιουργεί το σύνολο $[C]$ (Correct Set), στο οποίο περιλαμβάνονται όλοι οι κανόνες των οποίων η απόφαση συμφωνεί με την κλάση του δείγματος s, δηλαδή οι κανόνες που ταξινομούν ορθά το s. Παράλληλα, σχηματίζει και το $[!C]$ (Incorrect Set), το οποίο προκύπτει από την αφαίρεση του $[C]$ από το $[M]$ ($[!C] = [M] - [C]$), το σύνολο δηλαδή των κανόνων που ταξινομούν λανθασμένα το s.

Κάθε κανόνας αντιπροσωπεύεται από το σύνολο παραμέτρων $(c, a, tp, exp, num, cs, F)$, όπου

- c και a η συνθήκη και η απόφαση του κανόνα αντίστοιχα
- tp ο αριθμός των σωστών κατηγοριοποιήσεων του κανόνα
- exp ο συνολικός αριθμός των δειγμάτων που κλήθηκε να κατηγοριοποιήσει, δηλαδή ο αριθμός των $[M]$ στα οποία συμμετείχε ο κανόνας
- num η πληθικότητα του κανόνα, δηλαδή ο αριθμός των αντιγράφων του, στο σύνολο των κανόνων
- cs μία εκτίμηση του μέσου μεγέθους των $[C]$ στα οποία συμμετείχε ο κανόνας, και, τέλος,
- F η καταλληλότητά του.

Η *ακρίβεια* (accuracy) ενός κανόνα r ορίζεται ως το ποσοστό ορθών κατηγοριοποιήσεων του:

$$Accuracy(r) = \frac{tp(r)}{exp(r)} \tag{4.1}$$

Η *καταλληλότητα* (fitness) κάθε μεμονωμένου κανόνα (micro-classifier) δίνεται από τη σχέση:

$$F_{micro}(r) = Accuracy(r)^{\nu} \tag{4.2}$$

όπου ν μία σταθερά επιλεγμένη από το χρήστη που καθορίζει το βαθμό πίεσης προς ορθούς κανόνες, με συνηθισμένη τιμή το 10. Η συνολική καταλληλότητα ενός κανόνα (macro-classifier) προκύπτει από το άθροισμα των καταλληλοτήτων των αντιγράφων του:

$$F_{macro}(r) = num(r) \cdot F_{micro}(r) \tag{4.3}$$

Μετά την παρουσίαση ενός δείγματος $s \in D$ στο σύστημα, σχηματίζεται το $[M]$ και αυξάνεται κατά μία μονάδα η εμπειρία των κανόνων που συμμετέχουν σε αυτό. Στη συνέχεια, σχηματίζεται το $[C]$ από τους κανόνες των οποίων η συνθήκη συμφωνεί με την κλάση του s. Για αυτούς τους κανόνες, αυξάνεται κατά ένα ο αριθμός των ορθών κατηγοριοποιήσεων tp υπολογίζεται το μέγεθος cs και ενημερώνεται η καταλληλότητά τους.

Στον UCS, το σύνολο από όπου αντλεί ο Γενετικός Αλγόριθμος τους υποψήφιους γονείς είναι το $[C]$, δηλαδή το σύνολο των κανόνων που ταξινομούν ορθά ένα δεδομένο δείγμα εισόδου. Σε συνδυασμό μάλιστα με τη λειτουργία της διαγραφής που φροντίζει για την απομάκρυνση από τον πληθυσμό κανόνων, επιλέγοντάς τους βάσει της αντίστροφης καταλληλότητάς τους, καταλαβαίνουμε πως ο UCS εξελίσσει XBA, χαρτογραφώντας μόνο εκείνες τους κανόνες που προβλέπουν ορθά την κλάση των δειγμάτων που καλύπτουν. Η μεθοδολογία διαγραφής περιλαμβάνει την ανάθεση σε κανόνες του πληθυσμού μίας πιθανότητας διαγραφής ανάλογη προς την ποσότητα

$$d(i) = \begin{cases} \dfrac{cs(i) \cdot F_P}{F_{micro}(i)}, & experience(i) > \theta_{del} \text{ και } F_{micro}(i) < \delta \cdot F_P \\ cs(i), & \text{αλλού} \end{cases} \tag{4.4}$$

όπου F_P είναι το άθροισμα των καταλληλοτήτων των κανόνων του πληθυσμού $[P]$.

Ο αναγνώστης μπορεί να ανατρέξει στα [BMGG03, OPBM08] για περισσότερες λεπτομέρειες σχετικά με τον UCS.

*S-LCS: Γενικευμενα ΜαΣΤ για Εξορυξη Δεδομενων

Το *S-LCS αποτελεί ένα γενικότερο πλαίσιο επιβλεπόμενης μάθησης με ΜαΣΤ [Tzi12] για προβλήματα ενός βήματος και είναι ανεξάρτητο από τον υπολογισμό του τρόπου υπολογισμού της καταλληλότητας. Κληρονομεί το σύνολο των παραμέτρων. λειτουργιών, συνόλων και συνιστωσών του UCS και το επεκτείνει, χρησιμοποιώντας, εκτός από τον αριθμό των ορθών αποφάσεων ενός κανόνα tp και τον αριθμό των λανθασμένων αποφάσεών του fp, μία βαθμωτή ποσότητα str που εκτιμά τη μέση ανταμοιβή που λαμβάνει ανά βήμα και τη *χρονοσφραγίδα* ts (timestamp) που αποθηκεύει το χρονικό βήμα της τελευταίας συμμετοχής του σε σύνολο $[C]$.

Το τμήμα ενημέρωσης κληρονομείται και αυτό από τον UCS, αναλαμβάνοντας να ενημερώσει τις παραμέτρους των κανόνων στα $[M]$, όπως ακριβώς και ο UCS, με την προσθήκη της αύξησης του fp κατά ένα για τους κανόνες που ανήκουν στο $[!C]$. Το τμήμα εξερεύνησης περιλαμβάνει, εκτός από τη λειτουργία της κάλυψης, έναν Γενετικό Αλγόριθμο Σταθερής Κατάστασης, που εφαρμόζεται στα $[C]$ με μέσο ρυθμό θ_{GA}. Όταν ο μέσος όρος των χρονοσφραγίδων των κανόνων ενός $[C]$ ξεπερνά

την τρέχουσα τιμή χρονικού βήματος κατά θ_{GA}, τότε εφαρμόζεται ο Γενετικός Αλγόριθμος, επιλέγοντας γονείς με επιλογή τουρνουά. Η διαγραφή των κανόνων γίνεται, όπως και στον UCS, από το σύνολο των κανόνων του πληθυσμού, όταν αυτό ξεπερνάει κάποιο προκαθορισμένο μέγεθος, με ίδια μεθοδολογία ως προς τον UCS.

Ο *S-LCS διαθέτει δύο εκδόσεις: μία ακριβειο-κεντρική, τον AS-LCS, και μία δυναμο-κεντρική, τον SS-LCS. Ο AS-LCS, εκτός από τις παραπάνω αλλαγές, κληρονομεί αυτούσιες τις λειτουργίες του UCS. Ο SS-LCS προσπαθεί να προσεγγίσει το πρόβλημα της κατηγοριοποίησης με μία πιο παραδοσιακή μέθοδο, κάνοντας χρήση της δύναμης str η οποία χαρακτηρίζει τον κάθε κανόνα. Τα αποτελέσματα του SS-LCS πάνω σε πραγματικά σύνολα δεδομένων δείχνουν πως είναι τουλάχιστον ισοδύναμος, αν όχι καλύτερος, από αρκετές state-of-the-art προσεγγίσεις, παρέχοντας μοντέλα που ισορροπούν τις δύο απαιτήσεις σχεδίασής του: την αποδοτικότητα και την ερμηνευσιμότητα των παραγόμενων μοντέλων.

ΜΕΡΟΣ ΙΙ

Πολυκατηγορική Ταξινόμηση με Μανθάνοντα Συστήματα Ταξινομητών

5

Πολυκατηγορική Ταξινόμηση με τον GMl-ASLCS$_0$

Η παρούσα εργασία ερείδεται επί της Διπλωματικής Εργασίας του Μίλτου Αλλαμανή [Mil11] και του Μανθάνοντος Συστήματος Ταξινομητών που ανέπτυξε. Το ΜαΣΤ αυτό θα το ονομάσουμε, στα πλαίσια αυτής της εργασίας, GMl-ASLCS$_0$. Η ονομασία, εκ πρώτης όψης, προσομοιάζει στο ΜαΣΤ που αναφέραμε στο προηγούμενο κεφάλαιο, τον AS-LCS. Στην πραγματικότητα, είναι η επέκτασή του, η προσαρμογή του, στον πολυκατηγορικό (Ml) χώρο, όπως ακριβώς και οι αλγόριθμοι που παρουσιάσαμε στην Παρ. (2.5.2). Ο GMl-ASLCS$_0$ κληρονομεί τις βασικές λειτουργίες του πλαισίου λειτουργίας του AS-LCS, όπως τις παραμέτρους των κανόνων που εξελίσσει, και επεκτείνει άλλες, στο πολυκατηγορικό πεδίο. Η μηδενική υποσημείωση στην ονομασία του GMl-ASLCS$_0$ γίνεται για να καταδείξουμε ότι αυτό το σύστημα είναι η αφετηρία από όπου ξεκινάμε τη μελέτη, την ανάλυσή και την επέκτασή μας. Ο GMl-ASLCS$_0$ είναι μία από τις πρώτες προσεγγίσεις του προβλήματος της πολυκατηγορικής (ή πολυετικετικής) ταξινόμησης με μεθόδους Εξελικτικής Υπολογιστικής. Η μελέτη του στα πλάισια των [Mil11] και [ATM13] παρέχει τις πρώτες ενδείξεις πως η πολυκατηγορική ταξινόμηση με ΜαΣΤ είναι εφικτή, ενώ υπάρχει χώρος για βελτίωση, ώστε ο GMl-ASLCS$_0$ να μπορέσει να ανταγωνιστεί αποτελεσματικότερα τους state-of-art αλγορίθμους στο χώρο της πολυκατηγορικής ταξινόμησης.

Σε αυτό το κεφάλαιο:

- παρουσιάζουμε τις αλλαγές που ήταν απαραίτητο να γίνουν στον AS-LCS ώστε να είναι σε θέση να ταξινομεί πολυκατηγορικά δείγματα, παρέχοντας έτσι το υπόβαθρο για την κατανόηση του GML-ASLCS$_0$ και του μοντέλου που αναπτύξαμε,

- επισκεπτόμαστε, παράλληλα, τις βασικές συνιστώσες του GML-ASLCS$_0$, και

- μπαίνουμε σε μεγαλύτερο βάθος σε ορισμένες λειτουργίες του, ώστε να γίνουν κατανοητοί οι λόγοι για τους οποίους τις τροποποιούμε.

Τροποποιηση της Αναπαραστασης Κανονων

Η Αναπαράσταση των Κανόνων που χρησιμοποιούνται από ένα ΜαΣΤ μονοκατηγορικής ταξινόμησης τροποποιείται, ώστε στο τμήμα συνθήκης των κανόνων να περιλαμβάνεται η απόφαση για πολλαπλές ετικέτες. Προτού αναφερθούμε στην αναπαράσταση των ετικετών, εξετάζουμε την αναπαράσταση του τμήματος συνθήκης για τους διαφορετικούς τύπους γνωρισμάτων: τα *δυαδικά*, *ονομαστικά* και *πραγματικά* γνωρίσματα.

Αναπαράσταση Δυαδικών Γνωρισμάτων

Οι διαθέσιμες τιμές για ένα δυαδικό γνώρισμα είναι τρεις: *Αληθής*, *Ψευδής*, και *Αδιαφορία* (#). Ο χαρακτήρας # συμβολίζει την αδιαφορία ενός κανόνα για την τιμή του αντίστοιχου γνωρίσματος στα δεδομένα εισόδου. Συνεπώς, ένας κανόνας της μορφής 111# $\Rightarrow$ 101 καλύπτει *τουλάχιστον δύο* δείγματα: το 1111 $\Rightarrow$ 101, και το 1110 $\Rightarrow$ 001. Λόγω της ύπαρξης τριών πιθανών καταστάσεων για την τιμή ενός δυαδικού γνωρίσματος, χρησιμοποιούνται δύο δυαδικά ψηφία για την αναπαράστασή της. Ένα δυφίο, το λεγόμενο *Ψηφίο Ενεργοποίησης*, καθιστά το *Ψηφίο Τιμής* σχετικό ή άσχετο: εάν το Ψηφίο Ενεργοποίησης έχει τιμή μηδέν, ο κανόνας που το περιέχει αδιαφορεί για το συγκεκριμένο γνώρισμα. Σε αντίθετη περίπτωση, η τιμή του γνωρίσματος ταυτίζεται με την τιμή του Ψηφίου Τιμής.

$$\overbrace{\underbrace{b_1}_{\text{ψηφίο ενεργοποίησης}}\quad\underbrace{b_0}_{\text{ψηφίο τιμής}}}^{\text{Συνθήκη Δυαδικού Γνωρίσματος}}$$

Αναπαράσταση Ονομαστικών Γνωρισμάτων

Η αναπαράσταση ονομαστικών γνωρισμάτων αποτελεί γενίκευση αυτής των δυαδικών. Περιέχει και αυτή ένα ψηφίο ενεργοποίησης και, για την αναπαράσταση των n τιμών ενός ονομαστικού γνωρίσματος, χρησιμοποιεί μία δυαδική μάσκα n δυφίων. Μία μη μηδενική τιμή ενός από αυτά τα δυφία υποδηλώνει την αντίστοιχη τιμή του ονομαστικού γνωρίσματος. Όταν η μάσκα ενεργοποίησης τιμών εφαρμοσθεί σε ένα δείγμα και προκύψει τιμή διάφορη του μηδενός, τότε το δείγμα ικανοποιεί τη συγκεκριμένη συνθήκη του κανόνα.

$$\overbrace{\underbrace{b_n}_{\text{Ψηφίο Ενεργοποίησης}}\quad\underbrace{b_{n-1}b_{n-2}\ldots b_1b_0}_{\text{μάσκα τιμών}}}^{\text{Συνθήκη Ονομαστικού Γνωρίσματος}}$$

Αναπαράσταση Συνθηκών Πραγματικών Γνωρισμάτων

Για την αναπαράσταση συνθηκών πραγματικών γνωρισμάτων επιλέχθηκε η αναπαράσταση *διαστήματος τιμών*, η οποία ορίζει ένα διάστημα επιτρεπτών τιμών για το γνώρισμα στο οποίο αναφέρεται, καθιστώντας εφικτή και τη δημιουργία μονόπλευρων ανισοτήτων. Επιπρόσθετα, επιλέχτηκε η αναπαράσταση θέσης για την κωδικοποίηση των εμπλεκομένων πραγματικών αριθμών. Στην αναπαράσταση θέσης, επειδή η μέγιστη και η ελάχιστη τιμή ενός γνωρίσματος είναι εκ των προτέρων γνωστές, κβαντίζονται οι ενδιάμεσες τιμές και αντιστοιχίζονται στις προκύπτουσες στάθμες τα δυαδικά ψηφία που χρησιμοποιούνται για την αναπαράσταση ενός αριθμού. Με αυτόν τον τρόπο, χρησιμοποιούνται, μεν, όλα τα διαθέσιμα δυαδικά ψηφία χωρίς να υπάρχει η ανάγκη για διόρθωση, καθίσταται δε πολυπλοκότερος ο υπολογισμός του πραγματικού αριθμού στον οποίο αντιστοιχεί το γονίδιο. Πιο συγκεκριμένα:

$$x_i = min_i + \frac{int(gene_i)}{2^b}(max_i - min_i) \tag{5.1}$$

όπου x_i η πραγματική τιμή του γνωρίσματος i, max_i και min_i η καθολική μέγιστη και ελάχιστη τιμή του γνωρίσματος αντίστοιχα, b ο αριθμός των δυαδικών ψηφίων που χρησιμοποιήθηκαν στην αναπαράσταση και $gene_i$ η δυαδική αναπαράσταση του γονιδίου. Να σημειωθεί ότι το b είναι μια παράμετρος με την οποία μπορεί να ελεγχθεί η διακριτική ικανότητα του κανόνα. Μεγαλύτερες τιμές του b, οδηγούν σε μεγαλύτερη ακρίβεια, με κόστος την αύξηση του μεγέθους του γονιδίου. Συνολικά, με βάση την αναπαράσταση των πραγματικών τιμών που επιλέχθηκε, για 2^k στάθμες κβαντισμού, η δομή των συνθηκών πραγματικών γνωρισμάτων μπορεί να αποδοθεί σχηματικά ως εξής:

$$\overbrace{\underbrace{b_{2k}}_{\text{ψηφίο ενεργοποίησης}}\ \underbrace{b_{2k-1}b_{2k-2}\ldots b_{k+1}b_k}_{\text{άνω όριο}}\ \underbrace{b_{k-1}b_{k-2}\ldots b_1b_0}_{\text{κάτω όριο}}}^{\text{Συνθήκη Πραγμ. Γνωρισμάτων με Αναπαράσταση Διαστήματος Τιμών}}$$

Αναπαράσταση του Τμήματος Απόφασης

Η αναπαράσταση των κατηγοριών - ετικετών είναι σαφέστατα κρίσιμης σημασίας. Στην κατηγοριοποίηση μίας κατηγορίας, η αναπαράσταση είναι κάτι το τετριμμένο: οι πιθανές τιμές της μοναδικής ετικέτας - απόφασης αναπαριστώνται ως φυσικοί αριθμοί και ενσωματώνονται στην αναπαράσταση του γονιδίου ως δυαδικοί. Στην πολυκατηγορική ταξινόμηση υπάρχει μεγάλος χώρος για εναλλακτικές προσεγγίσεις - από την *αναπαράσταση μόνο μίας ετικέτας*, όπως στην απλή κατηγοριοποίηση, μέχρι τη *σαφή αναπαράσταση όλων των ετικετών* και την *αναπαράσταση ετικετών με αδιαφορίες*. Στον GMl-ASLCS$_0$ επιλέχθηκε η τελευταία, καθώς

η πρώτη προσεγγίζει τη μέθοδο μετασχηματισμού προβλημάτων BR, ή αν αναπαρασταθούν όλες οι ετικέτες στο ίδιο ΜαΣΤ, τη μέθοδο RT. Η δεύτερη προσέγγιση, από την άλλη, έχει ως αποτέλεσμα την παραγωγή εξαιρετικά συγκεκριμένων κανόνων, που αποφασίζουν πάντοτε υπέρ ή κατά της ταξινόμησης σε μία κατηγορία, αυξάνοντας το χώρο αναζήτησης των πιθανών καταστάσεων απόφασης και δυσκολεύοντας, έτσι, το Γενετικό Αλγόριθμο να συγκλίνει, ειδικά σε περιπτώσεις χαλαρής εξάρτησης μεταξύ των ετικετών.

Η αναπαράσταση ετικετών με αδιαφορίες προσομοιάζει στην αναπαράσταση συνθηκών δυαδικών γνωρισμάτων. Κάθε κανόνας διαθέτει τη δυνατότητα να αποφασίσει υπέρ ή κατά της ταξινόμησης σε μία δεδομένη ετικέτα, ή ακόμα να απέχει από τη διαδικασία απόφασης, αδιαφορώντας. Για $|L|$ ετικέτες, το τμήμα απόφασης των κανόνων παίρνει τη μορφή:

$a_0 l_0$	$a_1 l_1$	$\cdots$	$a_{\|L\|-1} l_{\|L\|-1}$

όπου a_i το ψηφίο ενεργοποίησης και l_i το δυφίο απόφασης για την ετικέτα i.

Σε αντίθεση με τη σαφή αναπαράσταση, οι βέλτιστοι κανόνες ορίζονται ως αυτοί με τη μέγιστη δυνατή κάλυψη και ταυτόχρονα το ειδικότερο δυνατό τμήμα απόφασης. Το σύστημα, λοιπόν, θα πρέπει να ισορροπήσει την εξερεύνηση του μεταξύ κανόνων γενικών, με πολλές αδιαφορίες στο τμήμα απόφασής τους, και κανόνων ειδικών με ελάχιστες αδιαφορίες. Ένας αποτελεσματικός τρόπος ώθησης του συστήματος προς αυτή την κατεύθυνση είναι η απαγόρευση συμμετοχής στα (σχηματιζόμενα ανά ετικέτα) Correct Sets των κανόνων που αδιαφορούν για την ετικέτα για την οποία σχηματίζεται το κάθε Correct Set (Η σχετική τροποποίηση της διαδικασίας ενημέρωσης και ο μηχανισμός ώθησης αναλύονται στην Εν. 5.2).

Συνολικά, είναι εμφανές ότι η χρήση της αναπαράστασης με αδιαφορίες μπορεί, ανάλογα με το πρόβλημα, να οδηγήσει σε μοντέλα που προσεγγίζουν όλο το φάσμα που ορίζεται από τις ακραίες περιπτώσεις των μετασχηματισμών BR (που χρησιμοποιεί μία ετικέτα) και LC (που χρησιμοποιεί όλους τους πιθανούς συνδυασμούς ετικετών). Ωστόσο, στα μειονεκτήματα αυτού του τρόπου αναπαράστασης θα πρέπει να προσμετρηθεί η πολυπλοκότητα που εισάγει στη διαδικασία εξερεύνησης (Εν. 5.3) και τις στρατηγικές συμπερασμού (Εν. 5.4). Τέλος, αξίζει να σημειωθεί ότι το γράμμα G στην ονομασία GMlAS-LCS$_0$ αναφέρεται σε ακριβώς αυτή την αναπαράσταση των ετικετών: την αναπαράσταση με χρήση αδιαφοριών.

Συνιστωσα Ενισχυσης

Η συνιστώσα ενίσχυσης και ενημέρωσης των παραμέτρων παίζει ρόλο-κλειδί στη μαθησιακή διαδικασία των ΜαΣΤ, καθώς είναι αυτή που αξιολογεί την ποιότητα των κανόνων, βάσει της οποίας εκτελούνται οι διαδικασίες της αναπαραγωγής, της διαγραφής και του συμπερασμού. Το τμήμα ενημέρωσης των πολυκατηγορικών ΜαΣΤ διαφοροποιείται από αυτό των μονοκατηγορικών εξαιτίας της πρόσθετης πολυπλοκότητας του προβλήματος προς επίλυση. Σημαντικές διαφορές που μπορούμε να εντοπίσουμε, μεταξύ άλλων, αφορούν:

- Στην *αξιολόγηση της ποιότητας των κανόνων*. Σε προβλήματα απλής κατηγοριοποίησης, ένας κανόνας ταξινομεί ένα δείγμα είτε απολύτως ορθά, είτε απολύτως λανθασμένα. Αντίθετα, στην πολυκατηγορική ταξινόμηση, τα πράγματα δεν είναι τόσο ξεκάθαρα, καθώς ένας κανόνας μπορεί να ταξινομεί ένα δείγμα ορθά σε μία ετικέτα, αλλά λανθασμένα σε μία άλλη. Συνεπώς, η μέτρηση της ποιότητας των κανόνων βάσει της απολύτως ορθής κατηγοριοποίησης αντικαθίσταται από αυτήν της *μερικώς ορθής* κατηγοριοποίησης, ώστε το σύστημα να διατηρήσει τη δυνατότητα σχετικής κατάταξης των κανόνων βάσει της ορθότητάς τους, χωρίς να γίνεται υπερβολικά αυστηρό.
- Στην *κάλυψη* των κανόνων. Η κάλυψη ενός κανόνα αναφέρεται στο ποσοστό των δειγμάτων που αυτός καλύπτει, δηλαδή το ποσοστό των δειγμάτων για τα οποία οι τιμές των γνωρισμάτων τους είναι ίσες με αυτές του τμήματος συνθήκης του κανόνα, ή δυνητικά ίσες (λόγω της αναπαράστασης κανόνων με αδιαφορίες στο τμήμα συνθήκης). Στον πολυκατηγορικό χώρο, λοιπόν, η κάλυψη, εκτός από το ποσοστό δειγμάτων που καλύπτει ο κανόνας, αναφέρεται πλέον και στο ποσοστό ετικετών που καλύπτει. Στη διαδικασία εξερεύνησης πρέπει να διασφαλίζεται η εξέλιξη συνολων κανόνων με πλήρη κάλυψη στα τμήματα συνθήκης και απόφασης.

Ο κύκλος εκπαίδευσης του GMl-ASLCS$_0$ παρουσιάζεται στον Αλγ. 5.1. Ο GMl-ASLCS$_0$ κληρονομεί τη *λογική* ενημέρωσης του AS-LCS, τροποποιώντας την αντίστοιχη διαδικασία (Αλγ. 5.4), ώστε να καλύπτει άμεσα τις ανάγκες της πολυκατηγορικής ταξινόμησης. Οι μέθοδοι εξαγωγής του Match Set για δεδομένο δείγμα $Instance$, του Correct Set για δεδομένο δείγμα $Instance$ και ετικέτα l και της αφαίρεσης κανόνων που δεν καλύπτουν κάποιο δείγμα του συνόλου δεδομένων $|D|$ περιγράφονται στους αλγορίθμους 5.2, 5.3 και 5.5 αντίστοιχα.

Αλγόριθμος 5.1 Ο κύκλος εκπαίδευσης του GMl-ASLCS$_0$

```
1: train(P, Instance)
2: M ← generateMatchset(P, Instance)
3: GMlASLCS0Update(P, M, Instance, D)
```

Αλγόριθμος 5.2 Παραγωγή του Match Set στον GMl-ASLCS$_0$

```
1: generateMatchSet(P, Instance)
2: initialize M
3: for each rule ∈ P do
4:    if rule covers Instance then
5:       M.add(rule)
6:    end if
7: end for
8: return M
```

Με την είσοδο του δείγματος εισόδου *Instance* στο ΜαΣΤ, οι κανόνες του πληθυσμού [P] που το καλύπτουν, σχηματίζουν το Match Set $[M]$ (Αλγ. 5.2). Στη συνέχεια, σχηματίζονται τόσα Correct Sets όσες είναι και οι ετικέτες των δειγμάτων

Αλγόριθμος 5.3 Παραγωγή του Correct Set στον GMl-ASLCS$_0$

generateLabelCorrectSet(M, l, $Instance$**)**
$initialize$ **C**
for each $rule \in M$ **do**
  if $rule.decision(l) = Instance.label(l)$ **then**
    C.$add(rule)$
  end if
end for
$return$ **C**

Αλγόριθμος 5.4 Συνιστώσα Ενημέρωσης του GMl-ASLCS$_0$

GMlASLCS$_0$Update(P, **M**, $Instance$, D**)**
for each $l \in L$ **do**
  C$[\, l \,] \leftarrow$ **generateLabelCorrectSet**(**M**$, l, Instance$)
end for
for each $rule \in$ **M do**
  updateFitness$(rule)$
  if $\exists l_i \in L : rule \in$ **C**$[\, l_i \,]$ **then**
    updateCs$(rule)$
  end if
end for
for each $l \in L$ **do**
  if C$[\, l \,] \neq \emptyset$ **then**
    if $timestamp - \overline{timestamp}([C_l]) > \theta_{GA}$ **then**
      $offspring \leftarrow evolve($**C**$[\, l \,])$
    end if
  else
    $offspring \leftarrow cover(Instance, l)$
  end if
  P.$insert(offspring)$
  if $|$**P**$| > maximumPopulationSize$ **then**
    $controlPopulation($**P**$)$
  end if
end for
if $random[0, 1] < 1/|D|$ **then**
  cleanUpZeroCoverageRules(**P**$, D$)
end if

Αλγόριθμος 5.5 Διαγραφή των κανόνων μηδενικής κάλυψης στον GMlASLCS$_0$

cleanUpZeroCoverageRules(P$, D$**)**
for each $rule \in$ **P do**
  if $rule.coveredInstances = 0$ AND $rule.presentedInstances = k \cdot |D|$ **then**
    P.$remove(rule)$
  end if
end for

του συνόλου δεδομένων (Αλγ. 5.3). Για μία τυχαία ετικέτα l, το $[C_l]$ περιλαμβάνει εκείνους τους κανόνες του $[M]$ των οποίων η απόφαση για την l ταυτίζεται με την πραγματική τιμή της l για το *Instance*. Οι κανόνες που αδιαφορούν για την l *δεν* συμμετέχουν στο $[C_l]$. Ο λόγος για αυτό αναλύεται στην Παρ. 6.4.1.

Για κάθε κανόνα που συμμετέχει στο $[M]$ (Αλγ. 5.6):

1. αυξάνεται κατά ένα η εμπειρία του,
2. ενημερώνονται οι μεταβλητές που είναι συναφείς με την καταλληλότητά του, μαζί με την ίδια την καταλληλότητα (συνάρτηση $updateFitness$ του Αλγ. 5.4) και
3. στην περίπτωση που έχει συμμετάσχει έστω και σε ένα $[C_l]$, ενημερώνεται η εκτίμηση του μεγέθους των $[C]$ στα οποία έχει συμμετάσχει μέχρι στιγμής (συνάρτηση $updateCs$ του Αλγ. 5.4).

Μετά το πέρας της διαδικασίας ενημέρωσης, ξεκινά η διαδικασία εξερεύνησης. Το τμήμα κάλυψης (συνάρτηση $evolve$ στη γρ. 17 του Αλγ. 5.4) ενεργοποιείται για κάθε $[C_l]$ το οποίο είναι κενό, ακολουθώντας τη μεθοδολογία που περιγράφεται στην Παρ. 5.3.2. Ο Γενετικός Αλγόριθμος (συνάρτηση $evolve$ στη γρ. 14 του Αλγ. 5.4) εκτελείται σε κάθε μη κενό $[C_l]$, επιλέγοντας δύο κανόνες προς αναπαραγωγή και δημιουργία δύο κανόνων-απογόνων, με τον τρόπο που περιγράφεται στην Παρ. 5.3.1.

Ενημέρωση Καταλληλότητας

Η ακριβής μεθοδολογία ενημέρωσης της καταλληλότητας και των μεταβλητών που σχετίζονται με αυτήν παρουσιάζεται στον Αλγ. 5.6.

Αλγόριθμος 5.6 Ενημέρωση της καταλληλότητας στον GMl-ASLCS$_0$

1: **updateFitness**($rule$)
2: $rule.exp \leftarrow rule.exp + 1$
3: **for each** $l \in L$ **do**
4: $\quad rule.tp \leftarrow rule.tp + correctness(rule, l)$
5: $\quad rule.msa \leftarrow rule.msa + 1$
6: **end for**
7: $rule.fitness \leftarrow \left(\frac{rule.tp}{rule.msa}\right)^{\nu}$

Για κάθε κανόνα στο Match Set εξετάζεται η ικανότητά του για κατηγοριοποίηση του $Instance$ σε κάθε ετικέτα l. Επίσης για κάθε ετικέτα, οι ποσότητες tp και msa αυξάνονται κατά ποσό ανάλογο της ικανότητας κατηγοριοποίησης ($correctness$) του κανόνα σε αυτή. Το μέγεθος tp αναπαριστά τον αριθμό ορθών κατηγοριοποιήσεων ενός κανόνα, ενώ το msa τον ολικό αριθμό κατηγοριοποιήσεων, σε αναλογία με το μέγεθος exp στη μονοκατηγορική ταξινόμηση.

Ο GMl-ASLCS$_0$ δεν επιτρέπει στους κανόνες που αδιαφορούν για την l να συμμετάσχουν στο αντίστοιχο $[C_l]$ για εξελικτικούς λόγους. Παρ' όλα αυτά, στην ενημέρωση του μεγέθους tp (γρ. 4 του Αλγ. 5.6) αποδίδει το ίδιο ποσό σε όλους τους κανόνες που δεν προβλέπουν την l λανθασμένα, είτε αυτοί αδιαφορούν, είτε αποφασίζουν σαφώς υπέρ της κατηγοριοποίησης στην l:

$$correctness(rule, l) = \begin{cases} 1, & \text{εάν ο } rule \text{ προβλέπει ορθά την } l \\ 0, & \text{εάν ο } rule \text{ προβλέπει λανθασμένα την } l \\ 1, & \text{εάν ο } rule \text{ αδιαφορεί για την } l \end{cases} \tag{5.2}$$

Ενημέρωση της εκτίμησης του μέσου μεγέθους των Correct Sets

Η μέθοδος $updateCs()$ ενημερώνει την εκτίμηση του μέσου μεγέθους των Correct Sets στα οποία συμμετέχει ένας κανόνας, κάθε φορά που αυτός συμμετέχει σε έστω και ένα $[C]$, για ένα δεδομένο $Instance$. Στην περίπτωση που συμμετέχει για το ίδιο $Instance$ σε περισσότερα του ενός $[C]$, η εκτίμηση του cs ενημερώνεται χρησιμοποιώντας την πληθικότητα $minCs$ του μικρότερου σε μέγεθος $[C]$, από αυτά στα οποία συμμετείχε ο κανόνας.

Αλγόριθμος 5.7 Ενημέρωση της εκτίμησης του μέσου μεγέθους των Correct Set στον GMl-ASLCS$_0$ (όπου β ο ρυθμός μάθησης, με σταθερή τιμή)

1: **updateCs**($rule$)
2: $minCs \leftarrow \min_{l \in L} \sum_{rule \in [C_l]} rule.num$
3: $rule.cs \leftarrow rule.cs + \beta \cdot (minCs - rule.cs)$

Συνιστωσα Εξερευνησης

Η συνιστώσα εξερεύνησης αποτελείται, όπως και στην απλή κατηγοριοποίηση, από δύο τμήματα: το Γενετικό Αλγόριθμο και το τμήμα Κάλυψης. Ο Γενετικός Αλγόριθμος δεν επιδέχεται σημαντικών αλλαγών σε σχέση με τον μονοκατηγορικό AS-LCS, καθώς μεταχειρίζεται απλώς χρωμοσώματα και αδιαφορεί για την αναπαράσταση του υποκείμενου προβλήματος. Συνεπώς, μπορεί να διατηρηθεί χωρίς δομικές τροποποιήσεις.

Γενετικός Αλγόριθμος

Ορίζουμε ως *γενετικό γεγονός ή γενετικό συμβάν* (genetic event) την παραγωγή ενός προκαθορισμένου αριθμού απογόνων, μέσω του Γενετικού Αλγορίθμου. Η συχνότητα των αποφάσεων για την τέλεση γενετικών γεγονότων υπαγορεύεται από την παράμετρο θ_{GA}, με τον ίδιο τρόπο όπως και στο πλαίσιο *S-LCS. Μικρές τιμές του θ_{GA} σημαίνουν μικρότερα διαστήματα ανάμεσα στην τρέχουσα τιμή του

χρονικού βήματος και του μέσου όρου των χρονοσφραγίδων των κανόνων που συμμετέχουν στο σύνολο στο οποίο γίνεται ο Γενετικός Αλγόριθμος. Δηλαδή, απαιτείται μικρότερο χρονικό διάστημα για την παραγωγή απογόνων και, συνεπώς, γίνονται περισσότερα γενετικά συμβάντα.

Ο GMl-ASLCS$_0$ χρησιμοποιεί επιλογή ρουλέτας (Παρ. 3.3.1), επιλέγοντας τον κανόνα i για να αποτελέσει υποψήφιο γονέα με πιθανότητα[1]

$$P(i) = \frac{num(i) \cdot fitness'(i)}{\sum_{j=1}^{n} \big(num(j) \cdot fitness'(j)\big)} \tag{5.3}$$

όπου n ο αριθμός των κανόνων του πληθυσμού.

Στην παραπάνω εξίσωση, το μέγεθος $fitness'$ είναι η καταλληλότητα του κανόνα, η οποία έχει υποστεί έκπτωση-βασισμένη-στην-εμπειρία (experience-based fitness discount), ακολουθώντας την πρακτική της βιβλιογραφίας, σύμφωνα με την εξίσωση:

$$fitness'(i) = \begin{cases} 0, & experience(i) < \theta_{exp} \\ (accuracy(i))^{\nu}, & \text{αλλού} \end{cases} \tag{5.4}$$

και

$$accuracy(i) = \frac{tp(i)}{msa(i)} \tag{5.5}$$

η ακρίβεια του κανόνα i, με $tp(i)$ τον αριθμό ορθών κατηγοριοποιήσεων του κανόνα και $msa(i)$ τον αριθμό συνολικών κατηγοριοποιήσεων του.

Στην ουσία πρόκειται για καθυστέρηση στην απόδοση εμπιστοσύνης προς τους κανόνες από το σύστημα, εξυπηρετώντας εν μέρει την αναπαραγωγή γενικών κανόνων[2] , καθώς η εμπειρία είναι ένα μέτρο της γενικότητας ενός κανόνα που παίζει ρόλο κυρίως στις πρώτες επαναλήψεις της εκπαίδευσης. Όσο πιο γενικός είναι ένας κανόνας, τόσα περισσότερα δείγματα καλύπτει και, άρα, σε τόσα περισσότερα $[M]$ συμμετέχει. Καθώς η εμπειρία ενός κανόνα αυξάνει κατά ένα σε κάθε $[M]$ που συμμετέχει, όσο πιο γενικός είναι, τόσο πιο γρήγορα θα ξεπεράσει το κατώφλι εμπειρίας θ_{exp}, άρα τόσο πιο γρήγορα θα του ανατεθεί μία μη μηδενική πιθανότητα επιλογής από το Γενετικό Αλγόριθμο.

Ένας ακόμα λόγος που συνηγορεί στη χρήση έκπτωσης βασισμένη στην εμπειρία είναι η απαίτηση μας το σύστημα να αποφασίζει για την επιλογή ενός κανόνα για αναπαραγωγή, βασισμένο σε μία πιο σαφή εικόνα για αυτόν και την ικανότητα κατηγοριοποίησής του. Σε ακραίες περιπτώσεις, θα μπορούσε να παρατηρηθεί το

[1] Ο αναγνώστης θα προσέξει τη διαφορά της χρήσης της πληθικότητας του κανόνα ανάμεσα στην παραπάνω εξίσωση και την Εξ. 3.1. Ο κανόνας στου οποίου την πιθανότητα επιλογής αναφερόμαστε, είναι ένας *μακροκανόνας*, και αυτή θα είναι η αναφορά της λέξης *κανόνας* από εδώ και στο εξής. Για την αναφορά σε έναν συγκεκριμένο κανόνα, δηλαδή σε έναν κανόνα που συνολικά είναι μέρος ενός μακροκανόνα, θα χρησιμοποιούμε τον όρο *μικρο-κανόνας*.

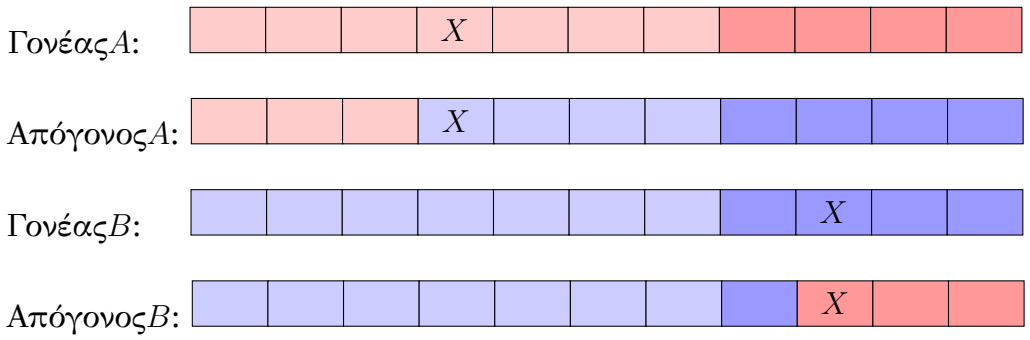
[2] Ακριβέστερα, αποτρέπει τη συμμετοχή υπερ-γενικών κανόνων στην εξελικτική διαδικασία, έως ότου η εμπειρία τους ξεπεράσει το κατώφλι θ_{exp}.

φαινόμενο ένας κανόνας να ταξινομεί ορθά ένα από τα πρώτα δείγματα που καλύπτει, αλλά να αποτυγχάνει πλήρως να ταξινομήσει μεταγενέστερα, λόγω της δομής του συνόλου δεδομένων. Αν το σύστημα τον εμπιστευόταν πλήρως εξαρχής, θα αναπαρήγαγε πιθανώς την αποτυχία του κανόνα στην ταξινόμηση εκείνων ακριβώς των δειγμάτων που θα προκαλούσαν τη μείωση της ακρίβειάς του, απομακρύνοντας έτσι το ΜαΣΤ από την αρχική μας απαίτηση για αναπαραγωγή βασισμένη στην ακρίβεια.

Τελεστές Διασταύρωσης και Μετάλλαξης

Μετά την επιλογή δύο κανόνων-γονέων, έστω A και B, εφαρμόζεται διασταύρωση ενός σημείου για την παραγωγή *δύο* απογόνων Α και Β. Το σημείο στο οποίο θα γίνει διασταύρωση επιλέγεται ξεχωριστά για τον κάθε απόγονο, ψευδο-τυχαία, και χωρίς βλάβη της γενικότητας έχει διαφορετική τιμή για τον καθένα. Μετά τον καθορισμό του σημείου διασταύρωσης, ο απόγονος A προκύπτει από τη σύνθεση α) του χρωμοσώματος του γονέα A μέχρι το σημείο διασταύρωσής του και β) του γονέα B από το σημείο διασταύρωσης μέχρι το τέλος του χρωμοσώματος. Αντίστοιχα, ο απόγονος B προκύπτει από τη σύνθεση α) του χρωμοσώματος του γονέα B μέχρι το σημείο διασταύρωσής του και β) του γονέα A από το σημείο διασταύρωσης μέχρι το τέλος του χρωμοσώματος.

Η διαδικασία απεικονίζεται στο Σχήμα 5.1, με ανοιχτό χρώμα για τα γνωρίσματα, σκούρο για τις ετικέτες και το X να υποδηλώνει το εκάστοτε σημείο διασταύρωσης.



Σχήμα 5.1: Διαδικασία Διασταύρωσης στον GMl-ASLCS$_0$.

Ο τελεστής της διασταύρωσης εφαρμόζεται με πιθανότητα $\chi = 0.8$. Σε περίπτωση απόφασης μη διασταύρωσης, οι απόγονοι βγαίνουν αλώβητοι από τη διαδικασία της διασταύρωσης, ως κλώνοι των γονέων τους. Σε κάθε περίπτωση, στη συνέχεια, και κατά τα γνωστά, εφαρμόζεται ο τελεστής της *ομοιόμορφης μετάλλαξης*, δηλαδή κάθε γνώρισμα και ετικέτα υπόκειται ξεχωριστά σε μετάλλαξη, με πιθανότητα $\mu = 0.04$.

Διαδικασία εισαγωγής απογόνων στον πληθυσμό

Κάθε κανόνας που έχει δημιουργηθεί μέσω του Γενετικού Αλγορίθμου, προτού εισαχθεί στον πληθυσμό (συνάρτηση $insert$ στον Αλγ. 5.4), ελέγχεται για *υπαγωγή* (subsumption) ως προς το σύνολο των κανόνων. Εάν υπάρχει κάποιος κανόνας r με εμπειρία μεγαλύτερη από ένα κατώφλι θ_{sub}, με καταλληλότητα μεγαλύτερη από ένα κατώφλι καταλληλότητας α_0, το ίδιο γενικός ή γενικότερος από το νέο κανόνα (απόγονο) στο τμήμα συνθήκης και, ταυτόχρονα, το ίδιο ειδικός ή ειδικότερος στο τμήμα απόφασης, ο απόγονος δεν εισάγεται στον πληθυσμό, αλλά αφομοιώνεται από τον r, αυξάνοντας την πληθικότητα του τελευταίου κατά ένα. Σε αντίθετη περίπτωση, ο κανόνας-απόγονος εισάγεται απευθείας στον πληθυσμό ως αυτοτελής κανόνας, χωρίς κάποια περαιτέρω διαδικασία.

Στην περίπτωση της δημιουργίας κανόνων μέσω της λειτουργίας κάλυψης, οι απόγονοι παρακάμπτουν τη διαδικασία αφομοίωσης και εισάγονται αυτοτελώς στον πληθυσμό.

Λειτουργία Διαγραφής

Όπως και στα σχήματα αναπαραγωγής των UCS και AS-LCS, έτσι και εδώ, το σύστημα εξελίσσει ένα σύνολο από κανόνες του οποίου το άθροισμα των πληθικοτήτων παραμένει κάτω από ένα συγκεκριμένο όριο, εκ των προτέρων επιλεγμένο από τον χρήστη. Κάθε φορά που συμβαίνει ένα γενετικό γεγονός, σύμφωνα με τα παραπάνω, έχουμε ως αποτέλεσμα την αύξηση του συνολικού αριθμού των κανόνων (και των μικρο-κανόνων) κατά δύο (δύο απόγονοι), είτε το αποτέλεσμα του γενετικού συμβάντος αφομοιωθεί, είτε εισαχθεί ακέραιο στον πληθυσμό. Στην περίπτωση που ο αριθμός των μικρο-κανόνων, πριν την εισαγωγή ενός απογόνου στον πληθυσμό είναι ίσος με το μέγιστο αριθμό μικρο-κανόνων που μπορεί να συγκρατήσει το σύστημα, ενεργοποιείται η λειτουργία της διαγραφής (γραμμές 20 και 21 στον Αλγ. 5.4), η οποία επιλέγει κανόνες προς διαγραφή χρησιμοποιώντας επιλογή ρουλέτας. Σε κάθε κανόνα ανατίθεται μία πιθανότητα διαγραφής ίση με

$$P(i) = \frac{num(i) \cdot d(i)}{\sum_{j=1}^{n} \big(num(j) \cdot d(j)\big)} \tag{5.6}$$

όπου i ένας τυχαίος κανόνας, n ο συνολικός αριθμός κανόνων του πληθυσμού και

$$d(i) = \begin{cases} 0, & coverage(i) = 0 \text{ ή } (coverage(i) = 1 \text{ και } experience(i) = 1) \\ \dfrac{1}{100 \cdot fitness(i)}, & experience(i) < \theta_{del} \\ e^{cs(i)} - 1, & \text{αλλού} \end{cases} \tag{5.7}$$

όπου $coverage(i)$ το ποσοστό δειγμάτων του συνόλου εκπαίδευσης που καλύπτει ο κανόνας i, $experience(i)$ η εμπειρία του και $fitness(i)$ η τιμή καταλληλότητάς του (χωρίς έκπτωση βασισμένη στην εμπειρία, όπως αναφέρεται στη γρ. 7 του Αλγ. 5.6).

Λειτουργία Κάλυψης

Η λειτουργία της κάλυψης έχει σκοπό την κατασκευή κανόνων με βάση το τρέχον δείγμα s, με το οποίο εκπαιδεύεται το σύστημα. Ενεργοποιείται για κάθε ετικέτα όταν δεν υπάρχει κανένας κανόνας που να καλύπτει το δείγμα s. Επιπλέον, ενεργοποιείται για τις ετικέτες για τις οποίες υπάρχουν κανόνες που καλύπτουν το δείγμα, αλλά η απόφασή τους δεν συνάδει με αυτή του δείγματος (κενό $[C_l]$). Στη μονοκατηγορική ταξινόμηση, ο τελεστής κάλυψης δημιουργεί έναν κανόνα γενικεύοντας τυχαία τα γνωρίσματα του s και μεταφέροντας αυτούσια την κλάση του δείγματος ως απόφαση του κανόνα. Στην πολυκατηγορική ταξινόμηση, όμως, δεδομένης της αναπαράστασης με αδιαφορίες που επεκτείνεται και στο τμήμα απόφασης των κανόνων, υπάρχει η δυνατότητα γενίκευσης και με βάση το τμήμα ετικετών των δειγμάτων. Έτσι, ένα γνώρισμα γενικεύεται (απενεργοποιείται) με πιθανότητα $P_{\#A}$, ενώ μία ετικέτα μπορεί να απενεργοποιηθεί και αυτή, με πιθανότητα $P_{\#L}$.

Στα περισσότερα πραγματικά προβλήματα, $P_{\#L} \ll P_{\#A}$, καθώς, δεν θα θέλαμε να ωθήσουμε το σύστημα σε μία κατεύθυνση όπου θα συγκρατούσε ένα πληθυσμό από γενικούς κανόνες στο τμήμα απόφασης, σε βάρος λίγων, ειδικών στο τμήμα απόφασής τους κανόνων. Αυτή η απαίτηση γίνεται κατανοητή και από τη σκοπιά της ψηφοφορίας, που περιγράφεται στην Παρ. 5.4.1. Ανεξαρτήτως, όμως, των πιθανοτήτων γενίκευσης $P_{\#L}$ και $P_{\#A}$, ένας τρόπος για να οδηγήσουμε το σύστημα στη δημιουργία κανόνων γενικών στο τμήμα συνθήκης και ειδικών στο τμήμα απόφασης είναι η "τιμωρία" των κανόνων για κάθε αδιαφορία στο τμήμα απόφασης τους.

Αφαίρεση των κανόνων μηδενικής κάλυψης

Ενδιαφέρον παρουσιάζουν οι γραμμές 24 και 25 του Αλγ. 5.4. Αφού ενημερωθούν οι παράμετροι των κανόνων που συμμετέχουν στο Match Set και τα διάφορα Correct Sets, φροντίζουμε να απομακρύνουμε από τον πληθυσμό τους κανόνες μηδενικής κάλυψης, δηλαδή εκείνους τους κανόνες στους οποίους έχουν παρουσιαστεί όλα τα δείγματα του συνόλου δεδομένων, αλλά δεν έχουν συμμετάσχει σε κανένα σχηματιζόμενο Match Set. Κάθε τέτοιος κανόνας χαρακτηρίζεται από την απουσία νοήματος πάνω στο πρόβλημα, καθώς δεν καλύπτει ούτε ένα δείγμα του συνόλου εκπαίδευσης. Τέτοιοι κανόνες παράγονται μόνο από το Γενετικό Αλγόριθμο, και όχι από το τμήμα Κάλυψης[3].

Ένα μέρος των κανόνων μηδενικής κάλυψης θα διαγραφεί από τον πληθυσμό, μετά το πέρας του κύκλου ενημέρωσης και παραγωγής κανόνων, με πιθανότητα αντιστρόφως ανάλογη του μεγέθους του συνόλου εκπαίδευσης D. Οι κανόνες που

[3]Η παραγωγή κανόνων μηδενικής κάλυψης, στην ουσία οφείλεται στη μη πληρότητα των πραγματικών συνόλων δεδομένων, δηλαδή στην απουσία δειγμάτων που να καθιστούν το σύνολο δεδομένων πλήρες ως προς όλους τους συνδυασμούς των τιμών των γνωρισμάτων.

θα διαγραφούν έχουν δύο ιδιότητες: α) κάθε κανόνας πρέπει να έχει κληθεί να συμμετάσχει τουλάχιστον μία φορά σε Match Set για *κάθε* δείγμα του D και β) ο χρόνος παραγωγής του κανόνα θα πρέπει να είναι τέτοιος ώστε ο αριθμός κλήσεων συμμετοχής του σε $[M]$, δηλαδή ο αριθμός δειγμάτων που έχουν παρουσιαστεί στον κανόνα, να είναι ακέραιο πολλαπλάσιο του αριθμού δειγμάτων του D.

Αρχικοποίηση Παραμέτρων Κανόνων

Οι κανόνες που παράγονται μέσω της λειτουργίας κάλυψης έχουν αρχικές παραμέτρους $(tp, msa, cs, fitness) \equiv (0, 0, 20, 0.5)$, ενώ αυτοί που δημιουργούνται μέσω του γενετικού αλγορίθμου $(0, 0, (parentA.cs + parentB.cs)/2, 0.5)$. Δηλαδή, ένας κανόνας που παράγεται μέσω του Γενετικού Αλγορίθμου κληρονομεί ως cs το μέσο όρο των cs των δύο γονέων του, $parentA$ και $parentB$. Η αρχική καταλληλότητα ασκεί (περιορισμένα) επιρροή μόνο στη λειτουργία της διαγραφής, μέχρι ένας κανόνας να συμμετάσχει σε κάποιο Match Set.

Συνιστωσα Επιδοσης

Η συνιστώσα επίδοσης, όπως αναφέραμε στην Παρ. 2.5.6, είναι υπεύθυνη για την κατηγοριοποίηση ενός αγνώστου δείγματος με βάση το σύνολο των κανόνων που έχει εξελίξει το ΜαΣΤ. Σε αυτή την παράγραφο αναλύουμε τη μέθοδο ψηφοφορίας των κανόνων που χρησιμοποιεί ο GMl-ASLCS$_0$ και τη διαδικασία ρύθμισης του κατωφλίου με τη μέθοδο PCut.

Μέθοδος Ψηφοφορίας

Η διαδικασία ξεκινάει με την είσοδο ενός άγνωστου δείγματος s στο σύστημα. Από το σύνολο των κανόνων του πληθυσμού, συγκεντρώνονται στο $[M]$, κατά τα γνωστά, οι κανόνες που καλύπτουν το s. Στη συνέχεια, σχηματίζεται ο μονοδιάστατος, κενός πίνακας ψηφοφορίας A, μεγέθους $|L|$, όπου $|L|$ ο αριθμός των ετικετών των δειγμάτων. Ακολούθως, κάθε κανόνας του $[M]$ καλείται να ψηφίσει για την κάθε ετικέτα ξεχωριστά, με τον εξής τρόπο: οι κανόνες που συνηγορούν στην κατηγοριοποίηση στην ετικέτα $l \in L$ ψηφίζουν με θετικό πρόσημο, με την ποσότητα $num \times fitness$. Οι κανόνες που αρνούνται την κατάταξη του δείγματος στην l ψηφίζουν με την ανωτέρω ποσότητα, αλλά με αρνητικό πρόσημο και αυτοί που αδιαφορούν για την l δεν συμβάλλουν καθόλου στην ψηφοφορία.

Για να γίνει περισσότερο κατανοητή η συνέχεια της διαδικασίας, ας υποθέσουμε ένα δείγμα $s_0 : x \Rightarrow 100$, $[M] = \{r_0, r_1\}$, $r_0 : x_0 \Rightarrow 110$ και $r_1 : x_1 \Rightarrow 1\#1$, με τις εξής παραμέτρους: $num_0 = 10$, $num_1 = 20$, $fitness_0 = 1$, $fitness_1 = 0.8$, για ένα πρόβλημα τριών ετικετών ($|L| = 3$). Τότε, ο πίνακας A_0 που σχηματίζεται μετά την ψηφοφορία για το δείγμα s_0 είναι ο εξής:

$1 \cdot 10 + 0.8 \cdot 20$	$-1 \cdot 10$	$1 \cdot 10 - 0.8 \cdot 20$

δηλαδή

26	-10	-6

Στη συνέχεια, η απόλυτη τιμή της μικρότερης αρνητικής ποσότητας ανάμεσα στις ετικέτες προστίθεται σε όλα τα αποτελέσματα και γίνεται κανονικοποίηση, χρησιμοποιώντας το άθροισμα των ψήφων:

36	0	4

και

36/40	0	4/40

καταλήγοντας στον πίνακα

0.9	0	0.1

Για κάθε δείγμα του συνόλου ελέγχου, λοιπόν, γίνεται η παραπάνω διαδικασία, με αποτέλεσμα ένα διδιάστατο πίνακα, με τόσες γραμμές όσα είναι τα δείγματα του συνόλου ελέγχου και τόσες στήλες όσες είναι ο αριθμός των ετικετών. Ακολούθως, υπολογίζεται το κατώφλι, το οποίο χαράσσει με οριζόντιο τρόπο την τελική κατηγοριοποίηση: δείγματα που στον πίνακα ψηφοφορίας τους A έχουν τιμές μεγαλύτερες από το κατώφλι στις θέσεις που αντιστοιχούν σε συγκεκριμένες ετικέτες, κατηγοριοποιούνται στις ετικέτες αυτές. Προφανώς, δείγματα που στον πίνακα A τους έχουν τιμές μικρότερες από το κατώφλι στις θέσεις που αντιστοιχούν σε συγκεκριμένες ετικέτες, δεν κατηγοριοποιούνται σε αυτές.

Σε συνέχεια του παραδείγματος, έστω ότι το υπολογιζόμενο κατώφλι παίρνει τιμή $t = 0.2$, τότε το δείγμα κατηγοριοποιείται πλήρως ορθά, καθώς το σύστημα αποφαίνεται πως πρέπει να κατηγοριοποιηθεί μόνο στην πρώτη ετικέτα και όχι στις άλλες δύο, όπως είναι και η πραγματική κατηγοριοποίηση για το s_0.

Στην παραπάνω διαδικασία, η καταλληλότητα $fitness$ είναι μεν συνάρτηση της ακρίβειας του κάθε κανόνα, μπορεί, όμως, να μην ταυτίζεται με το μέγεθος που χρησιμοποιήθηκε για την εξελικτική διαδικασία (Εξ. 5.4). Συγκεκριμένα, στο [MBK07] αναφέρεται πως διαφορετικά προβλήματα, απαιτούν διαφορετικές προσεγγίσεις για τη βελτιστοποίηση της κατηγοριοποίησης, όσον αφορά στην παράμετρο ν. Η παράμετρος ν χρησιμοποιείται για να δώσει μεγαλύτερη βαρύτητα στην ακρίβεια των κανόνων, ώστε να τους καταστήσει περισσότερο διαχωρίσιμους στην επιλογή τους από τις λειτουργίες του συστήματος, στρέφοντας την εξελικτική διαδικασία προς την ανακάλυψη κανόνων ακριβέστερων από τους προκατόχους τους. Στον GMl-ASLCS$_0$, η παράμετρος ν στη διαδικασία της ψηφοφορίας λαμβάνεται ίση με ένα, δηλαδή η ψηφοφορία, στην ουσία, γίνεται βάσει της *ακρίβειας* των κανόνων και όχι της καταλληλότητάς τους. Επιπλέον, δεν εφαρμόζεται κάποιου είδους έκπτωση βασισμένη στην εμπειρία, αλλά χρησιμοποιείται η πραγματική τιμή της ακρίβειας των κανόνων.

Από τη διαδικασία της ψηφοφορίας εξάγουμε και μία εικόνα για τη βαρύτητα που έχει η εξέλιξη κανόνων με αδιαφορίες στο τμήμα απόφασής τους, όσον αφορά στην ικανότητα ορθής κατηγοριοποίησης από ένα ΜαΣΤ. Εάν για ένα δείγμα ενεργοποιείται ένα σύνολο κανόνων με μεγάλο περιεχόμενο αδιαφοριών στις ετικέτες,

οι μη μηδενικές ψήφοι μειώνονται δραστικά, επιτρέποντας έτσι σε λίγους κανόνες να καθορίσουν στην ουσία, πιο άμεσα, την κατηγοριοποίηση των αγνώστων δειγμάτων και κάνοντας, επιπρόσθετα, τη διαδικασία εύρεσης κατωφλίου δυσχερέστερη. Αυτό το φαινόμενο ενισχύεται σε προβλήματα χαμηλής κατηγορικής πληθικότητας (σε προβλήματα δηλαδή όπου ενεργοποιείται ένας μικρός αριθμός ετικετών για κάθε δείγμα), καθώς η ελάχιστη αρνητική τιμή που αφαιρείται από όλα τα αποτελέσματα της ψηφοφορίας προστίθεται και στις ψήφους των ετικετών για τις οποίες δεν υπάρχει ουσιαστική ψηφοφορία (των ετικετών, δηλαδή, για τις οποίες υπάρχουν ελάχιστες μη μηδενικές ψήφοι). Τελικά, το δείγμα κατατάσσεται σε ένα σύνολο ετικετών που περιλαμβάνει ορθά τις λίγες ετικέτες στις οποίες ανήκει πραγματικά και, με μεγάλη πιθανότητα, λανθασμένα τις ετικέτες για τις οποίες υπήρξε μεγάλη αποχή στην ψηφοφορία.

Μέθοδος ρύθμισης κατωφλίου PCut

Όπως αναφέραμε και στην Παρ. 2.5.6, η μέθοδος PCut, η οποία παρουσιάζεται σε μορφή ψευδοκώδικα στον Αλγ. 5.8, προσπαθεί να επιλέξει το σωστό αριθμό ετικετών κατά μέσο όρο για ένα σύνολο δεδομένων G. Ο αλγόριθμος εκτελείται επαναληπτικά για *iterations* βήματα, και προσεγγίζει σταδιακά την αναζητούμενη τιμή κατωφλίου (παράμετρος $center$) με ολοένα μειούμενο βήμα *step*, αξιοποιώντας την κυρτότητα της συνάρτησης

$$err(th, LC) = \left| LC(D) - \left(\frac{1}{|G|} \sum_{i=1}^{|G|} |f_{th}(\bar{w}_i)| \right) \right| \tag{5.8}$$

όπου LC είναι η κατηγορική πληθικότητα του συνόλου εκπαίδευσης D και G το σύνολο δεδομένων με βάση το οποίο γίνεται η ρύθμιση του κατωφλίου.

Κλείνοντας, αξίζει να σημειώσουμε πως, στην περίπτωσή μας, τα σύνολα D και G ταυτίζονται πάντα, καθώς η χρήση του συνόλου ελέγχου στη θέση του G υπονοεί εκ των προτέρων γνώση της δομής των ετικετών σε μη επισημασμένα δείγματα και οδηγεί σε υπερβολικά επιεικείς αξιολογήσεις.

Αλγόριθμος 5.8 Μέθοδος ρύθμισης κατωφλίου Pcut

```
1: Pcut(center, step, iterations, LC)
2: if iterations = 0 then
3:    return
4: end if
5: fleft ← center − step/2
6: cleft ← center − step/4
7: fright ← center + step/2
8: cright ← center + step/4
9: next ← err(fleft, LC)
10: minError ← err(fleft, LC)
11: for point ∈ {cleft, center, cright, fright} do
12:    if minError > err(point, LC) then
13:       next ← point
14:       minError ← err(point, LC)
15:    end if
16: end for
17: Pcut(next, step/2, iterations − 1, LC)
```

6

GMl-ASLCS: Η εξέλιξη του GMl-ASLCS$_0$

Στο προηγούμενο κεφάλαιο εξετάσαμε βασικές λειτουργίες από τα ενδότερα του ΜαΣΤ GMl-ASLCS$_0$. Αυτό έγινε, αφενός, για να γίνουν περισσότερο κατανοητές οι εν λόγω λειτουργίες, αφετέρου, για να καταστούν κατανοητοί οι λόγοι για τις τροποποιήσεις που επιφέρουμε σε αυτές, καταλήγοντας στον αλγόριθμο GMl-ASLCS.

Όσον αφορά στην **αναπαράσταση των κανόνων** του πληθυσμού που εξελίσσει, ο GMl-ASLCS χρησιμοποιεί την ίδια μέθοδο αναπαράστασης με τον GMl-ASLCS$_0$. Ακόμα, η **συνιστώσα επίδοσης** παραμένει αναλλοίωτη, όπως και το **Τμήμα Κάλυψης** ως προς τη λειτουργία του. Τέλος, για την **πρόβλεψη των ετικετών αγνώστων δειγμάτων** χρησιμοποιούνται οι τρεις μέθοδοι που έχουν παρουσιαστεί συνολικά: η *επιλογή βέλτιστου κανόνα* (Παρ. 2.5.6), η *ψηφοφορία μέσου όρου με επιλογή κατωφλίου με τη μέθοδο PCut* (Παρ. 5.4.2) και η *ψηφοφορία μέσου όρου με επιλογή κατωφλίου με τη μέθοδο IVal* (Παρ. 2.5.6).

Σε αυτό το κεφάλαιο:

- Περιγράφουμε σε μεγαλύτερο βάθος την εκπαιδευτική διαδικασία του GMl-ASLCS
- Εστιάζουμε στους κανόνες μηδενικής κάλυψης και στα αποτελέσματα που επιφέρει η διατήρηση τέτοιων κανόνων στον πληθυσμό των ΜαΣΤ, προτείνοντας την εφαρμογή μίας μεθόδου που τα αντιμετωπίζει άμεσα.
- Παρουσιάζουμε τις δύο βασικές πρωτοτυπίες που εισάγει αυτή η εργασία: α) έναν ορθολογικό τελεστή διασταύρωσης που συνάδει με τη φύση των πολυκατηγορικών προβλημάτων και β) μία μέθοδο αύξησης του μέσου αριθμού δειγμάτων που καλύπτονται από τους κανόνες των ΜαΣΤ.

- Προτείνουμε μερικές περαιτέρω τροποποιήσεις που μπορούν να εφαρμοστούν αρθρωτά και αφορούν σε κατευθύνσεις που θα μπορούσαν να ακολουθηθούν στο μέλλον.

Ο Κυκλος Εκπαιδευσης του GML-ASLCS

Ο κύκλος εκπαίδευσης του GMl-ASLCS παρουσιάζεται στον Αλγ. 6.1. Εκ πρώτης όψεως, εμφανίζεται να είναι ίδιος με αυτόν του GMl-ASLCS$_0$ (Αλγ. 5.1 και 5.4), με τις κύριες διαφορές να εντοπίζονται στις γραμμές 3 έως 5, 16, 20, 26 έως 28 έως 30. Οι επόμενες ενότητες αναλύουν τις επιμέρους χρησιμοποιούμενες μεθόδους.

Αλγόριθμος 6.1 Ο κύκλος εκπαίδευσης του GMl-ASLCS

```
1: train(P, Instance)
2: M ← generateMatchSet(P, Instance)
3: if rouletteWheelDeletionsCommenced = true then
4:    controlMatchSet(P, M)
5: end if
6: for each l ∈ L do
7:    C[ l ] ← generateLabelCorrectSet(M, l, Instance)
8: end for
9: for each rule ∈ M do
10:   updateFitness(rule)
11:   if ∃l_i ∈ L : rule ∈ C[ l_i ] then
12:      updateCs(rule)
13:   end if
14: end for
15: S, S̄ ← ∅
16: for each l ∈ L do
17:   if C[ l ] ≠ ∅ then
18:      if timestamp − timestamp([C_l]) > θ_GA then
19:         evolve(C[ l ], P, S, S̄)
20:      end if
21:   else
22:      cover(Instance, l)
23:   end if
24: end for
25: P ← merge(P, S̄)
26: P.subsume(S)
27: if |P| > maximumPopulationSize then
28:   controlPopulation(P)
29: end if
```

Συνιστωσα Εξερευνησης

Η Συνιστώσα Εξερεύνησης περιλαμβάνει το Γενετικό Αλγόριθμο και τις επιμέρους λειτουργίες του, καθώς και το Τμήμα Κάλυψης, το οποίο, όπως ήδη αναφέρθηκε, παραμένει ίδιο με αυτό του GMl-ASLCS$_0$.

Ο Γενετικός Αλγόριθμος χρησιμοποιεί, όπως ακριβώς και στον GMl-ASLCS$_0$, επιλογή ρουλέτας για την εύρεση των υποψήφιων γονέων, με τον κανόνα i να έχει πιθανότητα επιλογής που δίνεται από τη σχέση:

$$P(i) = \frac{num(i) \cdot fitness'(i)}{\sum_{j=1}^{n} \big(num(j) \cdot fitness'(j)\big)} \tag{6.1}$$

όπου n ο αριθμός των κανόνων του πληθυσμού. Η καταλληλότητά των κανόνων είναι και εδώ συνάρτηση της εμπειρίας τους:

$$fitness'(i) = \begin{cases} 0, & experience(i) < \theta_{exp} \\ (accuracy(i))^{\nu}, & \text{αλλού} \end{cases} \tag{6.2}$$

Τελεστής Διασταύρωσης

Η μέχρι πρότινος προσέγγιση της διασταύρωσης μεταξύ των κανόνων-γονέων, χρησιμοποιούσε διασταύρωση ενός σημείου (single-point crossover), θεωρώντας ως μήκος του χρωμοσώματος το σύνολο του αριθμού των δυφίων που είναι απαραίτητα για την αναπαράσταση των γνωρισμάτων και των ετικετών του προβλήματος, $|attributes|$ και $|labels|$ αντίστοιχα. Το μέγεθος του χρωμοσώματος των κανόνων chromosomeSize είναι το άθροισμα των δύο παραπάνω μεγεθών:

$$chromosomeSize = \sum_{i=1}^{number\ of\ attributes} (bits\ of\ representation_i) + 2 \cdot (number\ of\ labels) \tag{6.3}$$

Η μέθοδος διασταύρωσης ενός σημείου, όπως είδαμε στην Παρ. 5.3.1, για την παραγωγή κάθε απογόνου, επιλέγει ψευδο-τυχαία ένα σημείο μέσα στο μήκος του χρωμοσώματος και εναλλάσσει *όλα* τα γονίδια, πέραν αυτού του σημείου, ανάμεσα στους δύο γονείς. Η πιθανότητα το σημείο διασταύρωσης χ να βρίσκεται μέσα στον χώρο των γνωρισμάτων είναι

$$P(\chi < |attributes|) = \frac{|attributes|}{|attributes| + |labels|} = \frac{|attributes|}{chromosomeSize} \tag{6.4}$$

Ωστόσο, στα περισσότερα πραγματικά σύνολα δεδομένων, ο αριθμός των γνωρισμάτων είναι τουλάχιστον μία τάξη μεγέθους μεγαλύτερος από τον αριθμό των

Πίνακας 6.1: Πιθανότητες επιλογής του σημείου διασταύρωσης μέσα στο χώρο των γνωρισμάτων για διασταύρωση ενός σημείου, χρησιμοποιώντας 5 bits για την αναπαράσταση αριθμητικών γνωρισμάτων.

Dataset	Attributes' type	Number of attributes	Number of labels	$P(\chi < \|attributes\|)$
Enron	nominal	1001	53	0.949
Medical	binary	1449	45	0.970
Yeast	nominal	103	14	0.976
Genbase	binary	1185	27	0.978
Music	nominal	72	6	0.985
Scene	binary	294	6	0.996

ετικετών, με αποτέλεσμα το σημείο διασταύρωσης να βρίσκεται μέσα στο χώρο των γνωρισμάτων, πολλές φορές, σχεδόν νομοτελειακά. Ενδεικτικές τιμές της πιθανότητας που δίνεται από την Εξ. 6.4, για τα πραγματικά σύνολα δεδομένων που μελετήσαμε σε αυτή την εργασία[1], παρουσιάζονται, με αύξουσα σειρά, στον Πίνακα 6.1.

Εφόσον, λοιπόν, δεδομένης της επιλογής για διασταύρωση των γονέων, ο Γενετικός Αλγόριθμος διασταυρώνει τους γονείς κατά κύριο λόγο στο τμήμα των γνωρισμάτων, αυτό σημαίνει πως το τμήμα των αποφάσεων μεταφέρεται αυτούσιο από κάθε γονέα στον απόγονο που του αντιστοιχεί. Αυτό δυσχεραίνει εξελικτικά το έργο της εξερεύνησης, καθώς αναπαράγονται και πιθανόν λανθασμένες αποφάσεις για ετικέτες διαφορετικές από αυτήν για την οποία σχηματίστηκε το Correct Set από το οποίο επιλέχθηκαν οι γονείς. Αυτές οι λανθασμένες αποφάσεις, όμως, είναι σύμφυτες με την ικανότητα γενίκευσης των κανόνων: διαισθητικά, αντιλαμβάνεται κανείς εύκολα ότι όσα περισσότερα δείγματα καλύπτει ένας κανόνας, τόσο λιγότερο πιθανό είναι να αποφασίζει ορθά για όλες τις ετικέτες αυτών των δειγμάτων. Υπερ-ειδικοί κανόνες έχουν το πλεονέκτημα εξαιρετικής επίδοσης μάθησης πάνω στα δεδομένα εκπαίδευσης, ωστόσο αποτυγχάνουν να κατηγοριοποιήσουν ορθά δείγματα τα οποία δεν έχουν παρουσιαστεί στο σύστημα, λόγω ακριβώς της πολύ στενής τους σχέσης με πολύ συγκεκριμένες παρατηρήσεις, οδηγώντας έτσι στην εμφάνιση του φαινομένου της υπερ-εκπαίδευσης (over-fitting).

Ουσιαστικά, λοιπόν, το σύστημα πρέπει να ισορροπήσει πάνω στις απαιτήσεις της ακρίβειας και της γενίκευσης. Εφόσον σκοπός μας είναι η όσο το δυνατόν ακριβέστερη γενίκευση (ή, αλλιώς, η κατασκευή ορθών μοντέλων με ταυτόχρονη ικανότητα γενίκευσης), θεωρούμε δεδομένη την παρουσία και λανθασμένων αποφάσεων από τους κανόνες. Δεδομένης της πλήρους εναλλαγής του συνόλου των ετικετών στην πλειοψηφία των διασταυρώσεων, το αποτέλεσμα είναι η ανακύκλωση ίδιων συνόλων αποφάσεων, ενώ οι πιθανόν λανθασμένες αποφάσεις κληροδοτούνται από

[1]Χρησιμοποιούμε 11 bits για την αναπαράσταση αριθμητικών γνωρισμάτων, λόγω της χρήσης αναπαράστασης άνω και κάτω φράγματος. Για την αναπαράσταση των δύο θέσεων χρησιμοποιούνται $5 \cdot 2$ bits, συν ένα για το ψηφίο ενεργοποίησης. Χρησιμοποιούνται 2 bits για δυαδικα γνωρίσματα και 2 bits για τις ετικέτες

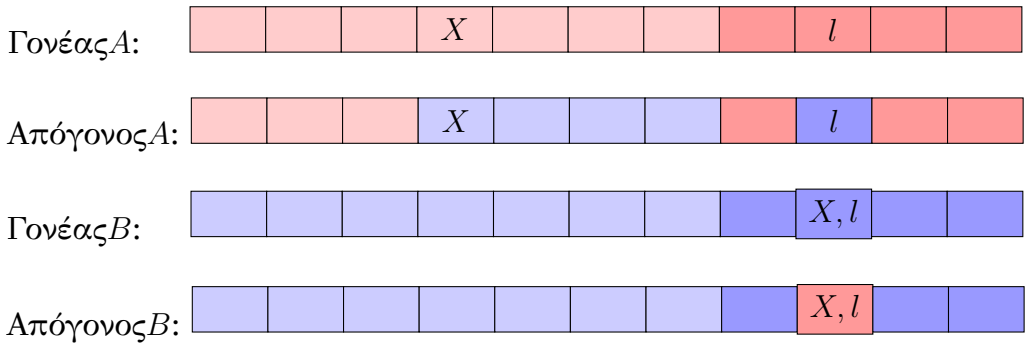
τους γονείς προς τους απογόνους, δυσκολεύοντας περαιτέρω την εξερεύνηση, τη σύγκλιση του Γενετικού Αλγορίθμου και την εξελικτική πορεία προς κανόνες περισσότερο κατάλληλους, στο πέρασμα των επαναλήψεων.

Από μία διαφορετική οπτική γωνία, λόγω της μεταφοράς του συνόλου των ετικετών των γονέων προς τους απογόνους, η διασταύρωση ενός σημείου υποθέτει πως κανόνες που συμμετέχουν σε ένα Correct Set παίρνουν το ίδιο σύνολο αποφάσεων για όλες τις ετικέτες ενός μεγάλου φάσματος δειγμάτων. Για κάτι τέτοιο, όμως, δε φαίνεται να υπάρχει κάποια βάση, είτε αναλύοντας εσωτερικά τα σύνολα δεδομένων, είτε κατά τη διαδικασία εκπαίδευσης. Δεν φαίνεται, δηλαδή, ότι μπορούν να παραχθούν κανόνες που να γενικεύουν με απολύτως ορθό τρόπο πάνω στα δείγματα του συνόλου δεδομένων.

Προσβλέποντας στην αναχαίτιση όλων των ανωτέρω ζητημάτων, κατασκευάσαμε έναν ορθολογικότερο και περισσότερο συμβατό με τη φύση των πολυκατηγορικών προβλημάτων τρόπο για την εκτέλεση της διασταύρωσης.

Η *Διασταύρωση Δύο Τμημάτων* θεωρεί ως μέγεθος του χρωμοσώματος των κανόνων τον αριθμό δυφίων αναπαράστασης των γνωρισμάτων του, συν μία ετικέτα, δηλαδή συν δύο δυφία. Αυτή η ετικέτα αντιστοιχεί στην ετικέτα για την οποία σχηματίστηκε το Correct Set από το οποίο επιλέχθηκαν οι γονείς προς αναπαραγωγή. Το σημείο διασταύρωσης επιλέγεται ψευδο-τυχαία ως μία ακέραιη θέση μέσα στο παραπάνω μέγεθος. Στη συνέχεια, αν το σημείο διασταύρωσης βρεθεί, όπως είναι και το πιο πιθανό, στο χώρο των γνωρισμάτων του χρωμοσώματος, οι δύο κανόνες-γονείς εναλλάσσουν α) τα δύο τμήματα των συνθηκών τους που οριοθετούνται από την αρχή του χρωμοσώματος, το σημείο διασταύρωσης και το τέλος του τμήματος συνθηκών του χρωμοσώματος και β) την απόφαση για την ετικέτα l για την οποία σχηματίστηκε το $[C_l]$. Στην περίπτωση που το σημείο διασταύρωσης τύχει να αντιστοιχεί στην l, εναλλάσσεται μόνο αυτή και κανένα γνώρισμα. Η υιοθέτηση πολλαπλών τμημάτων διασταύρωσης από τη Διασταύρωση Δύο Τμημάτων αναμένουμε ότι θα αποτελέσει ένα βοηθητικό παράγοντα στην καλύτερη εξερεύνηση του χώρου αναζήτησης.

Η Διασταύρωση Δύο Τμημάτων, για ξεχωριστό σημείο διασταύρωσης X για τον κάθε απόγονο και την ετικέτα για την οποία σχηματίστηκε το Correct Set $[C_l]$ από το οποίο επιλέχθηκαν οι γονείς να επισημαίνεται ως l, παρουσιάζεται στο Σχήμα 6.1.

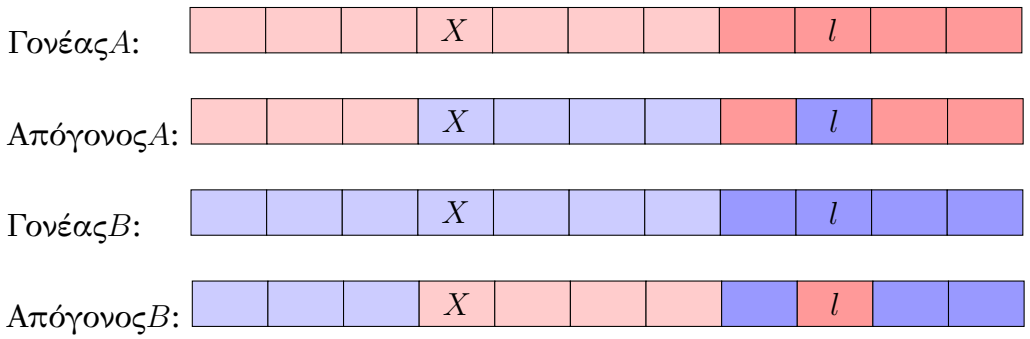


Σχήμα 6.1: Διασταύρωση Δύο Τμημάτων, με διαφορετικά σημεία διασταύρωσης ανά απόγονο στον GMl-ASLCS

Με τον νέο τελεστή Διασταύρωσης Δύο Τμημάτων καταφέρνουμε να μετασχηματίσουμε τη μεταφορά των αποφάσεων για το σύνολο των ετικετών (στο οποίο ενδέχεται να περιλαμβάνονται και λανθασμένες αποφάσεις) σε μεταφορά μόνο των ορθών αποφάσεων για κάθε ετικέτα ξεχωριστά. Πιο συγκεκριμένα, με τη Διασταύρωση Δύο Τμημάτων εναλλάσσονται οι προβλέψεις τις οποίες θεωρούμε τόσο ορθές, ώστε να περιλάβουμε στα αντίστοιχα Correct Set τους κανόνες που τις λαμβάνουν. Για την ακρίβεια, ο αναγνώστης θα παρατηρήσει πως στην πραγματικότητα, δεδομένου ότι δεν επιτρέπουμε στους κανόνες που αδιαφορούν για ετικέτες να συμμετάσχουν στα $[C_l]$, εναλλάσσονται οι ίδιες, σαφώς ορθές αποφάσεις.

Συνολικά, η λειτουργία της διασταύρωσης δύο τμημάτων δε βασίζεται μόνο στην κληροδότηση ορθών αποφάσεων, αλλά και στη μη κληροδότηση λανθασμένων αποφάσεων που παίρνουν κανόνες-γονείς, για τους οποίους δεν μπορούμε να υποθέσουμε ότι παίρνουν το ίδιο σύνολο αποφάσεων, ενώ καλύπτουν, εν γένει, διαφορετικά σύνολα δειγμάτων και συμμετέχουν σε διαφορετικά $[M]$ και $[C]$ ο καθένας. Η Διασταύρωση Δύο Τμημάτων μπορεί, έτσι, να θεωρηθεί η άμεση επέκταση στον πολυκατηγορικό χώρο του τελεστή Διασταύρωσης Ενός Σημείου στη μονοκατηγορική ταξινόμηση, που χρησιμοποιείται στους UCS και AS-LCS, ενώ συμβάλει στην επίτευξη ανεξαρτησίας από την πολυκατηγορική πληθικότητα του συνόλου εκπαίδευσης. Η εναλλαγή των γονιδίων περιλαμβάνει μόνο αυτά που αντιστοιχούν στα γνωρίσματα των κανόνων-γονέων και όχι στις ετικέτες για τις οποίες οι δύο γονείς, εν γένει, μπορούν να λαμβάνουν διαφορετικές αποφάσεις, ενώ δεν διαθέτουμε κάποια a priori γνώση για να υποθέσουμε το αντίθετο.

Τέλος, αξίζει να σημειωθεί ότι ο GMl-ASLCS δεν χρησιμοποιεί ένα σημείο διασταύρωσης για κάθε απόγονο, αλλά ένα κοινό, παράγοντας έτσι απογόνους, κατά μία έννοια, συμπληρωματικούς. Στο Σχήμα 6.2 φαίνεται η Διασταύρωση Δύο Τμημάτων με κοινό σημείο διασταύρωσης στον GMl-ASLCS.



Σχήμα 6.2: Διασταύρωση Δύο Τμημάτων, με ένα σημείο διασταύρωσης στον GMl-ASLCS

Οι Αλγ. 6.2 και 6.3 περιγράφουν στην πλήρη του λειτουργία τον Γενετικό Αλγόριθμο του GMl-ASLCS. Ο ΓΑ εφαρμόζεται σε ένα Correct Set κάθε φορά που η τιμή της τρέχουσας χρονοσφραγίδας *timestamp* είναι μεγαλύτερη από τη μέση τιμή των χρονοσφραγίδων δημιουργίας των κανόνων που συμμετέχουν στο εν λόγω Correct

Set, κατά τη σταθερά θ_{GA} (Αλγ. 6.1, γρ. 18). Η χρονοσφραγίδα $timestamp$ λειτουργεί ως ένας μετρητής των γενετικών συμβάντων, αυξανόμενη κατά μία μονάδα κάθε φορά που συμβαίνει ένα από αυτά (Αλγ. 6.2, γρ. 8).

Αλγόριθμος 6.2 Η λειτουργία του Γενετικού Αλγορίθμου στον GMl-ASLCS

1: **evolve(**$evolve(\mathbf{C}[\,l\,]), \mathbf{P}, \mathbf{S}, \overline{\mathbf{S}}$**)**
2: $parentA \leftarrow rouletteWheel(\mathbf{C}[\,l\,])$
3: $parentB \leftarrow rouletteWheel(\mathbf{C}[\,l\,])$
4: **if** $random[0,1] > \chi$ **then**
5: $offspringA \leftarrow parentA$
6: $offspringB \leftarrow parentB$
7: **else**
8: $timestamp \leftarrow timestamp + 1$
9: $chromosomeSize_c = |attributes| + 2$
10: $X \leftarrow random[0,1] \cdot chromosomeSize_c$
11: **if** $X < chromosomeSize_c$ **then**
12: **crossoverConditions**$(parentA, parentB, X)$
13: **end if**
14: **crossoverLabels**$(parentA, parentB, l)$
15: **end if**
16: **for each** $offspring$ **do**
17: $mutate(offspring)$
18: $fixChromosome(offspring)$
19: $offspring.coverage = checkForZeroCoverage(offspring)$
20: **if** $offspring.coverage \neq 0$ **then**
21: **checkForSubsumption**$(offspring, parents, \mathbf{P}, \mathbf{S}, \overline{\mathbf{S}})$
22: **end if**
23: **end for**

Αλγόριθμος 6.3 Η διαδικασία Διασταύρωσης Δύο Τμημάτων στον GMl-ASLCS. Η σημειογραφία που χρησιμοποιείται είναι αυτή της αρχικής θέσης και μήκους μεταβολής.

1: **crossoverConditions(**$parentA, parentB, \chi$**)**
2: $offspringA(0, \chi - 1) \leftarrow parentA(0, \chi - 1)$
3: $offspringA(\chi, |attributes| - \chi + 1) \leftarrow parentB(\chi, |attributes| - \chi + 1)$
4: $offspringB(0, \chi - 1) \leftarrow parentB(0, \chi - 1)$
5: $offspringB(\chi, |attributes| - \chi + 1) \leftarrow parentA(\chi, |attributes| - \chi + 1)$

1: **crossoverLabels(**$parentA, parentB, l$**)**
2: $offspringA(|attributes| + 2 \cdot l, 2) \leftarrow parentB(|attributes| + 2 \cdot l, 2)$
3: $offspringB(|attributes| + 2 \cdot l, 2) \leftarrow parentA(|attributes| + 2 \cdot l, 2)$

Η διασταύρωση εφαρμόζεται με πιθανότητα χ, ενώ σε περίπτωση απόφασης μη διασταύρωσης, οι απόγονοι προκύπτουν ως αντίγραφα των γονέων τους. Σε περίπτωση απόφασης διασταύρωσης, η συνάρτηση $crossoverConditions$ αναλαμβάνει

τη διασταύρωση των τμημάτων συνθηκών των δύο κανόνων-γονέων, $parentA$ και $parentB$, για δεδομένο σημείο διασταύρωσης X, ενώ η $crossoverLabels$ τη διασταύρωση των τμημάτων απόφασης. Οι δύο παραπάνω συναρτήσεις παρατίθενται στον Αλγ. 6.3.

Και οι δύο συναρτήσεις χρησιμοποιούν το συνολικό αριθμό των δυφίων που απαιτούνται για την αναπαράσταση των γνωρισμάτων του προβλήματος, ανάλογα με το είδος τους, δηλαδή

$$|attributes| = \sum_{i=1}^{number\ of\ attributes} bits\ of\ representation_i \tag{6.5}$$

και τον αριθμό των δυφίων που απαιτούνται για την αναπαράσταση μίας ετικέτας, δηλαδή 2, λόγω της δυαδικής αναπαράστασης ετικετών με χρήση αδιαφοριών. Συνεπώς, το μήκος του χρωμοσώματος που χρησιμοποιείται για τη διασταύρωση είναι:

$$chromosomeSize_c = |attributes| + 2 \tag{6.6}$$

Μετά τη διασταύρωση, κάθε απόγονος υπόκειται σε ομοιόμορφη μετάλλαξη των γονιδίων του (συνάρτηση $mutate$), με πιθανότητα μ. Λόγω της τυχαιότητας της διασταύρωσης και της μετάλλαξης, οι τελικοί απόγονοι ενδέχεται να αποκτήσουν γονίδια χωρίς φυσική σημασία, για αυτό και απαιτείται σε αυτό το σημείο ο έλεγχος τους με επιπλέον υπολογιστικό κόστος (συνάρτηση $fixChromosome$).

Αντιμετώπιση Απογόνων Μηδενικής Κάλυψης

Σε αντίθεση με τον GMl-ASLCS$_0$, ο GMl-ASLCS δεν διατηρεί τους κανόνες μηδενικής κάλυψης στον πληθυσμό του, αλλά τους απομακρύνει εν τη γενέσει τους (Αλγ. 6.2, γρ. 19-20). Κάθε απόγονος που προκύπτει από τον ΓΑ[2], μετά την εφαρμογή του τελεστή μετάλλαξης πάνω του, ελέγχεται ως προς την ύπαρξη νοήματός του πάνω στο πρόβλημα. Η παρουσία ενός κανόνα στον πληθυσμό έχει νόημα, εάν αυτός καλύπτει τουλάχιστον ένα δείγμα του συνόλου δεδομένων εκπαίδευσης, δηλαδή εάν κωδικοποιεί ένα μέρος του προβλήματος προς επίλυση από το ΜαΣΤ.

Αξίζει να συγκρίνουμε την παραπάνω προσέγγιση με τον τρόπο αντιμετώπισης της παραγωγής κανόνων με μηδενική κάλυψη στον GMl-ASLCS$_0$, όπου κάθε τέτοιος κανόνας διαγράφεται πιθανοτικά, με πιθανότητα αντιστρόφως ανάλογη του αριθμού των δειγμάτων του συνόλου εκπαίδευσης D, αφού πρώτα κληθεί να συμμετάσχει τουλάχιστον σε τόσα Match Sets όσα είναι και ο αριθμός των δειγμάτων του συνόλου δεδομένων. Αυτό σημαίνει πως ένας κανόνας μηδενικής κάλυψης θα διατηρηθεί στον πληθυσμό τουλάχιστον για μία ολόκληρη επανάληψη εκπαίδευσης προτού διαγραφεί. Ακόμα χειρότερα, ένας κανόνας που πληροί τις παραπάνω προϋποθέσεις, ενδέχεται να μη διαγραφεί ακόμα και τότε, διότι θα πρέπει να του έχουν

[2]Κανόνες μηδενικής κάλυψης μπορούν να προκύψουν μόνο μέσω του Γενετικού Αλγορίθμου και όχι από το τμήμα κάλυψης, καθώς αυτό δημιουργεί κανόνες γενικεύοντας πάνω στα γνωρίσματα ήδη υπαρχόντων δειγμάτων εκπαίδευσης.

παρουσιαστεί ακριβώς $k \cdot |D|$, $k \in \mathbb{Z}$ δείγματα τη χρονική στιγμή κατά την οποία τυχαίνει να εκκαθαριστεί ο πληθυσμός του ΜαΣΤ από κανόνες μηδενικής κάλυψης, λόγω σχεδιαστικού σφάλματος στον κώδικα υλοποίησης.

Ας δούμε, όμως, γιατί η διατήρηση κανόνων μηδενικής κάλυψης στον πληθυσμό ενός ΜαΣΤ αποτελεί πρόβλημα για τη λειτουργία του. Ένα ΜαΣΤ διατηρεί έναν πεπερασμένο αριθμό κανόνων, που έχει τεθεί εξαρχής. Υπό ορισμένες συνθήκες, όπως για σύνολα δεδομένων με μεγάλο αριθμό γνωρισμάτων και περιορισμένο αριθμό δειγμάτων σε σχέση με το συνολικό αριθμό των δυνατών συνδυασμών των τιμών των γνωρισμάτων[3], είναι δυνατή η εμφάνιση ενός φαινομένου στραγγαλισμού από άποψης του απαιτούμενου αριθμού κανόνων για την επίλυση του προβλήματος. Αυτό οφείλεται στο γεγονός ότι, όσο γενικοί και αν είναι οι κανόνες που αναπαράγονται, υπάρχουν περιορισμοί στην παραγωγή κανόνων μη μηδενικής κάλυψης μέσω της διασταύρωσης, λόγω της αραιότητας του συνόλου δεδομένων. Αυτοί οι περιορισμοί έχουν ως αποτέλεσμα ο πληθυσμός να αποτελείται κατά ένα μέρος από κανόνες "χρήσιμους" και κατά ένα μέρος από κανόνες μηδενικής κάλυψης, κάθε άλλο παρά αμελητέας πληθικότητας.

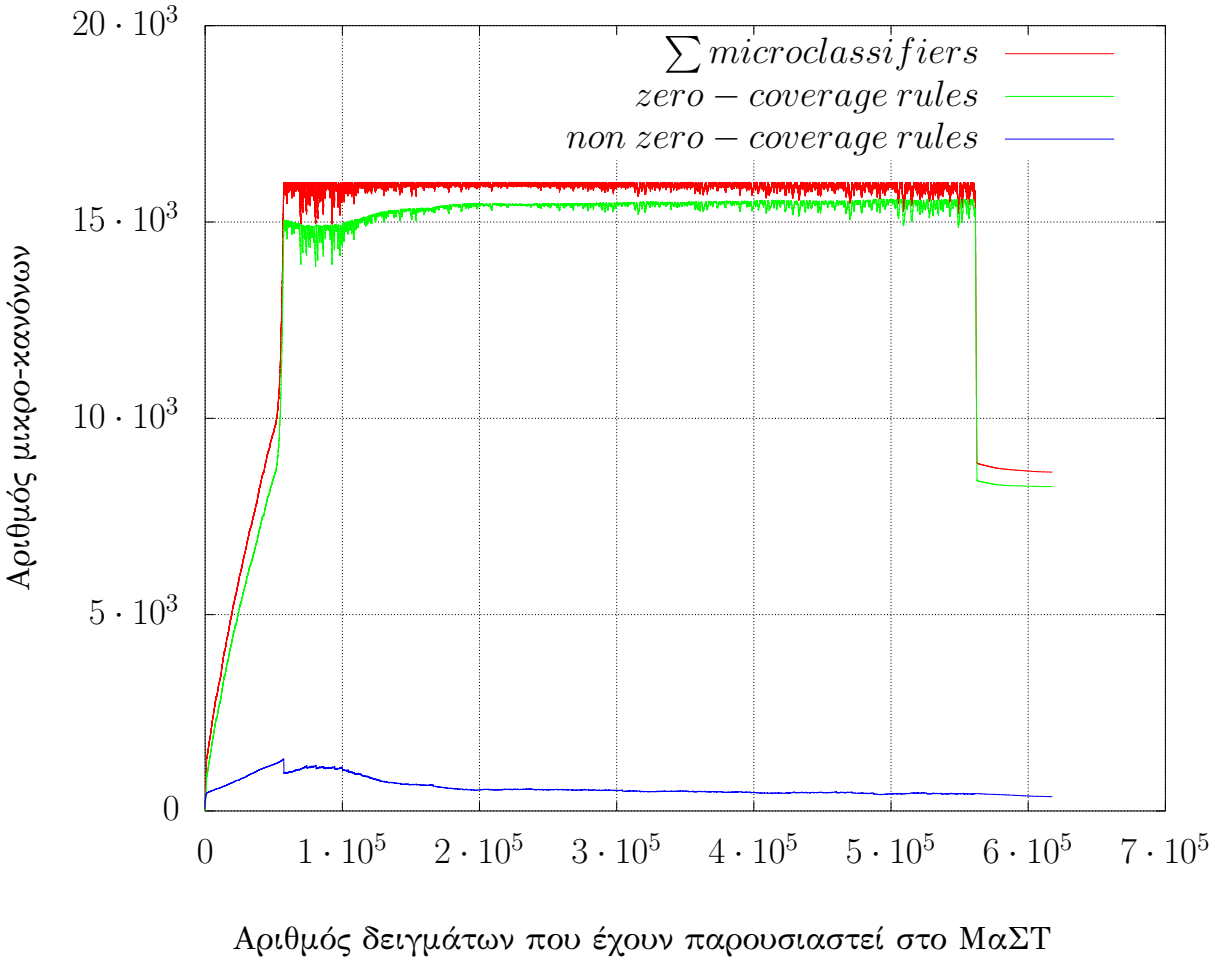
Χαρακτηριστικά, στο σύνολο *enron* που αποτελείται από 1123 δείγματα, με 1001 δυαδικά γνωρίσματα, χρησιμοποιώντας πληθυσμό 16000 μικρο-κανόνων για τον GMl-ASLCS$_0$, σε κάθε απόγονο μη μηδενικής κάλυψης που παράγεται, αντιστοιχούν περίπου τριάντα (30) κανόνες που διατηρούνται στον πληθυσμό σε κάθε δεδομένη στιγμή, μη έχοντας αντικείμενο μέσα στο ΜαΣΤ. Αυτό σημαίνει πως κάθε δεδομένη στιγμή, ο πληθυσμός αποτελείται, μεσοσταθμικά, κατά 97% από κανόνες μηδενικής κάλυψης, για το διάστημα στο οποίο ο αριθμός των μικρο-κανόνων έχει σταθεροποιηθεί. Συνολικά, στο συγκεκριμένο πείραμα, διαγράφονται περί τους $2.5 \cdot 10^6$ κανόνες μηδενικής κάλυψης, ενώ συνολικά παράγονται περί τους $42 \cdot 10^3$ κανόνες μη μηδενικής κάλυψης. Με άλλα λόγια, ο ρυθμός παραγωγής κανόνων μηδενικής κάλυψης βρίσκεται στο 98.3%[4].

Εν τέλει, παρατηρείται ένα φαινόμενο συμπίεσης του αριθμού των πραγματικών κανόνων που θα επιλύσουν το πρόβλημα και μάλιστα ένα φαινόμενο το οποίο δεν είναι ελέγξιμο, μας εμποδίζει στην αποτελεσματική επίλυση του προβλήματος ταξινόμησης και επιβαρύνει την προσέγγισή μας λόγω της αβεβαιότητας που επιφέρει για την κατάσταση του συστήματος. Επιπρόσθετα, η παραγωγή κανόνων μηδενικής κάλυψης σε υπερβολικό βαθμό επιβαρύνει το χρόνο εκπαίδευσης, καθώς το σύστημα αναλώνεται σε μεγάλο βαθμό στην εισαγωγή και διαγραφή κανόνων χωρίς χρησιμότητα.

Το γράφημα του Σχήματος 6.3 αναπαριστά, στην πορεία του χρόνου, τον αριθμό των μικρο-κανόνων του πληθυσμού $[P]$ με κόκκινο χρώμα, τον αριθμό των κανόνων μηδενικής κάλυψης με πράσινο και τον αριθμό των κανόνων μη μηδενικής κάλυψης με μπλε χρώμα, για το παραπάνω πείραμα στο σύνολο δεδομένων *enron* με πληθυσμό 16000 μικρο-κανόνων, όπως περιγράψαμε παραπάνω.

[3]Παραδείγματα αποτελούν τα πραγματικά σύνολα *enron* και *medical* που μελετώνται στην παρούσα εργασία.

[4]Είναι εύκολο κανείς να δει πως ο συνολικός αριθμός παραχθέντων κανόνων s είναι το άθροισμα των κανόνων που διαγράφηκαν λόγω μηδενικής κάλυψης z, των κανόνων που διαγράφηκαν μέσω της λειτουργίας διαγραφής κανόνων με επιλογή ρουλέτας d και του αριθμού των κανόνων του τελικού πληθυσμού p. Ο ρυθμός παραγωγής κανόνων μηδενικής κάλυψης είναι $zeroCoverageProductionRate = 1 - p/s$.



Σχήμα 6.3: Διακυμάνσεις του αριθμού των κανόνων του πληθυσμού $[P]$ και συμπίεση των χρήσιμων κανόνων του λόγω της υπέρμετρης δημιουργίας κανόνων μηδενικής κάλυψης στον GMl-ASLCS$_0$.

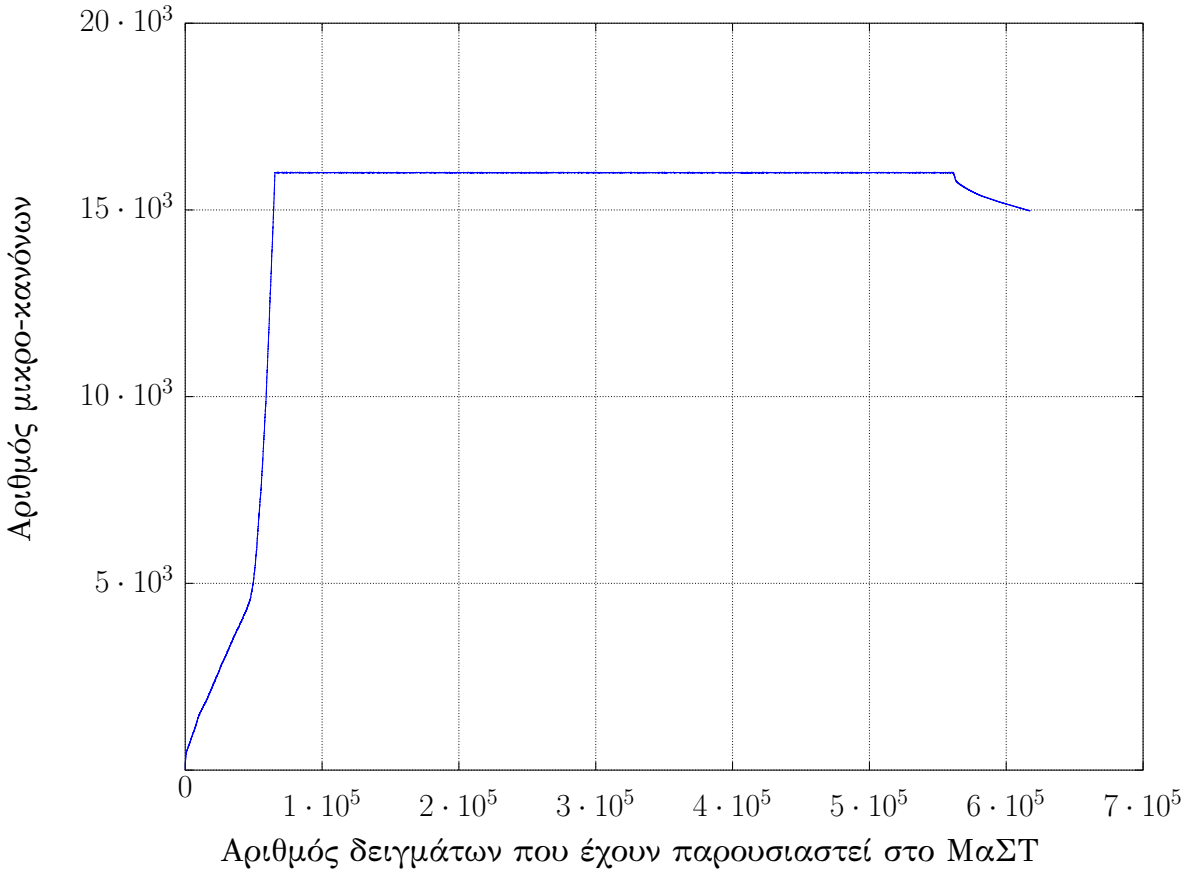
Στο γράφημα παρατηρούμε:

1. τις διακυμάνσεις του συνολικού αριθμού μικρο-κανόνων του πληθυσμού, που φτάνει να χάνει μέχρι και 1000 κανόνες μηδενικής κάλυψης σε ορισμένα χρονικά σημεία, και τη μείωση στο μισό του αριθμού των προκαθορισμένων κανόνων του πληθυσμού προς το τέλος της εκπαιδευτικής διαδικασίας,

2. το βαθμό συμπίεσης που υφίσταται ο πληθυσμός των χρήσιμων κανόνων, λόγω της υπέρ-παραγωγής κανόνων μηδενικής κάλυψης και της αδυναμίας του συστήματος να τους εντοπίσει και να τους αφαιρέσει έγκαιρα, και

3. το γεγονός ότι ο τελικός πληθυσμός κανόνων δεν αποτελείται αμιγώς από κανόνες μη μηδενικής κάλυψης, λόγω της στοχαστικής φύσης της διαγραφής κανόνων μηδενικής κάλυψης και της σχεδιαστικής συνθήκης που επιβάλει ότι σε ένα κανόνα μηδενικής κάλυψης πρέπει να έχει παρουσιαστεί ακέραιο πολλαπλάσιο του αριθμού δειγμάτων $|D|$ τη στιγμή επιλογής κανόνων προς διαγραφή, ώστε να αφαιρεθεί από τον πληθυσμό.

Μία λύση, βεβαίως, θα ήταν να αυξήσουμε το όριο του πληθυσμού σε σημείο τέτοιο ώστε οι πραγματικοί κανόνες του πληθυσμού να φτάσουν τον αριθμό των εξαρχής επιθυμητών κανόνων. Κάτι τέτοιο όμως θα αποτελούσε ημίμετρο, καθώς δεν θα ήρε την απροσδιοριστία για το πραγματικό μέγεθος του πληθυσμού που απαιτεί το πρόβλημα και θα αύξανε κατά πολύ το χρόνο εκπαίδευσης.



Σχήμα 6.4: Ομαλοποίηση του αριθμού των κανόνων του πληθυσμού $[P]$, μέσω της αποφυγής πρόσθεσης κανόνων μηδενικής κάλυψης στον πληθυσμό, στον GMl-ASLCS.

Αντί αυτού, η προσέγγισή μας στα πλαίσια του GMl-ASLCS αντιμετωπίζει το πρόβλημα στη ρίζα του. Κάθε απόγονος, προτού εισαχθεί στον πληθυσμό, ελέγχεται, όπως προαναφέραμε, ως προς την κάλυψη δειγμάτων του συνόλου δεδομένων. Στην περίπτωση μηδενικής κάλυψης ο κανόνας δεν εισάγεται στον πληθυσμό και η εκτέλεση του αλγορίθμου συνεχίζει κανονικά. Με αυτόν τον τρόπο επιτυγχάνεται:

- η άρση της απροσδιοριστίας για την κατάσταση του συστήματος, ώστε να καταστεί περισσότερο ελέγξιμο, βοηθώντας στην κατανόηση και μελέτη των λειτουργιών του, και καθιστώντας ευκολότερη τη ρύθμιση παραμέτρων που θα οδηγήσουν σε ορθές λύσεις μέσω του ακριβέστερου προσδιορισμού των παραμέτρων του ΜαΣΤ,

- η μείωση του συνολικού αριθμού των κανόνων που απαιτούνται για την επίλυση ενός συγκεκριμένου προβλήματος για την εξαγωγή ίδιων αποτελεσμάτων, σε σχέση με την προηγούμενη προσέγγιση του GMl-ASLCS$_0$ (σε στοχαστικά πλαίσια), και

- η μείωση του χρόνου εκπαίδευσης για τον ίδιο αριθμό συνολικών μικροκανόνων, σε σχέση με τον GMl-ASLCS$_0$, ανάλογα με το βαθμό φορτικότητας του φαινομένου μηδενικής κάλυψης (όσο μεγαλύτερος ο ρυθμός παραγωγής κανόνων μηδενικής κάλυψης, τόσο μεγαλύτερη και η μείωση του χρόνου χρόνου εκπαίδευσης) και αντιστρόφως ανάλογα με το μέγεθος του συνόλου εκπαίδευσης.

  Με την απαγόρευση εισαγωγής κανόνων μηδενικής κάλυψης στον πληθυσμό, πρέπει να γίνει έλεγχος κάλυψης $|D|$ δειγμάτων για τους κανόνες μηδενικής κάλυψης και ενός απροσδιόριστου αριθμού δειγμάτων $instancesChecked \leq |D|$ για αυτούς που είναι μη μηδενικής κάλυψης, μετά τη δημιουργία τους από τον ΓΑ. Ενδεικτικά, στο πρόβλημα *enron*, για άνω όριο συνολικών μικροκανόνων ίσο με 16000, ο χρόνος εκπαίδευσης του GMl-ASLCS είναι λίγα λεπτά χαμηλότερος από αυτόν του GMl-ASLCS$_0$, που είναι είναι τρεις ώρες[5].

Στο γράφημα του Σχήματος 6.4 φαίνεται η ομαλοποίηση του αριθμού των μικροκανόνων του πληθυσμού, λόγω της αποφυγής πρόσθεσης κανόνων μηδενικής κάλυψης στον πληθυσμό, και η ταύτισή του με τον αριθμό συνολικών μικρο-κανόνων που έχει τεθεί εκ των προτέρων. Η πτώση του παραπάνω αριθμού στο τέλος της εκπαίδευσης εξηγείται στην Παρ. 6.3.2.

Αφομοίωση-Υπαγωγή Απογόνων

Οι κανόνες που περνούν τον έλεγχο μηδενικής κάλυψης (γραμμές 18, 19 της συνάρτησης *evolve* στον Αλγ. 6.2) είναι πλέον έτοιμοι προς εισαγωγή στον πληθυσμό $[P]$. Η εισαγωγή στον πληθυσμό, όμως, δε σημαίνει απαραίτητα την αυτοτελή πρόσθεση του κανόνα σε αυτόν. Κάθε απόγονος, όπως περιγράφεται από τη συνάρτηση $checkForSubsumption()$ του Αλγ. 6.4, ελέγχεται πρώτα για *υπαγωγή* από τους γονείς του και ύστερα, σε περίπτωση αποτυχίας, από όλο τον πληθυσμό. Εάν κανένας κανόνας του πληθυσμού δεν πληροί τις προϋποθέσεις για να αφομοιώσει τον απόγονο που παρήχθη, αυτός εισάγεται ως αυτοτελής κανόνας. Τα σύνολα $\mathbf{S}$ και $\overline{\mathbf{S}}$ στον Αλγ. 6.2 περιέχουν τους απογόνους που θα αφομοιωθούν εν τέλει από κανόνες του πληθυσμού και τους κανόνες που θα εισαχθούν ως αυτοτελείς κανόνες, αντίστοιχα.

Αναγκαίες συνθήκες ώστε ένας κανόνας r_s να αφομοιώσει έναν κανόνα r είναι:

- Το τμήμα συνθήκης του r_s να είναι το ίδιο γενικό ή γενικότερο από του r

- Το τμήμα απόφασης του r_s να είναι το ίδιο ειδικό ή ειδικότερο από του r

- Η εμπειρία του r_s να ξεπερνάει ένα κατώφλι θ_{sub}

[5] Ο χρόνος εκπαίδευσης αυξάνει εκθετικά με την αύξηση του μεγέθους του πληθυσμού.

Αλγόριθμος 6.4 Η λειτουργία αφομοίωσης-υπαγωγής στον GMl-ASLCS

```
1: checkForSubsumption(offspring, parents, P, S, S̄)
2: if (parentA.isEligibleToSubsume(offspring) AND
   parentB.isEligibleToSubsume(offspring)) then
3:    subsumer ← fittestAndMostExperienced(parents)
4:    S.insert(offspring)
5:    return
6: else
7:    if parentA.isEligibleToSubsume(offspring) then
8:       subsumer ← parentA
9:       S.insert(offspring)
10:      return
11:   end if
12:   if parentB.isEligibleToSubsume(offspring) then
13:      subsumer ← parentB
14:      S.insert(offspring)
15:      return
16:   end if
17:   for each rule ∈ P do
18:      candidateSubsumers ← rule.isEligibleToSubsume(offspring)
19:   end for
20:   subsumer ← fittestAndMostExperienced(candidateSubsumers)
21:   if subsumer found then
22:      S.insert(offspring)
23:   else
24:      S̄.insert(offspring)
25:   end if
26: end if
```

Όσον αφορά στην αφομοίωση των απογόνων από τους γονείς τους, ένας απόγονος θα αφομοιωθεί από τον καταλληλότερο γονέα που πληροί τις παραπάνω συνθήκες και, σε περίπτωση ίσης καταλληλότητας, από αυτόν με τη μεγαλύτερη εμπειρία. Αν δεν ικανοποιούνται οι παραπάνω τρεις συνθήκες για τους δύο γονείς, η απόφαση υπαγωγής μεταφέρεται στον πληθυσμό. Εκεί, προτεραιότητα έχουν οι κανόνες με γενικότερο τμήμα συνθήκης και τη μεγαλύτερη καταλληλότητα, καταλήγοντας σε αυτούς με το ίδιο μέγεθος κάλυψης, με την καταλληλότητα των κανόνων να παίζει τον καθοριστικό ρόλο, όπως και πριν. Στην περίπτωση που υπάρχουν παραπάνω του ενός κανόνες, αυτός που θα αφομοιώσει τον απόγονο θα είναι αυτός που δημιουργήθηκε τελευταίος χρονικά. Οι κανόνες που έχουν προκύψει μέσω του Τμήματος Κάλυψης δεν ελέγχονται για υπαγωγή και εισέρχονται αυτοτελώς στον πληθυσμό του GMl-ASLCS.

Σε κάθε περίπτωση, αφού βρεθεί ο κανόνας r_s που θα αφομοιώσει τον απόγονο r, ο r δεν εισάγεται στον πληθυσμό και η πληθικότητα του r_s αυξάνει κατά ένα. Σε αντίθετη περίπτωση, ο r εισάγεται ως αυτόνομος κανόνας στον πληθυσμό $[P]$, με αρχικές συνθήκες $(tp, msa, cs, fitness) = (0, 0, 1, 1)$.

Η λειτουργία της αφομοίωσης αποτρέπει τη συσσώρευση υπέρ-ειδικών κανόνων, λειτουργώντας, έτσι, ενισχυτικά προς τη διατήρηση κανόνων σε ένα επίπεδο γενικότητας. Ακριβώς για τον ίδιο λόγο, όμως, μπορεί να δυσχεραίνει την εξερεύνηση του χώρου αναζήτησης, αποτρέποντας υπέρ-ειδικούς κανόνες, που κωδικοποιούν ένα μέρος της λύσης και λειτουργούν βοηθητικά προς αυτή, από το να συμμετάσχουν στον πληθυσμό.

Διαδικασία Εισαγωγής Απογόνων στον Πληθυσμό

Για ένα πρόβλημα πολυκατηγορικής ταξινόμησης με $|L|$ ετικέτες, η εισαγωγή ενός δείγματος για εκπαίδευση στο ΜαΣΤ έχει ως αποτέλεσμα το σχηματισμό $|L|$ Correct Sets, από το καθένα από τα οποία είτε θα προκύψουν ακριβώς δύο απόγονοι, ή κανένας. Συνεπώς, συνολικά, η εισαγωγή ενός δείγματος εκπαίδευσης έχει ως αποτέλεσμα την παραγωγή k απογόνων, όπου $k = 2 \cdot m$ με $m \in \mathbb{N}$ και $m \in [0, |L|]$. Ενώ στον GMl-ASLCS$_0$ η διαδικασία εισαγωγής ενσωματώνει κάθε κανόνα στον πληθυσμό απευθείας μετά την παραγωγή του, ο GMl-ASLCS χρησιμοποιεί ένα σχήμα μαζικής εισαγωγής των απογόνων που παράγονται συνολικά από την είσοδο ενός δείγματος στο ΜαΣΤ.

Αφενός, αυτό γίνεται για λόγους οικονομίας χρόνου και υπολογισμών. Στην πρώτη περίπτωση, για τη συνολική εισαγωγή k απογόνων στον πληθυσμό $[P]$, εάν έχει ξεπεραστεί ο μέγιστος αριθμός κανόνων που έχει θέσει ο χρήστης, σε κάθε γενετικό συμβάν πρέπει να υπολογιστούν

$$k \cdot (|P| + 1) \tag{6.7}$$

πιθανότητες διαγραφής, ενώ στη δεύτερη

$$|P| + k \tag{6.8}$$

Αφετέρου, τα διενεργηθέντα πειράματα σε τεχνητά και πραγματικά σύνολα δεδομένων δείχνουν πως η δεύτερη μέθοδος εισαγωγής κανόνων στον πληθυσμό επιδεικνύει συνολικά καλύτερες επιδόσεις από την πρώτη.

Συνολικά, όπως φαίνεται και στον Αλγ. 6.1, η συνάρτηση $controlPopulation$ καλείται μία φορά ανά δείγμα σε περίπτωση που ο αριθμός των μικρο-κανόνων έχει ξεπεράσει το όριο $maximumPopulationSize$ που έχουμε θέσει (Αλγ. 6.1, γρ. 27, 28), αντί για k φορές (Αλγ. 5.4, γρ. 20, 21). Στους αλγορίθμους 6.1 και 6.4, τα σύνολα κανόνων $\mathbf{S}$ και $\bar{\mathbf{S}}$ υποδηλώνουν το σύνολο στο οποίο τοποθετούνται οι κανόνες που προορίζονται προς αφομοίωση και αυτό στο οποίο τοποθετούνται κανόνες που θα εισαχθούν στον πληθυσμό ως αυτόνομοι, αντίστοιχα, με $|\mathbf{S}| + |\bar{\mathbf{S}}| = k$.

Λειτουργία Διαγραφής

Όπως προαναφέραμε, η λειτουργία διαγραφής ενεργοποιείται κάθε φορά που είναι αληθής η συνθήκη

$$\sum microClassifiers > maximumPopulationSize \tag{6.9}$$

όπου $maximumPopulationSize$ το άνω όριο για το συνολικό αριθμό μικρο-κανόνων που διατηρεί το ΜαΣΤ στον πληθυσμό $[P]$. Μετά από κάθε εισαγωγή απογόνων στον πληθυσμό, ελέγχεται η ισχύς της (6.9) και, όταν βρεθεί αληθής, επιλέγονται για διαγραφή τόσοι κανόνες όση η διαφορά

$$\sum microClassifiers - maximumPopulationSize \tag{6.10}$$

χρησιμοποιώντας επιλογή ρουλέτας.

Σε κάθε κανόνα i του πληθυσμού $[P]$ ανατίθεται μία πιθανότητα επιλογής $P(i)$:

$$P(i) = \frac{num(i) \cdot d(i)}{\sum_{j=1}^{n} \big(num(j) \cdot d(j)\big)} \tag{6.11}$$

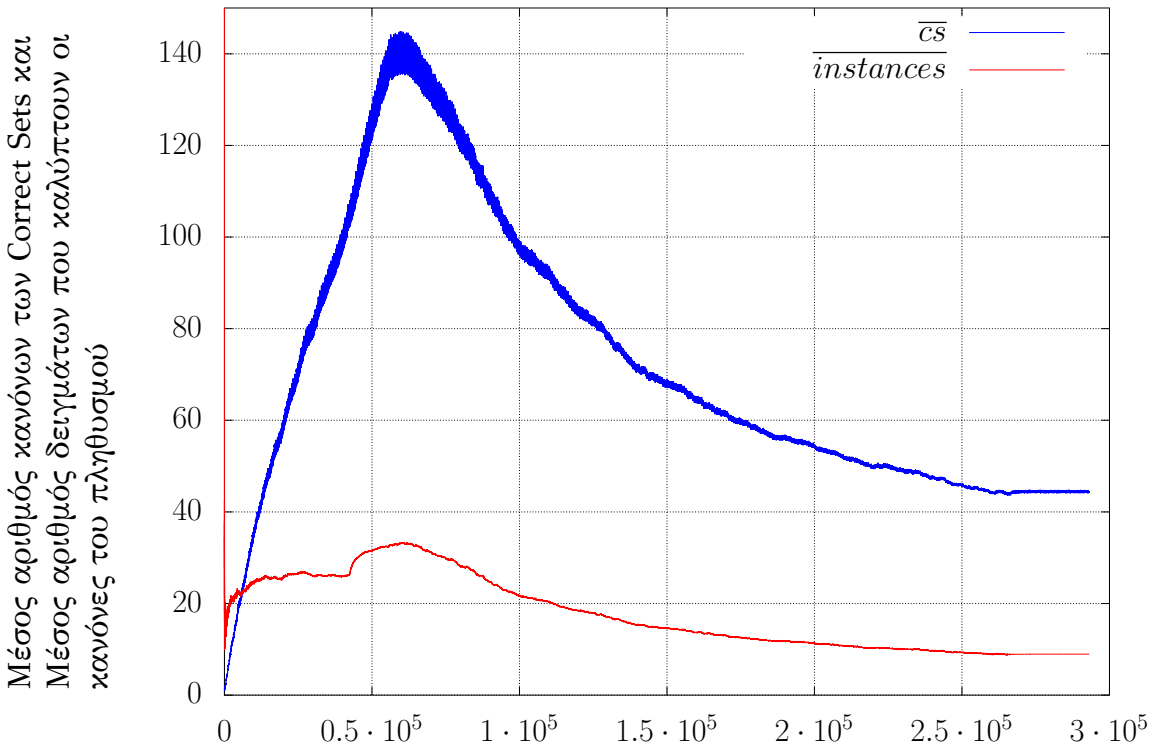
με $d(i)$:

$$d(i) = \begin{cases} e^{\dfrac{1}{fitness(i)}}, & experience(i) < \theta_{del} \\ \dfrac{e^{cs(i)} - 1}{fitness(i)}, & \text{αλλού} \end{cases} \tag{6.12}$$

Η Εξ. 6.12 δείχνει ότι αποφεύγουμε συνολικά τη χρήση της μέσης καταλληλότητας του πληθυσμού F_P, καθώς και τη χρήση του cs για τους κανόνες των οποίων η εμπειρία είναι μικρή, δηλαδή έχουν συμμετάσχει σε λιγότερα από θ_{del} Match Sets. Η χρήση της μέσης καταλληλότητας F_P βρέθηκε πειραματικά ότι δεν έχει κάποια βοηθητική

δράση στις επιδόσεις του GMl-ASLCS (συγκεκριμένα για τις μετρικές της Ακρίβειας και της Μέσης Κάλυψης δειγμάτων από τους κανόνες του), στο υφιστάμενο πλαίσιο διαγραφής, ενώ ο υπολογισμός της εισάγει πρόσθετη χρονική πολυπλοκότητα. Όσον αφορά στην εκτίμηση του μεγέθους των Correct Sets cs στα οποία συμμετέχει ένας κανόνας, αυτή δεν χρησιμοποιείται για τους κανόνες με $experience < \theta_{del}$, λόγω της υπερεκτίμησης του πραγματικού μεγέθους των Correct Sets στα οποία συμμετέχουν, η οποία μπορεί να οδηγήσει σε μη αξιόπιστες διαγραφές.

Το γράφημα του Σχήματος 6.5 παριστά με μπλε χρώμα το μέσο όρο των cs των κανόνων του πληθυσμού και με κόκκινο το μέσο αριθμό καλυπτόμενων δειγμάτων από τον πληθυσμό $[P]$ των κανόνων, για το πρόβλημα $music$. Ο κατακόρυφος άξονας συμβολίζει το μέσο αριθμό κανόνων των Correct Sets για την καμπύλη με μπλε χρώμα και το μέσο αριθμό δειγμάτων που καλύπτουν οι κανόνες του πληθυσμού για την καμπύλη με κόκκινο χρώμα. Παρατηρούμε δύο διακριτά διαστήματα: το πρώτο, όπου το $\overline{cs}$ αυξάνει γραμμικά, και το δεύτερο, στο οποίο φθίνει εκθετικά.



Σχήμα 6.5: Καμπύλες εξέλιξης του μέσου cs και του μέσου αριθμού καλυπτόμενων δειγμάτων από τους κανόνες του πληθυσμού στο πρόβλημα music.

Στο πρώτο διάστημα, ο αριθμός των κανόνων αυξάνει, χωρίς να έχει φτάσει το άνω όριό του και, άρα, χωρίς να έχουν γίνει διαγραφές κανόνων. Όσο αυξάνει ο αριθμός των κανόνων, αυξάνει και ο αριθμός των κανόνων που συμμετέχουν σε Correct Sets, λόγω της αναπαραγωγής κανόνων με βάση τους κανόνες που συμμετέχουν σε αυτά. Έτσι, για κάθε κανόνα, αυξάνει και η εκτίμηση για τον αριθμό

κανόνων cs στα Correct Sets που συμμετέχει, αυξάνοντας συνεπώς και το μέσο cs ($\overline{cs}$) των κανόνων του πληθυσμού. Η εκτίμηση cs του μεγέθους των $[C]$ βρίσκεται πάντα χαμηλότερα από το πραγματικό cs, όσο αυξάνει το $minCs$, λόγω της σχέσης

$$rule.cs \leftarrow rule.cs + \beta \cdot (minCs - rule.cs) \tag{6.13}$$

που χρησιμοποιείται. Μάλιστα, η υποεκτίμηση αυξάνει σε μέγεθος όσο μεγαλύτερες είναι οι διαφορές $minCs - rule.cs$. Στους πίνακες 6.2 και 6.3 παρατίθενται τεχνητά μεγέθη για το ελάχιστο μέγεθος των Correct Sets στα οποία συμμετέχει ένας κανόνας, η πραγματική τιμή του μέσου μεγέθους cs, υπολογισμένη ως ο μέσος όρος των $minCs$ και η εκτίμηση για το μέσο cs των κανόνων. Ο αναγνώστης θα παρατηρήσει πως όσο αυξάνει ο αριθμός συμμετοχών ενός κανόνα σε Correct Set, τόσο μικρότερο είναι το σφάλμα για το πραγματικό μέσο μέγεθος των αριθμών των κανόνων των Correct Sets στα οποία αυτός συμμετέχει.

Πίνακας 6.2: Υποεκτίμηση του πραγματικού cs, για $\beta = 0.2$, πριν την έναρξη των διαγραφών με επιλογή ρουλέτας.

$min(cs)$	Πραγματικό cs	cs
20	20	20
40	30	24
60	40	31.2
80	50	40.96
100	60	52.768

Πίνακας 6.3: Υπερεκτίμηση του πραγματικού cs, για $\beta = 0.2$, μετά την έναρξη των διαγραφών με επιλογή ρουλέτας.

$min(cs)$	Πραγματικό cs	cs
100	100	100
80	90	96
60	80	88.8
40	70	79.04
20	60	67.232

Το δεύτερο διάστημα, όπου το μέσο cs φθίνει εκθετικά, είναι αυτό στο οποίο διαγράφονται κανόνες. Οι διαγραφές ξεκινούν στο ολικό μέγιστο της καμπύλης του Σχήματος 6.5. Στον Πίνακα 6.3 βλέπουμε, αντίστοιχα με την υποεκτίμηση που συμβαίνει στο πρώτο διάστημα, την υπερεκτίμηση του μέσου μεγέθους cs των κανόνων του πληθυσμού, για το διάστημα στο οποίο διαγράφονται κανόνες σύμφωνα με την Εξ. 6.12, στο οποίο παρατηρείται η πτώση του $\overline{cs}$ των κανόνων. Εάν διαγράφαμε κανόνες που είναι σχετικά άπειροι, με $experience < \theta_{del}$, που δημιουργήθηκαν μετά την έναρξη των διαγραφών, έχοντας ως κριτήριο και την εκτίμηση cs, τότε, θα στερούσαμε από τον πληθυσμό καινούριους κανόνες οι οποίοι ίσως να ήταν χρήσιμοι αλλά η εκτίμηση του πραγματικού cs τους είναι σαφέστατα μεγαλύτερη από την πραγματική τιμή του. Παράλληλα, δεν θα είχε σταθεροποιηθεί η εκτίμηση ακόμα, τόσο, ώστε να τους ανατεθεί μία ακριβέστερη πιθανότητα διαγραφής. Αν εξετάσουμε το ζήτημα υπό το πρίσμα του δευτέρου μέρους της Εξ. 6.12, είναι δυνατόν, το μέγεθος cs κανόνων σχετικά νέων στον πληθυσμό, καθώς φθίνει γενικότερα το μέγεθος cs των κανόνων, να είχε τιμή μεγαλύτερη από αυτό κανόνων έμπειρων, που δημιουργήθηκαν πριν την έναρξη των διαγραφών, των οποίων το μέγεθος cs είναι χαμηλότερο λόγω της υποεκτίμησης του πραγματικού cs στο πρώτο διάστημα.

Για αυτό το λόγο, επιλέγουμε να αναθέσουμε σε κανόνες των οποίων η εμπειρία είναι χαμηλότερη από το κατώφλι θ_{del}, μία πιθανότητα διαγραφής που δε χρησιμοποιεί την εκτίμηση cs, αλλά είναι αντιστρόφως ανάλογη προς την καταλληλότητά τους και σε εκθετική μορφή, ώστε νέοι κανόνες, χαμηλής καταλληλότητας, να έχουν συγκρίσιμη πιθανότητα διαγραφής ως προς αυτούς που είναι εμπειρότεροι. Θέτοντας μάλιστα το θ_{del} σε μία τιμή λογική ως προς το μέσο ποσοστό κάλυψης των κανόνων, φροντίζουμε ώστε οι κανόνες που έχουν παραχθεί πρόσφατα σε σχέση με τη στιγμή που τους ανατίθενται πιθανότητες διαγραφής, να έχουν αξιολογηθεί ως προς τα περισσότερα, αν όχι όλα, τα δείγματα που καλύπτουν, έτσι ώστε να έχει σταθεροποιηθεί η καταλληλότητά τους και να μην υπάρχουν κανόνες που αδικούνται από το σχήμα διαγραφής.

Όσον αφορά στο δεύτερο μέρος της Εξ. 6.12, αυτό βασίζεται, όπως και στις προσεγγίσεις της βιβλιογραφίας, στην εκτίμηση cs. Η δική μας προσέγγιση χρησιμοποιεί, ως πρόσθετο βάρος στην πιθανότητα διαγραφής, την καταλληλότητα του κανόνα, αλλά με μικρότερο βαθμό επιρροής σε σχέση με αυτόν του cs. Η, σχεδόν αποκλειστική, χρήση του μεγέθους cs οφείλεται σε δύο λόγους:

1. Διαγράφοντας έχοντας ως κριτήριο το cs, τείνουν να εξισορροπηθούν σε μέγεθος τα διαφορετικά Correct Sets, κατανέμοντας στο καθένα έναν ίσο αριθμό κανόνων με όλα τα υπόλοιπα [Wil95]. Πράγματι, ένας κανόνας έχει μεγαλύτερη πιθανότητα διαγραφής όταν συμμετέχει σε $[C]$ με μεγάλο μέγεθος. Με αυτό τον τρόπο, επαναληπτικά, μειώνονται οι κανόνες που συμμετέχουν σε υπερπληθή $[C]$, εξισορροπώντας τα μεγέθη των διαφόρων Correct Sets.

2. Από τα αποτελέσματα και τις παρατηρήσεις μας, φαίνεται ότι αποτελεί ένα αξιόπιστο κριτήριο διαγραφής.

   Αυτή η εργασία, επιλέγει να χρησιμοποιήσει ως ένα δεύτερο κριτήριο και το μέγεθος της καταλληλότητας, ώστε να διαχωρίζονται καλύτερα κανόνες που συμμετέχουν σε Correct Sets με παρόμοιο μέγεθος, βοηθώντας στην προαγωγή των καταλληλότερων κανόνων. Το ζήτημα της υπερεκτίμησης είναι παρόν και εδώ, αλλά σε μικρότερο βαθμό από ότι για κανόνες με εμπειρία μικρότερη από θ_{del} και, μάλιστα, με μειούμενη επιρροή όσο μεγαλύτερη είναι η εμπειρία ενός κανόνα. Όσο μεγαλύτερη είναι η εμπειρία του, τόσες περισσότερες φορές έχει εκτιμηθεί η πραγματική τιμή του cs, άρα τόσο "μικρότερη" είναι η υπερεκτίμηση για αυτό, όπως φαίνεται και στους Πίνακες 6.2 και 6.3.

Ο αναγνώστης θα παρατηρήσει πως και στα δύο μέρη της Εξ. 6.12 χρησιμοποιούμε ως εκθέτες με βάση e τα κριτήρια στα οποία στηρίζονται οι διαγραφές. Η εκθετική μορφή χρησιμοποιείται ώστε να διαχωρίσει ακριβέστερα και με μεγαλύτερη αποφασιστικότητα κανόνες με μικρές διαφορές (ως προς το μέγεθος cs ή ανάμεσα στην καταλληλότητα και το cs), ανάλογα και με το μέγεθος της υπερεκτίμησης για τον κάθε κανόνα με $experience \geq \theta_{del}$.

Συνιστωσα Ενισχυσης

Η Συνιστώσα Ενίσχυσης τροποποιείται σημαντικά στον GMl-ASLCS, με δύο σημεία διαφοροποίησης, τα οποία μπορούμε να εντοπίσουμε στον Αλγ. 6.1. Η πρώτη διαφοροποίηση βρίσκεται στον έλεγχο και τη λειτουργία της συνάρτησης $controlMatchSet$ (γραμμές 3 και 4, αντίστοιχα, του αλγορίθμου 6.1). Η λειτουργία και οι λόγοι για την επινόηση της διαδικασίας που υλοποιείται στη μέθοδο $controlMatchSet$ περιγράφονται στην Παρ. 6.3.2. Η δεύτερη τροποποίηση παρουσιάζεται στην επόμενη ενότητα και εντοπίζεται στη συνάρτηση $updateFitness$ (γρ. 10 του Αλγ. 6.1). Αφορά στον τρόπο χειρισμού των αδιαφοριών για ετικέτες στο καίριο σημείο της λογικής μεταβολής της καταλληλότητας ενός κανόνα που συμμετέχει σε ένα Match Set, αλλά αδιαφορεί για μία ετικέτα l.

Ενημέρωση Καταλληλότητας

Η ακριβής μεθοδολογία ενημέρωσης των μεταβλητών που σχετίζονται με την καταλληλότητα περιγράφεται από τον Αλγ. 6.5.

Αλγόριθμος 6.5 Ενημέρωση της καταλληλότητας στον GMl-ASLCS.

1: **updateFitness**($rule$)
2: $rule.exp \leftarrow rule.exp + 1$
3: **for each** $l \in L$ **do**
4: $\quad rule.tp \leftarrow rule.tp + correctness(rule, l)$
5: $\quad rule.msa \leftarrow rule.msa + msaValue(rule, l)$
6: **end for**
7: $rule.fitness \leftarrow \left(\frac{rule.tp}{rule.msa}\right)^{\nu}$

Για κάθε κανόνα στο Match Set, εξετάζεται η ικανότητά του για κατηγοριοποίηση του $Instance$ σε κάθε ετικέτα l. Επίσης για κάθε ετικέτα, οι ποσότητες tp και msa αυξάνονται κατά ποσό ανάλογο της ικανότητας κατηγοριοποίησης ($correctness$) του κανόνα σε αυτή. Τα μεγέθη tp και msa αναπαριστούν και εδώ τον αριθμό ορθών κατηγοριοποιήσεων ενός κανόνα και τον ολικό αριθμό κατηγοριοποιήσεών του, αντίστοιχα. Λόγω της χρήσης αναπαράστασης ετικετών με αδιαφορίες, όμως, είναι αναγκαίο να εξετάσουμε το μέγεθος της ποσότητας "ανταμοιβής" που θα πρέπει να λάβει ένας κανόνας, σε περίπτωση αδιαφορίας για μία ετικέτα. Φορμαλιστικά:

$$correctness(rule, l) = \begin{cases} 1, & \text{εάν ο } rule \text{ προβλέπει ορθά την } l \\ 0, & \text{εάν ο } rule \text{ προβλέπει λανθασμένα την } l \\ \omega, & \text{εάν ο } rule \text{ αδιαφορεί για την } l \end{cases} \tag{6.14}$$

Προφανώς, σε περίπτωση ορθής κατηγοριοποίησης το μέγεθος tp θα αυξηθεί κατά ένα, και σε περίπτωση λανθασμένης θα μείνει αμετάβλητο. Στην περίπτωση της αδιαφορίας, όμως, δε θα ήταν σχεδιαστικά και εξελικτικά ορθό να χρησιμοποι-

ήσουμε κάποια από αυτές τις δύο ακραίες τιμές: ο κανόνας δεν κατηγοριοποιεί ούτε ορθά για να λάβει το πλήρες ποσό της ανταμοιβής, ούτε λανθασμένα ώστε να μην λάβει κάποια ανταμοιβή.

Επιπρόσθετα, η επιζητούμενη ποσότητα θα πρέπει να ιδωθεί και υπό το πρίσμα της έκπτωσης στην καταλληλότητα που προσδίδει η παράμετρος ν. Ακόμα και αν θέταμε την τιμή του ω στο μέσο όρο των δύο ακραίων τιμών ορθότητας, υποθέτοντας έστω και μία αδιαφορία, η ύψωση στη νιοστή δύναμη[6] της ακρίβειας $acc = tp/msa$ θα καθιστούσε τον κανόνα λιγότερο επιλέξιμο πιθανοτικά για αναπαραγωγή και περισσότερο για διαγραφή από όσο του "αξίζει", κάνοντας την προσέγγισή μας πολύ αυστηρή.

Παράλληλα, ο καθορισμός μίας ορθολογικής τιμής για το ω θα πρέπει να γίνει σε συνδυασμό με τον καθορισμό της τιμής ϕ:

$$msaValue(rule, l) = \begin{cases} \phi, & \text{εάν ο } rule \text{ αδιαφορεί για την } l \\ 1, & \text{αλλού} \end{cases} \tag{6.15}$$

Η προσέγγιση που ακολουθείται εν τέλει, θέτει το $\omega = 0.9$ και το $\phi = 1$. Έτσι, για έναν κανόνα, για κάθε ετικέτα που αδιαφορεί, γίνεται έκπτωση 0.1 από την τιμή της ορθής κατηγοριοποίησης, μεταβάλλοντας την καταλληλότητά του στην:

$$fitness = \left(\frac{tp + 0.9}{msa + 1}\right)^{\nu} \tag{6.16}$$

Διεύρυνση της Μέσης Κάλυψης

Όπως έχει προαναφερθεί, το (κατά το δυνατόν) υψηλό ποσοστό κάλυψης δειγμάτων από τους κανόνες είναι μία από τις προϋποθέσεις για την αποτελεσματικότητα και ακρίβεια πρόβλεψης κάθε Μανθάνοντος Συστήματος Ταξινομητών. Ο GMl-ASLCS$_0$, παρ' όλα αυτά, εξέλισσε κανόνες των οποίων ο μέσος αριθμός δειγμάτων που κάλυπταν βρίσκονταν σε σχετικά χαμηλά επίπεδα, όπως σημειώνεται και στον Πίνακα 6.4. Η σχετικότητα έγκειται αφενός στη φύση και τα χαρακτηριστικά ενός δεδομένου προβλήματος, αφετέρου στις παραμέτρους και τις επιμέρους λειτουργίες του ίδιου του ΜαΣΤ.

Ιδανικά, στόχος μας είναι το τελικό μοντέλο να αποτελεί μία συμπαγή αναπαράσταση γνώσης, ένα πλούσιο σύνολο κανόνων από πλευράς πληροφορίας και ποικιλότητας κανόνων, με ταυτόχρονη υψηλή προβλεπτική ικανότητα. Αυτή η εργασία κινείται με άξονα τους παραπάνω στόχους, προσπαθώντας να συγκεράσει τις απαιτήσεις για για ακρίβεια και γενίκευση, διατηρώντας τουλάχιστον στα ίδια επίπεδα την ακρίβεια που παρουσιάζει ο GMl-ASLCS$_0$ για τα σύνολα δεδομένων στα οποία δοκιμάζεται, αυξάνοντας τον μέσο αριθμό δειγμάτων που καλύπτει, κατά μέσο όρο, κάθε κανόνας του ΜαΣΤ.

Η προσέγγιση που ακολουθήσαμε βασίζεται στο παρακάτω σκεπτικό. Ας υποθέσουμε την ύπαρξη ενός δεδομένου συνόλου κανόνων $[S]$, όπως το Match Set $[M]$, ή το Correct Set $[C]$, με επαρκή πληθικότητα. Το $[S]$ αποτελείται, εν γένει, από κανόνες

[6] Η τιμή που χρησιμοποιείται σχεδόν αποκλειστικά στη βιβλιογραφία είναι $\nu = 10$.

Πίνακας 6.4: Μέση Κάλυψη δειγμάτων από τους κανόνες, όπως προκύπτει πειραματικά για τον GMl-ASLCS$_0$, για τα έξι σύνολα πολυκατηγορικών δεδομένων που χρησιμοποιήθηκαν.

dataset	coverage %	*instances*
music	0.2827	1.51
yeast	0.1361	2.96
genbase	2.1677	12.90
scene	0.1628	1.97
medical	4.7601	15.76
enron	0.3545	3.98

που βρίσκονται σε διαφορετικά επίπεδα κάλυψης[7], ανάλογα και με το βαθμό στον οποίο έχει γενικεύσει το ΜαΣΤ. Επιπλέον, σε κάθε επίπεδο κάλυψης, βρίσκονται κανόνες διαφορετικής καταλληλότητας. Εάν σε κάθε επίπεδο κάλυψης υπάρχουν δύο ή περισσότεροι κανόνες, τότε, ο κανόνας με τη μικρότερη προβλεπτική ικανότητα ανάμεσά τους, ενδεχομένως να μην έχει νόημα ύπαρξης στον πληθυσμό $[P]$, καθώς υπάρχουν περισσότεροι και καταλληλότεροι κανόνες που καλύπτουν το ίδιο δείγμα για το οποίο σχηματίστηκε το $[S]$. Μπορεί, επομένως, να διαγραφεί, χωρίς να βλάψει την απόδοση του συστήματος.

Η εφαρμογή της παραπάνω λογικής, μόνο στο χαμηλότερο επίπεδο κάλυψης για το $[M]$, δηλαδή μόνο για τους κανόνες που καλύπτουν το μικρότερο αριθμό δειγμάτων από τους κανόνες του Match Set για δεδομένο δείγμα i, είναι η λειτουργία της συνάρτησης $controlMatchSet$ στη γραμμή 4 του αλγορίθμου 6.1. Η επιλογή του $[S] \equiv [M]$ έναντι του $[C]$ έγινε διότι:

- θεωρούμε το $[M]$ ως ένα πρώτο βήμα μελέτης αυτού του μηχανισμού διαγραφής
- το $[M]$ δε συγκεντρώνει κανόνες ανά ετικέτα, συνεπώς διαγράφοντας με βάση την καταλληλότητα, αφαιρούμε κανόνες βάσει της συνολικής τους επίδοσης, από μία μαζικότερη δεξαμενή κανόνων[8]
- ενδεχομένως να ήταν άδικο να διαγράφουμε κανόνες από τα Correct Sets, καθώς εκεί περιέχονται οι κανόνες που αποφασίζουν ορθά για τις διάφορες ετικέτες
- ανάλογα με τον αριθμό ετικετών του συνόλου δεδομένων, η διαγραφή από τα $[C]$ ίσως γίνει φορτικότερη, ενώ αποδυναμώνει την ποικιλότητα κανόνων στο χαμηλότερο επίπεδο κάλυψης ανά σύνολο.

Ας εξετάσουμε τη διαδικασία διαγραφής κανόνων από πιο κοντά. Έστω το δείγμα i του συνόλου δεδομένων $|D|$ και $[P]$ ο πληθυσμός των κανόνων του ΜαΣΤ. Με την παρουσίαση του i στο ΜαΣΤ, οι κανόνες του $[P]$ που το καλύπτουν σχηματί-

[7]Ονομάζουμε επίπεδο κάλυψης δειγμάτων τον αριθμό δειγμάτων που καλύπτει ένας κανόνας. Η φράση τοποθετείται σε συμφραζόμενα συνόλου κανόνων.

[8]Χωρίς βλάβη της γενικότητας, για δεδομένο δείγμα και για το σύνολο των ετικετών, $[C] \subseteq [M]$.

ζουν το Match Set $[M]$, πληθικότητας M κανόνων. Ο Πίνακας 6.5 αναπαριστά το σύνολο $[M]$, με f_k και cov_k την καταλληλότητα και τον αριθμό δειγμάτων που καλύπτει ο κανόνας k του $[M]$, αντίστοιχα. Έστω cov_{min} ο ελάχιστος αριθμός δειγμάτων που καλύπτεται από τους κανόνες του $[M]$. Συγκεντρώνουμε τους N κανόνες που καλύπτουν αριθμό δειγμάτων ίσο με cov_{min} στο σύνολο $[MD]$ (Πίνακας 6.6), διατάσσοντάς τους κατά μειούμενη καταλληλότητα $f_{m0} \geq f_{m1} \geq f_{m2} \geq \ldots \geq f_{mN-2} \geq f_{mN-1}$. Η τιμή του πεδίου $D.checked$ υποδηλώνει το εάν σε έναν κανόνα έχουν παρουσιαστεί όλα τα δείγματα του $|D|$ τουλάχιστον μία φορά. Κανόνες με τιμή $D.checked = false$ σημαίνει ότι έχουν δημιουργηθεί προτού παρουσιαστούν στο ΜαΣΤ $|D|$ δείγματα σε σχέση με το χρόνο σχηματισμού του $[M]$.

Πίνακας 6.5: Ενδεικτικό Match Set με κανόνες σε διάφορα επίπεδα κάλυψης. Μπορεί να ισχύει $cov_i = cov_j$ για $i \neq j$.

$rule\#$	$fitness$	$coverage(instances)$
0	f_0	cov_0
1	f_1	cov_1
2	f_2	cov_2
$\vdots$	$\vdots$	$\vdots$
$M-2$	f_{M-2}	cov_{M-2}
$M-1$	f_{M-1}	cov_{M-1}

Πίνακας 6.6: Υποσύνολο κανόνων του Match Set του Πίνακα 6.5 που ανήκουν στο χαμηλότερο επίπεδο κάλυψης cov_{min}.

$fitness$	$coverage$	$D.checked$
f_{m0}	cov_{min}	$true$
f_{m1}	cov_{min}	$false$
f_{m2}	cov_{min}	$true$
$\vdots$	$\vdots$	$\vdots$
f_{mN-2}	cov_{min}	$true$
f_{mN-1}	cov_{min}	$false$

Η διαδικασία διαγραφής από το Match Set λειτουργεί αφαιρώντας από τον πληθυσμό $[P]$, τον κανόνα k εκείνο του συνόλου $[MD]$ ο οποίος έχει $D.checked_k = true$ και τη χαμηλότερη τιμή καταλληλότητας (ανάμεσα στους υπόλοιπους κανόνες του $[MD]$ οι οποίοι έχουν κληθεί να συμμετάσχουν σε περισσότερα από $|D|$ Match Sets μέχρι τη στιγμή δημιουργίας του $[M]$). Στην περίπτωση του Πίνακα 6.6, αυτός ο κανόνας θα ήταν ο f_{mN-2}. Εάν υπάρχει μόνο ένας κανόνας που να ικανοποιεί τις παραπάνω συνθήκες, αυτός δε διαγράφεται, καθώς είτε είναι ο μόνος στο χαμηλό-

τερο επίπεδο κάλυψης, οπότε θα θέλαμε να τον συμπεριλάβουμε σε κάθε περίπτωση στον $[P]$, είτε υπάρχουν περισσότεροι κανόνες στο $[MD]$, αλλά με $D.checked = false$, οπότε για το συγκεκριμένο $[M]$ η διαδικασία θα ενεργοποιηθεί σε αργότερο χρόνο.

Αυτό που καταφέρνουμε, εν τέλει, με την παραπάνω προσέγγιση είναι να εξελιχθεί ένας πληθυσμός κανόνων που κινείται σε δύο συνιστώσες: α) ένα τμήμα του αποτελείται από ικανώς γενικούς κανόνες και β) ένα άλλο αποτελείται από ειδικούς κανόνες υψηλής καταλληλότητας, που λειτουργεί συμπληρωματικά ως προς το πρώτο, εξερευνώντας ικανοποιητικά το χώρο που το πρώτο δεν έχει τη δυνατότητα να καλύψει.

Στον Αλγ. 6.1, παρατηρούμε πως η λειτουργία διαγραφής στα Match Sets ενεργοποιείται όταν ικανοποιείται η συνθήκη

$$rouletteWheelDeletionsCommenced = true \tag{6.17}$$

Οι διαγραφές στα Match Sets, δηλαδή, πραγματοποιούνται στο τμήμα της εκπαίδευσης εκείνο, στο οποίο ο πληθυσμός έχει φτάσει στο άνω αριθμητικό του όριο και, συνεπώς, έχει ξεκινήσει η διαγραφή κανόνων από τον πληθυσμό $[P]$. Με άλλα λόγια, οι δύο λειτουργίες διαγραφής εκκινούν (τυπικά) ταυτόχρονα.

Πειραματικά, αυτή η προσέγγιση φαίνεται ορθότερη από ότι αν η διαδικασία διαγραφής ξεκινούσε από την αρχή της εκπαίδευσης. Αυτή η κίνηση θα παρέκκλινε από τη φιλοσοφία των ΜαΣΤ, όσον αφορά στο κομμάτι της εξερεύνησης, προτού ξεκινήσει οποιαδήποτε μέθοδος διαγραφής κανόνων. Κάτι τέτοιο είναι φυσιολογικό, καθώς, θα παρεμβαίναμε, και μάλιστα από τα πρώτα βήματα της εξερεύνησης, στη διαδικασία εύρεσης (λειτουργία Κάλυψης) και εξέλιξης των κανόνων (Γενετικός Αλγόριθμος). Στα πρώτα στάδια εκπαίδευσης, ανάλογα και με τον ρυθμό γενίκευσης γνωρισμάτων $P_{\#A}$, θα διαγράφαμε κανόνες πλησιέστερους προς τα δείγματα, μειώνοντας τις πιθανότητες διασταύρωσης με κάποιον αντίστοιχα λιγότερο γενικό κανόνα, ώστε να παράξουν από κοινού τους ζητούμενους απογόνους που καλύπτουν το χώρο που δεν μπορούν οι περισσότερο γενικοί κανόνες. Έπειτα, αν ξεκινούσαν από την αρχή της εκπαίδευσης οι διαγραφές στα Match Sets, λόγω της μικρής πληθικότητας αυτών, θα υπήρχε μεγάλη πιθανότητα διαγραφής κανόνων που συμμετέχουν σε ένα $[M]_i$ με άλλους κανόνες, αλλά είναι μοναδικοί για κάποιο άλλο $[M]_j$, παραβιάζοντας τη συνθήκη μη διαγραφής κανόνων μοναδικών στο χαμηλότερο επίπεδο κάλυψης. Στατιστικά, αυτή η πιθανότητα είναι μη μηδενική και στην περίπτωση διαγραφής για το διάστημα όπου $rouletteWheelDeletionsCommenced = true$, αλλά σίγουρα μικρότερη από προηγουμένως[9].

[9] Η πιθανότητα αυτή εξαρτάται από το μέγεθος του πληθυσμού $[P]$, τον αριθμό των επαναλήψεων εκπαίδευσης $|I|$, και την παράμετρο θ_{GA}:

$$P_M(wrongfulDeletion) \propto \frac{|I|}{\theta_{GA} \cdot |\mathbf{P}|} \tag{6.18}$$

Παρατηρησεις πανω σε λειτουργιες του GMl-ASLCS

Στην παρούσα ενότητα κάνουμε μία σειρά παρατηρήσεων πάνω στη συμπεριφορά μεμονωμένων λειτουργιών του GMl-ASLCS, αλλά και στη συνολική συμπεριφορά του χρησιμοποιώντας ποιοτικά κριτήρια.

Μη Συμμετοχή των Αδιαφοριών Στα Correct Sets

Επειδή το σύνολο από όπου αντλεί ο Γενετικός Αλγόριθμος τους κανόνες που εξελίσσει είναι το $[C_l]$ (για δεδομένη ετικέτα l), με τον αποκλεισμό των κανόνων που αδιαφορούν για την l από τη συμμετοχή σε αυτό, αποκλείονται ενεργητικά από την εξελικτική διαδικασία οι κανόνες που αδιαφορούν για την l. Έτσι, ελαχιστοποιείται η εξάπλωση των αδιαφοριών στο τμήμα της απόφασης των κανόνων, καθώς δεν υπάρχει πίεση για παραγωγή κανόνων που να αδιαφορούν για κάθε ετικέτα l. Τονίζουμε τη λέξη ενεργητικά παραπάνω, γιατί κανόνες που αποφασίζουν σαφώς για την ετικέτα l (και συμμετέχουν, συνεπώς, στο $C[l]$), μπορούν κάλλιστα να αδιαφορούν για άλλες ετικέτες. Ουσιαστικά, λοιπόν, η παρουσία των αδιαφοριών δεν εξαλείφεται - μόνο ελαχιστοποιείται.

Οι παραπάνω υποθέσεις μας αποδεικνύονται πειραματικά, καθώς με την άρση της απαγόρευσης συμμετοχής των κανόνων που αδιαφορούν για μία δεδομένη ετικέτα στο αντίστοιχο $[C_l]$, παρατηρούμε σημαντική αύξηση του μέσου ποσοστού αδιαφορίας στις ετικέτες, αύξηση των αδιαφοριών στις συνθήκες και, ταυτόχρονα, όπως είναι αναμενόμενο, πτώση της ακρίβειας του ΜαΣΤ. Η εξήγηση είναι σχετικά απλή. Καταρχήν, ο Γενετικός Αλγόριθμος οδηγεί τους κανόνες προς τη γενικότητα, καθώς όσο περισσότερο γενικός είναι ένας κανόνας, σε τόσα περισσότερα $[M]$ θα συμμετέχει, άρα σε τόσα περισσότερα $[C]$, και άρα τόσο περισσότερες πιθανότητες θα έχει για να αναπαραχθεί. Αν χρησιμοποιούσαμε διασταύρωση ενός σημείου, τα πράγματα θα ήταν χειρότερα, καθώς οι αδιαφορίες στο τμήμα απόφασης των γονέων θα ανταλλάσσονταν μεταξύ τους, αναπαράγοντας το φαινόμενο της αύξησης αδιαφοριών στις ετικέτες και καταλήγοντας σε ακόμα μεγαλύτερα $[C]$.

Επιπλέον, ανεξάρτητα από τον χρησιμοποιούμενο τελεστή διασταύρωσης, εφόσον πριμοδοτούμε κάθε αδιαφορία στις ετικέτες με $\omega = 0.9$, η διαφορά των κανόνων που αποφασίζουν σαφώς και "ανταμείβονται" με $\omega = 1$ δεν είναι αρκετή για να κάνει το Γενετικό Αλγόριθμο να γείρει την πλάστιγγα υπέρ τους. Έτσι, στην ουσία, υποσκάπτονται οι κανόνες με σαφείς αποφάσεις, τόσο στην εξελικτική διαδικασία, όσο και όσον αφορά στη διαγραφή τους (όσο περισσότεροι κανόνες συμμετέχουν στα $[C]$, τόσο μεγαλύτερη είναι και η εκτίμηση για το μέγεθος cs). Το αποτέλεσμα είναι η εξέλιξη υπέρ-γενικών κανόνων, τόσο στο τμήμα της συνθήκης, όσο και στο τμήμα τις απόφασης, με καταστροφικό αποτέλεσμα στην αποτελεσματικότητα του συστήματος.

Ένας τρόπος αντιμετώπισης αυτού του φαινομένου θα φαινόταν ότι είναι η χρήση χαμηλότερης τιμής του ω για την ανταμοιβή των αδιαφοριών. Σε αυτή την περίπτωση, όμως, κανόνες εν γένει ακριβείς που αδιαφορούν έστω και για ένα μικρό αριθμό από ετικέτες, δεδομένου της μεγάλης τιμής του ν στον υπολογισμό της καταλληλότητας, θα θεωρούνταν αναξιόπιστοι και, συν τοις άλλοις δεν θα είχαν αρκετές πιθανότητες να επιλεγούν για αναπαραγωγή από το Γενετικό Αλγόριθμο.

Εν τέλει, θα ωθούσαμε το σύστημα να αξιοποιήσει κανόνες που θα ήταν πλήρως σαφείς στις αποφάσεις τους, με πιθανό κόστος τη χαμηλή κάλυψή τους, οδηγώντας και το σύνολο του πληθυσμού σε χαμηλότερα επίπεδα μέσης κάλυψης.

Σε αυτό το σημείο αξίζει να αναφερθούμε στην αξία της γενίκευσης ετικετών κατά τη λειτουργία κάλυψης που αναφέραμε στην Παρ. 5.3.2, σε συνδυασμό με τον αποκλεισμό των κανόνων που δεν αποφασίζουν για κάποια ετικέτα στη συμμετοχή στα $[C]$. Στις περιπτώσεις όπου ένα δείγμα καλύπτεται από ένα σύνολο κανόνων το οποίο δεν αποφασίζει ευθέως για μία ετικέτα l ενεργοποιείται ο τελεστής κάλυψης, αφού προκύπτει κενό C_l. Δεδομένης της χαμηλής τιμής της πιθανότητας γενίκευσης $P_{\#L}$, το σύστημα είναι εξαιρετικά πιθανό να δημιουργήσει μέσω κάλυψης (κάλλιστα ο εν λόγω κανόνας μπορεί να προκύψει και μέσω του Γενετικού Αλγορίθμου) έναν κανόνα που να αποφασίζει σαφώς για την l και, μάλιστα, η απόφασή του θα είναι ίδια με την τιμή της l. Λόγω και της επαναληπτικής φύσης των ΜαΣΤ, το σύστημα θα εξελίξει σταδιακά κανόνες που καλύπτουν όλα τα δείγματα του συνόλου δεδομένων και, επιπρόσθετα, όλες τις ετικέτες τους, παρέχοντας με αυτόν τον τρόπο πλήρη κάλυψη, τόσο γνωρισμάτων όσο και ετικετών, από το σύνολο των κανόνων.

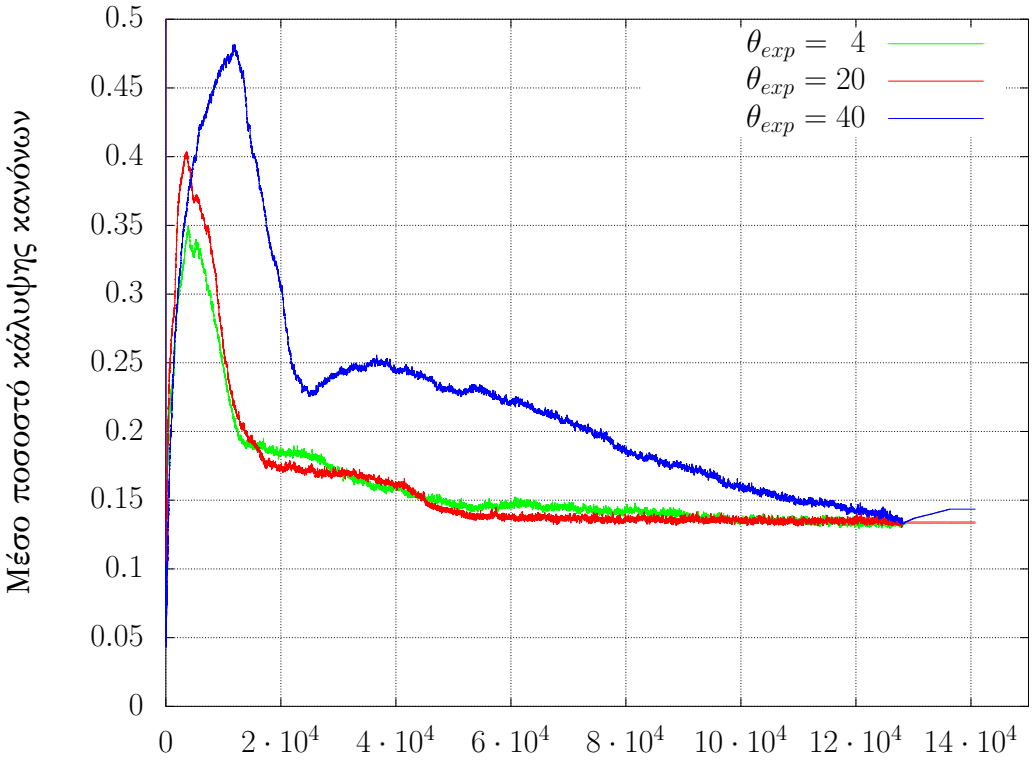
Για την ΄Εκπτωση Καταλληλότητας στη Συνιστώσα Εξερεύνησης

Όπως έχουμε προαναφέρει, η συνιστώσα εξερεύνησης και πιο συγκεκριμένα ο Γενετικός Αλγόριθμος στον GMl-ASLCS, επιλέγει κανόνες βάσει της καταλληλότητάς τους, μετά την εφαρμογή έκπτωσης με βάση την εμπειρία (Εξ. 5.4). Η παράμετρος θ_{exp} προστίθεται και αυτή στο σύνολο των παραμέτρων που πρέπει να ρυθμιστούν κατάλληλα, σε συνδυασμό και με τις υπόλοιπες, για το κάθε σύνολο δεδομένων, ανάλογα με τους στόχους ακρίβειας ή γενίκευσης που έχουμε θέσει.

Στο γράφημα του Σχήματος 6.6 παρατίθενται οι καμπύλες εξέλιξης του μέσου ποσοστού κάλυψης δειγμάτων από τους κανόνες του GMl-ASLCS για το πρόβλημα $mlPosition7$. Με πράσινο χρώμα παριστάνεται η πορεία της μέσης κάλυψης για $\theta_{exp} = 4$, με κόκκινο για $\theta_{exp} = 20$ και με μπλε για $\theta_{exp} = 40$. Οι υπόλοιπες παράμετροι των τριών διακριτών πειραμάτων κρατήθηκαν ίδιες και η τελική ακρίβεια είχε την ίδια τιμή και για τα τρία.

Μερικά πράγματα που μπορούμε να παρατηρήσουμε, είναι:

- Αύξηση του κατωφλίου εμπειρίας θ_{exp} σημαίνει πως θα περάσουν περισσότερες επαναλήψεις μέχρι το ΜαΣΤ να εμπιστευτεί έναν κανόνα και να του επιτρέψει να συμμετάσχει στην εξελικτική διαδικασία. Όπως είναι φυσικό, κανόνες που είναι γενικότεροι, καλύπτουν περισσότερα δείγματα του συνόλου δεδομένων, άρα συμμετέχουν σε περισσότερα Match Sets και, συνεπώς, η εμπειρία τους αυξάνεται με μεγαλύτερο ρυθμό από τους ειδικότερους από αυτούς κανόνες. Εφόσον οι κανόνες που σπάνε το φράγμα του θ_{exp} πρώτοι και συχνότερα είναι γενικοί, αυτό σημαίνει πως ο πληθυσμός των κανόνων οδηγείται από νωρίς σε αυξημένα επίπεδα γενίκευσης. Από μόνο του αυτό το γεγονός δεν είναι απαραίτητα ζημιογόνο, αλλά σε περιπτώσεις ανισοκατανομής των διακριτών συνδυασμών ετικετών ή/και επιλογής υπό-βέλτιστου συνδυασμού παραμέτρων $(|I|, \theta_{GA}, maxPopulationSize, P_{\#A})$ είναι δυνατό ο Γενετικός Αλ-



Σχήμα 6.6: Καμπύλες εξέλιξης του μέσου αριθμού κάλυψης δειγμάτων των κανόνων του πληθυσμού για τιμές $\theta_{exp} = \{4, 20, 40\}$ στο σύνολο δεδομένων $mlPosition7$.

γόριθμος να μην μπορέσει να παράξει τους ειδικούς κανόνες εκείνους που θα καλύψουν τους διακριτούς συνδυασμούς ετικετών που εκπροσωπούνται από τα λιγότερα δείγματα στο σύνολο δεδομένων.

- Όσο αυξάνει το κατώφλι θ_{exp}, τόσο μετατοπίζεται αργότερα το σημείο στο οποίο ο πληθυσμός των κανόνων συναντά το ανώτατο του όριο, και άρα εκκινούν οι διαγραφές. Όσο γενικότερος ένας κανόνας, σε τόσα περισσότερα Correct Sets συμμετέχει, άρα τόσο περισσότερο κοντά στην τρέχουσα τιμή της χρονοσφραγίδας $timestamp$ του συστήματος θα βρίσκεται αυτή του κανόνα. Όταν οι κανόνες που καταφέρνουν αρχικά να ξεπεράσουν το κατώφλι είναι αναγκαστικά γενικότεροι, αυτό σημαίνει πως και τα περισσότερα Correct Sets θα πληρώνονται με κανόνες που καθιστούν μικρότερη τη διαφορά $timestamp - \overline{timestamp}([C_l])$, για δεδομένο l, και συνεπώς ο ρυθμός παραγωγής κανόνων καθίσταται μικρότερος από αυτόν για μικρότερες τιμές του κατωφλίου εμπειρίας θ_{exp}.

- Όπως φανερώνει η καμπύλη με μπλε χρώμα, όσο αυξάνει το κατώφλι θ_{exp}, τόσο περισσότερο χρόνο χρειάζεται το ΜαΣΤ για να συγκλίνει προς τη βέλτιστη λύση. Αυτό σημαίνει πως αυξημένες τιμές του κατωφλίου, δυνητικά να χρειάζονται και μεγαλύτερο αριθμό επαναλήψεων εκπαίδευσης.

- Αν παρατηρήσουμε το διάστημα ενημέρωσης, στο τελευταίο μέρος της εκπαίδευσης (για συνολική παρουσίαση περίπου $13 \cdot 10^4$ δειγμάτων), θα δούμε ότι το ΜαΣΤ, για $\theta_{exp} = 40$, λόγω ίσως και της αργής του σύγκλισης, αποβάλλει τους κανόνες στο χαμηλότερο επίπεδο κάλυψης στα Match Sets (σε αντίθεση με τις προσεγγίσεις για $\theta_{exp} = \{4, 20\}$ που τους έχουν αποβάλει ήδη), αυξάνοντας έτσι, εν τέλει, το μέσο ποσοστό κάλυψης των κανόνων του τελικού μοντέλου για το ίδιο πρόβλημα.

Οι παραπάνω παρατηρήσεις μας οδηγούν στα εξής συμπεράσματα: α) Ο ακριβής προσδιορισμός του κατωφλίου θ_{exp} θα πρέπει να γίνει σε σχέση με τις υπόλοιπες παραμέτρους του ΜαΣΤ, β) η διαδικασία προσδιορισμού του, ανάλογα με το σύνολο δεδομένων, είναι χρονοβόρα διότι και εδώ θα πρέπει να εφαρμόσουμε μια διαδικασία trial-and-error, γ) ίσως θα ήταν συμφερότερο να χρησιμοποιήσουμε δύο κατώφλια εμπειρίας, ένα για το διάστημα όπου η πληθικότητα του πληθυσμού είναι μόνιμα κάτω από το άνω όριό του και ένα για το διάστημα μετά την έναρξη των διαγραφών, αν και τότε θα έπρεπε να προσδιορίσουμε δύο κατώφλια αντί για ένα, και δ) ίσως θα έπρεπε να εγκαταλείψουμε εξ ολοκλήρου την προσέγγιση που θέτει ένα αυθαίρετο κατώφλι εμπειρίας και να βρούμε μία αντικειμενική συνθήκη για την έκπτωση της καταλληλότητας.

Μία τέτοια μεθοδολογία, όπως προσδιορίζεται από το τελευταίο συμπέρασμα, θα μπορούσε να βασίζεται στην αντικατάσταση του κατωφλίου εμπειρίας, και άρα της σχετικότητας του βαθμού γενίκευσης ενός κανόνα, από μία δυαδική σχέση που θα βασίζεται στο αν ο κανόνας έχει κληθεί να συμμετάσχει σε παραπάνω Match Sets από ότι είναι ο αριθμός των δειγμάτων του συνόλου δεδομένων, ή όχι.

$$fitness'(i) = \begin{cases} 0, & D.checked = false \\ \left(\frac{tp(i)}{msa(i)}\right)^{\nu}, & \text{αλλού} \end{cases} \tag{6.19}$$

Με αυτό τον τρόπο, απομακρύνουμε τον προσδιορισμό ακόμα μίας παραμέτρου (δύο για την ακρίβεια, όπως θα φανεί παρακάτω) και θέτουμε ένα αντικειμενικό κριτήριο εμπιστοσύνης του συστήματος προς τους κανόνες. Ιδίως εάν το πρόβλημα διαθέτει μικρό αριθμό ετικετών, όμως, ίσως χρειαστούν περισσότερες επαναλήψεις για τη σύγκλιση σε κάποια βέλτιστη λύση, καθώς θα καθυστερείται η ανανέωση του πληθυσμού των κανόνων.

Τέλος, θα πρέπει να επισημάνουμε πως το κατώφλι θ_{exp} στη συνιστώσα εξερεύνησης και το κατώφλι θ_{del} στη λειτουργία διαγραφής είναι οι δύο όψεις του ίδιου νομίσματος και ο προσδιορισμός του ενός πρέπει να γίνει σε συσχέτιση με τον προσδιορισμό του άλλου. Εάν κρατήσουμε τη δομή της Εξ. 6.12 (που αφορά στην ποσότητα $d(i)$ η οποία προσδιορίζει την πιθανότητα διαγραφής ενός κανόνα) και σταθερό το θ_{exp}, ανεξάρτητα από τις ακριβείς πιθανότητες διαγραφής, η διαφορά

$$\theta_{del} - \theta_{exp} > 0 \tag{6.20}$$

θα προσδιορίσει το διάστημα χάριτος στο οποίο κανόνες θα έχουν τη δυνατότητα, ανάλογα με την καταλληλότητά τους, να αποτελέσουν υποψήφιους γονείς και να διαιωνίσουν τα γονίδιά τους, χωρίς να κινδυνεύουν να διαγραφούν πρόωρα σε αντίθετη περίπτωση.

Διάστημα Ενημέρωσης

Στο γράφημα του Σχήματος 6.4 παρατηρούμε την πτώση του αριθμού των κανόνων του πληθυσμού στις τελευταίες επαναλήψεις. Η διαδικασία εκπαίδευσης περιλαμβάνει δύο διαδοχικά στάδια: α) την εκπαίδευση του ΜαΣΤ, όπου κάθε δείγμα του συνόλου δεδομένων εισάγεται σε αυτό $|I|$ φορές και στην οποία η λειτουργία του Γενετικού Αλγορίθμου είναι ενεργοποιημένη και β) το διάστημα ενημέρωσης, $u \cdot |I|$ επαναλήψεων, με $u \in [0, 1]$, όπου η παραγωγή νέων κανόνων είναι απενεργοποιημένη (άρα και η λειτουργία διαγραφής κανόνων μέσω επιλογής ρουλέτας).

Το διάστημα ενημέρωσης ενσωματώνεται στη διαδικασία εκπαίδευσης ώστε να παρέλθει αρκετός χρόνος, δηλαδή ικανός αριθμός επαναλήψεων, ώστε η πραγματική τιμή της καταλληλότητας των κανόνων και η καταλληλότητα μετά την έκπτωσή της που βασίζεται στην εμπειρία Εξ. 5.4, να ισούνται αριθμητικά. Γενικότερα, αυτό το διάστημα εξυπηρετεί στη σταθεροποίηση των παραμέτρων των κανόνων (όπως η εκτίμηση cs ή, σημαντικότερα, η καταλληλότητά τους) που δημιουργήθηκαν στην τελευταία επανάληψη της διαδικασίας εκπαίδευσης. Σε κάθε περίπτωση, το διάστημα ενημέρωσης, ανεξάρτητα από το μήκος του, είναι αναπόσπαστο κομμάτι της διαδικασίας δημιουργίας του τελικού μοντέλου, ώστε η συνιστώσα επίδοσης να χρησιμοποιεί τις ακριβείς τιμές καταλληλότητας των κανόνων και να ταξινομήσει ορθά, με βάση το πλήρες σύνολο κανόνων που παρήγε το ΜαΣΤ. Αν η συνιστώσα επίδοσης ενός ΜαΣΤ κάνει και αυτή χρήση της έκπτωσης της καταλληλότητας με βάση την εμπειρία, θα πρέπει να ισχύει:

$$u \cdot |I| > \theta_{exp} \tag{6.21}$$

Σε διαφορετική περίπτωση, αρκεί έστω και μία τελική επανάληψη με το Γενετικό Αλγόριθμο απενεργοποιημένο, ώστε οι κανόνες που δημιουργήθηκαν στην τελευταία επανάληψη να εξετάσουν για μία φορά το σύνολο δεδομένων και να σταθεροποιήσουν τις παραμέτρους τους.

Αν και η συνιστώσα επίδοσης του GMl-ASLCS χρησιμοποιεί απευθείας τις τιμές της ακρίβειας (και όχι της καταλληλότητας) των κανόνων για το συμπερασμό των ετικετών των δειγμάτων του συνόλου ελέγχου, εδώ χρησιμοποιούμε το διάστημα ενημέρωσης για να απομακρύνουμε από τον πληθυσμό κανόνες που πληρούν τις προϋποθέσεις αφαίρεσής τους από τον πληθυσμό, μέσω του μηχανισμού Διεύρυνσης της Μέσης Κάλυψης που περιγράφηκε στην Παρ. 6.3.2. Η μείωση του αριθμού των κανόνων που παρατηρούμε στο γράφημα του Σχήματος 6.4 στις τελευταίες επαναλήψεις, οφείλεται ακριβώς στην εφαρμογή του μηχανισμού αυτού.

Παρατηρήσεις πάνω στη μέση κάλυψη

Μία εικόνα για τον αριθμό δειγμάτων που καλύπτουν οι κανόνες του πληθυσμού λαμβάνουμε από το μέσο ποσοστό κάλυψης δειγμάτων των κανόνων, το οποίο παριστάνεται με κόκκινο χρώμα για το σύνολο δεδομένων $emotions$ στο γράφημα του Σχήματος 6.5. Αυτό που παρατηρούμε είναι ότι το μέσο ποσοστό δειγμάτων που καλύπτουν οι κανόνες ακολουθεί τη μορφή της πορείας του μέσου αριθμού κανόνων των Corrects Sets: αυξάνει μέχρι ο πληθυσμός να φτάσει για πρώτη φορά το άνω όριό του και για το διάστημα που ακολουθεί, φθίνει, μέχρι το τέλος της εκπαίδευσης. Όσο πιο γενικός είναι ένας κανόνας, τόσα περισσότερα δείγματα καλύπτει, άρα συμμετέχει σε περισσότερα Match Sets και άρα σε περισσότερα Correct Sets. Στους γενικότερους κανόνες δίνονται περισσότερες ευκαιρίες αναπαραγωγής λόγω ακριβώς του γεγονότος ότι συμμετέχουν σε περισσότερα $[C]$ από ότι οι υπόλοιποι κανόνες, παράγοντας απογόνους που διατηρούν και διευρύνουν αυτή τη γενίκευση πάνω στα δείγματα. Αυτή η ροπή συμμετοχής σε περισσότερα $[C]$, προκαλεί την αύξηση του μεγέθους των Correct Sets. Όμως, το κυρίαρχο κριτήριο διαγραφής για έναν κανόνα είναι το μέγεθος των Correct Sets στα οποία αυτός συμμετέχει, το οποίο είναι και ικανή συνθήκη για να διαγραφεί, αλλά όχι αναγκαία. Ανάμεσα στους κανόνες που διαγράφονται, βρίσκονται και αυτοί που χαρακτηρίζονται από αυξημένη γενικότητα σε σχέση με τους υπόλοιπους. Για αυτό ακριβώς χρησιμοποιούμε και ως κριτήριο διαγραφής την καταλληλότητα, ώστε να διαχωρίσουμε τους κανόνες που είναι υπέρ-γενικοί και η προβλεπτική τους ικανότητα είναι χαμηλή, από τους κανόνες που καταφέρνουν να ισορροπήσουν ανάμεσα στην επαρκή γενικότητα και την ικανοποιητική καταλληλότητα[10].

Είναι εύκολο να παρατηρήσει κανείς, με βάση και το περιεχόμενο της παραπάνω παραγράφου, ότι μεγαλύτερα μεγέθη πληθυσμών, με σταθερό ρυθμό παραγωγής κανόνων θ_{GA} και αριθμό επαναλήψεων, θα είχαν ως αποτέλεσμα μεγαλύτερα επίπεδα τιμών μέσου μεγέθους των Correct Sets $\overline{cs}$ και μεγαλύτερα επίπεδα κάλυψης.

Πράγματι, θέτοντας το μέγιστο αριθμό μικρο-κανόνων που μπορεί να συγκρατήσει ο πληθυσμός σε μεγαλύτερες τιμές, το σημείο καμπής στο οποίο εκκινούν οι διαγραφές κανόνων μετατοπίζεται σε αργότερο χρόνο, όπου, εν γένει, τα διάφορα Correct Sets αποτελούνται από περισσότερους κανόνες και συνεπώς όπου το $\overline{cs}$ είναι αυξημένο, όπως και η μέση κάλυψη των τελικών κανόνων. Αυτή, βέβαια, δε δύναται να είναι συνεχώς αύξουσα: αν δεν ετίθετο όριο στο μέγεθος του πληθυσμού και των επαναλήψεων δε θα έφτανε στην τιμή 1.

Μικρότερη τιμή της παραμέτρου θ_{GA}, για ίσο αριθμό μικρο-κανόνων, θα είχε ως αποτέλεσμα τη μετατόπιση του παραπάνω σημείου νωρίτερα, λόγω του μεγαλύτερου ρυθμού παραγωγής κανόνων και άρα της πιο γρήγορης άφιξης του πληθυσμού στο όριό του. Αν αυξάναμε τον αριθμό των επαναλήψεων, κρατώντας σταθερά το αριθμητικό όριο του πληθυσμού και το θ_{GA}, θα παρατηρούσαμε μία συνεχή πτώση τόσο του μέσου μεγέθους των $[C]$, όσο και της μέσης κάλυψης των κανόνων. Η πτώση αυτή, επειδή είναι εκθετική, προκαλεί μείωση των παραπάνω μεγεθών με όλο και μικρότερο βαθμό, δηλαδή παρατηρείται μικρότερος ρυθμός πτώσης τους όσο αυξάνουν οι επαναλήψεις.

[10] Εν γένει, αυτή η εργασία αποφεύγει τη χρήση απόλυτων όρων, καθώς, η λύση κάθε προβλήματος απαιτεί διαφορετικά επίπεδα γενικότητας ή/και καταλληλότητας των κανόνων.

Παράλληλα, ο ρυθμός γενίκευσης γνωρισμάτων $P_{\#A}$ είναι και αυτός ένας καθοριστικός παράγοντας στη διαδικασία εκπαίδευσης, καθορίζοντας άμεσα την ισορροπία ανάμεσα στη γενίκευση και την ακρίβεια του τελικού μοντέλου. Κάθε σύνολο δεδομένων απαιτεί και διαφορετικό $P_{\#A}$, ανάλογα με το βαθμό πληρότητας του, τον καθορισμό των παραπάνω τριών μεταβλητών και τη δυνατότητα των δειγμάτων του συνόλου δεδομένων να "ομαδοποιηθούν" στην αναπαράστασή τους από κανόνες, τόσο στο τμήμα συνθήκης όσο και στο τμήμα απόφασής τους.

Σε κάθε περίπτωση, η διαδικασία εύρεσης ενός αξιόπιστου συνόλου

$$(|I|, \theta_{GA}, maxPopulationSize, P_{\#A})$$

είναι επίπονη για κάθε σύνολο δεδομένων D, λόγω της εσωτερικής δομής του D και της στοχαστικής φύσης των ΜαΣΤ.

Εν γένει, η τιμή της παραμέτρου θ_{GA} θα πρέπει να καθορίζεται με γνώμονα τη μη βιαιότητα ανανέωσης κανόνων στο εσωτερικό του πληθυσμού. Μικρές τιμές της σημαίνει ότι ο ρυθμός παραγωγής κανόνων γίνεται μεγαλύτερος, συνεπώς και ότι ο ρυθμός διαγραφής κανόνων γίνεται μεγαλύτερος. Σε αυτό το πλαίσιο λειτουργίας, στο ζυγό της καταλληλότητας και της γενικότητας, η πλάστιγγα γέρνει υπέρ της καταλληλότητας για δύο λόγους. Πρώτον, χρειάζονται λιγότερες επαναλήψεις για την επίτευξη του ισοδύναμου μεγέθους μέσης κάλυψης λόγω της μεγαλύτερης παραγωγής και διαγραφής κανόνων και βάσει των παρατηρήσεων των δύο πρώτων παραγράφων αυτής της ενότητας και, δεύτερον, υπάρχει μεγαλύτερη πίεση για εύρεση γονέων πλέον κατάλληλων και, συνεπώς, λιγότερο γενικών. Αντίστροφα, μία μεγάλη τιμή του θ_{GA} δε θα παρείχε επαρκή πίεση προς την εύρεση κατάλληλων κανόνων, λόγω και της αραιότερης παραγωγής κανόνων, ενώ θα μετατόπιζε τον ορίζοντα διαγραφών αργότερα, έχοντας σαν αποτέλεσμα αυξημένη γενικότητα κανόνων, λόγω της αραιής παραγωγής και διαγραφής κανόνων, σε βάρος της εξέλιξης των κατάλληλων λύσεων. Εν κατακλείδι, η διαδικασία καθορισμού μίας λειτουργικής τετράδας παραμέτρων $(|I|, \theta_{GA}, maxPopulationSize, P_{\#A})$ δεν μπορεί να ιδωθεί ως μία αρθρωτή διαδικασία, ξεχωριστή για την καθεμία, αλλά ως ένα συνεργατικό έργο.

Για το ρυθμό μεταβολής του μεγέθους του πληθυσμού στα πρώιμα στάδια της εκπαίδευσης

Στα Σχήματα 6.3 και 6.4, παρατηρούμε την απότομη άνοδο του αριθμού των μικρο-κανόνων του πληθυσμού μετά την είσοδο περίπου $5 \cdot 10^4$ δειγμάτων στο ΜαΣΤ. Μέχρι αυτό το κρίσιμο σημείο, το ΜαΣΤ έχει παράξει έναν περιορισμένο αριθμό κανόνων μέσω της λειτουργίας κάλυψης, ενώ η πλειοψηφία των κανόνων έχει δημιουργηθεί μέσω του Γενετικού Αλγορίθμου. Ωστόσο, λόγω του σχετικά μεγάλου αριθμού ετικετών του προβλήματος enron, με $|L| = 53$, και της χαμηλής τιμής της τρέχουσας χρονοσφραγίδας, η κατανομή των κανόνων ανά $[C]$ είναι τέτοια που ο μέσος όρος των χρονοσφραγίδων των κανόνων ξεπερνά το κατώφλι $timestamp([C]) - \theta_{GA}$ για μικρό αριθμό από Correct Sets. Υπενθυμίζουμε πως για να ενεργοποιηθεί η διαδικασία παραγωγής κανόνων μέσω του Γενετικού Αλγορίθμου, θα πρέπει

$$timestamp - \overline{timestamp}([C]) > \theta_{GA} \tag{6.22}$$

όπου $timestamp$ η τρέχουσα τιμής χρονοσφραγίδας τους συστήματος, $\overline{timestamp}([C])$ ο μέσος όρος των χρονοσφραγίδων δημιουργίας των κανόνων ενός Correct Set και θ_{GA} το κατώφλι χρόνου.

Το κατώφλι αυτό βλέπουμε ότι ξεπερνιέται όταν περάσει χρονικό διάστημα τέτοιο ώστε να υπάρχει ποικιλότητα ως προς το χρόνο δημιουργίας των κανόνων. Η ποικιλότητα αυτή μειώνει το μέσο όρο των χρονοσφραγίδων δημιουργίας των κανόνων στα $[C]$, σε σχέση με τη διαφορά $timestamp([C]) - \theta_{GA}$.

Αρθρωτες Τροποποιησεις

Μέχρις αυτού του σημείου, έχουμε περιγράψει πλήρως τις βασικές λειτουργίες του GMl-ASLCS και τα σημεία διαφοροποίησης του από τον GMl-ASLCS$_0$. Σε αυτή την ενότητα θα εξετάσουμε μερικά περαιτέρω σημεία στα οποία θα μπορούσαμε να διαφοροποιήσουμε τη λειτουργία του GMl-ASLCS, αλλά δεν περιλαμβάνονται στο βασικό ορισμό του. Τα σημεία αυτά θα μελετηθούν στη συνέχεια σε σχέση με αυτόν.

Τροποποίηση της Ενημέρωσης Καταλληλότητας στη Συνιστώσα Ενίσχυσης

Όπως περιγράφηκε στη γραμμή 7 του Αλγ. 6.5, ο GMl-ASLCS υπολογίζει την καταλληλότητα ενός κανόνα απευθείας από το λόγο ορθών κατηγοριοποιήσεων προς τις συνολικές του κατηγοριοποιήσεις, υψωμένο σε μία δύναμη ν. Σαφέστατα, είναι δυνατόν να υπάρξουν ποικίλοι τρόποι υπολογισμού της καταλληλότητας ενός κανόνα, αλλά σε αυτή την εργασία εξετάζουμε μόνο άλλη μία. Αυτή βασίζεται στην τεχνική *Q-learning* [Wil94] που χρησιμοποιείται κατά κόρον στην ενισχυτική μάθηση και προσομοιάζει περισσότερο στο διαμοιρασμό καταλληλότητας που χρησιμοποιεί ο UCS. Έτσι, για δεδομένο κανόνα $rule$, αν t και $t-1$ διακριτές διαδοχικές χρονικές στιγμές στις οποίες ενημερώνεται η καταλληλότητα του $rule$, $rule.tp$, ο αριθμός των ορθών και $rule.msa$ συνολικών κατηγοριοποιήσεών του, και β ο ρυθμός μάθησης, η καταλληλότητα του $rule$ ενημερώνεται ως εξής:

$$fitness(rule, t) = fitness(rule, t-1) + \beta \cdot \left(\left(\frac{rule.tp}{rule.msa} \right)^{\nu} - fitness(rule, t-1) \right) \tag{6.23}$$

Τροποποίηση της Λειτουργίας Διαγραφής

Αν και κάθε σύνολο δεδομένων έχει διαφορετικές ανάγκες ως προς τη διαγραφή κανόνων, το Σχήμα διαγραφής ενός ΜαΣΤ θα πρέπει να ενσωματώνει μία μέθοδο διαγραφής που να μπορεί να είναι εύρωστη για διαφορετικά σύνολα δεδομένων. Εφόσον η μέθοδος διαγραφής του UCS (Εξ. 6.24) αποδεικνύεται αξιόπιστη για προβλήματα μονοκατηγορικής ταξινόμησης και δεδομένης της ανάγκης μας για καλύτερο διαχωρισμό των υπό διαγραφή κανόνων μέσω της ανάθεσης πιθανοτήτων

διαγραφής που βασίζονται στην ύψωση του αριθμού Euler στις παραμέτρους που αποτελούν τα κριτήρια διαγραφής, μπορούμε να ενσωματώσουμε τη μέθοδο διαγραφής του UCS στο πλαίσιο διαγραφής του GMl-ASLCS, αντικαθιστώντας την Εξ. 6.12 στον υπολογισμό της πιθανότητας διαγραφής $P(i)$ ενός κανόνα i, με την:

$$d(i) = \begin{cases} e^{\frac{cs(i) \cdot F_P}{F_{micro}(i)}}, & experience(i) > \theta_{del} \text{ και } F_{micro}(i) < \delta \cdot F_P \\ e^{cs(i)}, & \text{αλλού} \end{cases} \tag{6.24}$$

Εναλλακτικές τιμές των μεταβλητών ω, ϕ

Όπως είδαμε στην Παρ. 5.2.1, οι μεταβλητές ω, ϕ καθορίζουν το ποσό μεταβολής της ακρίβειας ενός κανόνα για κάθε ετικέτα για την οποία αδιαφορεί. Ο GMl-ASLCS ακολουθεί μία ασφαλή προσέγγιση, χρησιμοποιώντας ως $(\omega.\phi) \equiv (0.9, 1)$, ωστόσο, αυτό δεν είναι το μοναδικό ζευγάρι παραμέτρων το οποίο θα οδηγούσε στην ορθή αξιολόγηση κανόνων που περιλαμβάνουν αδιαφορίες στο τμήμα απόφασής τους. Ένα δεύτερο ζεύγος τιμών για τις παραμέτρους ω, ϕ, με το οποίο πειραματιστήκαμε είναι το $(\omega.\phi) \equiv (0, 0)$. Με αυτό τον τρόπο παραβλέπονται οι αδιαφορίες των κανόνων για ετικέτες και ο υπολογισμός της καταλληλότητάς τους βασίζεται μόνο στις σαφείς αποφάσεις που λαμβάνουν, υπέρ ή κατά αυτών. Σε συνδυασμό με τον αποκλεισμό των κανόνων που αδιαφορούν για μία δεδομένη ετικέτα l από το να συμμετέχουν στο αντίστοιχο $[C_l]$ και χαμηλές τιμές πιθανότητας γενίκευσης ετικετών από το τμήμα κάλυψης, $P_{\#L} < 0.1$, αυτή η μέθοδος ενημέρωσης της καταλληλότητας των κανόνων φαίνεται ότι αποτελεί μία ορθή και αξιόπιστη εναλλακτική προσέγγιση στο ζήτημα του προσδιορισμού των δύο αυτών παραμέτρων (βλ. πειραματικά αποτελέσματα στην Παρ. 8.4.3).

Αρχικοποίηση Πληθυσμού μέσω Ομαδοποίησης

Η λειτουργία της κάλυψης, στα πλαίσια ορισμού των ΜαΣΤ, δημιουργεί κανόνες ταξινόμησης ως γενικευμένες εκδόσεις των δειγμάτων του συνόλου εκπαίδευσης D στις περιπτώσεις που το ΜαΣΤ δε διαθέτει κάποιον κανόνα που να ενεργοποιείται για ένα δεδομένο παρεχόμενο δείγμα του D. Ο τελεστής κάλυψης συμπληρώνει το σύνολο κανόνων προοδευτικά και *κατά τη διάρκεια* της εκπαίδευσης, δηλαδή σε στενή αλληλεπίδραση με τις διαδικασίες εξελικτικής αναζήτησης και διαγραφής που χρησιμοποιούνται από τα ΜαΣΤ. Στο [Tzi12], γίνεται η πρώτη αναφορά σε αρχικοποίηση του πληθυσμού κανόνων *πριν τη διαδικασία εκπαίδευσης*, για ΜαΣΤ μονοκατηγορικής ταξινόμησης. Η *βασισμένη στην Ομαδοποίηση Αρχικοποίηση* εφαρμόζεται πριν την εκπαίδευση και συμπληρώνει τον τελεστή κάλυψης, έχοντας ως στόχο την παροχή ικανών αρχικών λύσεων για την εξελικτική συνιστώσα ανακάλυψης κανόνων.

Εν γένει, η μέθοδος αρχικοποίησης μέσω ομαδοποίησης προσπαθεί να εκμεταλλευτεί τις δυνατότητες των αλγορίθμων ομαδοποίησης (clustering) για να παράξει ένα μη-τυχαίο σύνολο κανόνων το οποίο μπορεί να βοηθήσει την εξελικτική διαδι-

κασία, προσδιορίζοντας τη "βάση" των κανόνων που εξελίσσει, ώστε το ΜαΣΤ να εστιάσει αποτελεσματικά στα βέλτιστα σημεία του χώρου αναζήτησης (στο βέλτιστο σύνολο κανόνων για το υπό μελέτη πρόβλημα κατηγοριοποίησης). Διαισθητικά, αυτό το μη-τυχαίο σύνολο αρχικών κανόνων θα πρέπει να βασίζεται στη διαθέσιμη πληροφορία για το πρόβλημα και να παρέχει μία περιεκτική περίληψη της γνώσης που περιέχεται σε αυτό. Στόχος είναι η βελτίωση συνολικά της προβλεπτικής ικανότητας των ΜαΣΤ και η ερμηνευσιμότητα των παραγόμενων μοντέλων.

Η μέθοδος επιλέγει ένα αντιπροσωπευτικό σύνολο σημείων - τα κεντροειδή - από το D και τα μετασχηματίζει σε κανόνες κατάλληλους για την αρχικοποίηση του πληθυσμού του ΜαΣΤ, μέσω της παρακάτω διαδικασίας:

1. Το σύνολο εκπαίδευσης D διαμερίζεται σε N υποσύνολα, όπου N είναι ο αριθμός των διακριτών συνδυασμών τιμών των ετικετών των δειγμάτων του D. Κάθε υποσύνολο $Partition_i$, $1 \leq i \leq N$, περιλαμβάνει όλα τα δείγματα του διακριτού συνδυασμού τιμών των ετικετών i του D:

$$\sum_{i=1}^{N} Partition_i = D \tag{6.25}$$

2. Για κάθε $Partition_i$:

   (α′) Τα δείγματά του ομαδοποιούνται σε $M_i = \lceil \gamma \cdot |Partition_i| \rceil$ συστάδες, όπου $|Partition_i|$ είναι ο αριθμός δειγμάτων του $Partition_i$ και $\gamma \leq 1$ μία οριζόμενη από το χρήστη παράμετρος.

   (β′) Για κάθε συστάδα $cluster_j$, $1 \leq j \leq M_i$ που αναγνωρίστηκε στο βήμα (2α'), βρίσκεται το κεντροειδές της μέσω του αλγορίθμου k-means και βάσει των γνωρισμάτων του δημιουργείται ένας κανόνας του οποίου το τμήμα συνθήκης αποτελεί, εν γένει, μία γενίκευσή του, ανά γνώρισμα, όπως και στην περίπτωση εφαρμογής του τελεστή κάλυψης. Το τμήμα απόφασης του κανόνα τίθεται ίδιο με το διακριτό συνδυασμό τιμών ετικετών που συσχετίζεται με το $Partition_i$.

3. Όλοι οι

$$K = \sum_{i=1}^{N} \sum_{j=1}^{|M_i|} rule_{ij}$$

κανόνες της μορφής

$$rule_{ij} : ruleCondition_{ij} \rightarrow distinctLabelCombination_i$$

που δημιουργήθηκαν από την ομαδοποίηση του συνόλου εκπαίδευσης, συγχωνεύονται για να δημιουργήσουν το σύνολο κανόνων που θα χρησιμοποιηθεί για την αρχικοποίηση της εκπαιδευτικής διαδικασίας.

Συνοψη

Τα βασικά σημεία που παρουσιάστηκαν σε αυτό το κεφάλαιο συνοψίζονται παρακάτω.

I Παρουσιάσαμε τους λόγους για την επινόηση του τελεστή Διασταύρωσης Δύο Τμημάτων που προσιδιάζει περισσότερο στη φύση των προβλημάτων πολυκατηγορικής ταξινόμησης και που αποφεύγει την εναλλαγή διαφορετικών ετικετών από αυτή για την οποία σχηματίστηκε το Correct Set στο οποίο εφαρμόζεται η λειτουργία του Γενετικού Αλγορίθμου (Παρ. 6.2.1).

II Εκθέσαμε τα προβλήματα που επιφέρει η συγκράτηση κανόνων μηδενικής κάλυψης στον πληθυσμό κανόνων των ΜαΣΤ και προτείναμε μία μέθοδο μη εισαγωγής τους στον πληθυσμό, που είναι ανεξάρτητη από τον αριθμό των ετικετών του προβλήματος και συνεπώς μπορεί να εφαρμοστεί ακόμα και σε μονοκατηγορικά ΜαΣΤ (Παρ. 6.2.2).

III Περιγράψαμε την ανανεωμένη διαδικασία εισαγωγής απογόνων στον πληθυσμό κανόνων (Παρ. 6.2.3) μετά τον έλεγχό τους για υπαγωγή (Παρ. 6.2.3).

IV Εισήγαμε μία νέα μέθοδο διαγραφής κανόνων, εξηγήσαμε γιατί είναι προτιμότερο να μην χρησιμοποιείται η προσέγγιση του μεγέθους των Correct Sets στα οποία συμμετέχει ένας κανόνας στα πρώτα του βήματα μέσα στον κύκλο εκπαίδευσης των ΜαΣΤ όσον αφορά στη διαγραφή του και τους λόγους χρήσης του αριθμού Euler ως βάση για την ύψωση σε αυτή των κριτηρίων διαγραφής (Παρ. 6.2.5).

V Διαπιστώσαμε το χαμηλό επίπεδο μέσης κάλυψης των κανόνων του τελικού μοντέλου του GMl-ASLCS$_0$ και προτείναμε την εφαρμογή μίας πρωτότυπης μεθόδου για την αύξησή του (που και αυτή είναι ανεξάρτητη του αριθμού των ετικετών του προβλήματος και συνεπώς μπορεί να βρει εφαρμογή και σε προβλήματα απλής ταξινόμησης), κάνοντας την υπόθεση πως για δεδομένο δείγμα εκπαίδευσης είναι δυνατόν να βρούμε τους συσχετισμούς εκείνους που θα μας επιτρέψουν να αφαιρέσουμε κανόνες που στην ουσία αποτελούν πλεονασμό πληροφορίας (Παρ. 6.3.2).

VI Σχολιάσαμε τη λειτουργία ορισμένων αρθρωτών τμημάτων των ΜαΣΤ και προτείναμε περαιτέρω τροποποίησής τους, βάσει των παραπάνω σχολίων, προς κατευθύνσεις που θα μπορούσαν να εξερευνηθούν και να επεκταθούν στο μέλλον (εν. 6.5).

7

Πειράματα Τεχνητών Συνόλων Δεδομένων

Η πρώτη φάση της πειραματικής διαδικασίας της παρούσας εργασίας χρησιμοποιεί τεχνητά προβλήματα πολυκατηγορικής ταξινόμησης, με δυαδικά γνωρίσματα, με σκοπό μία περισσότερο στοχευμένη αξιολόγηση της ικανότητας ταξινόμησης του GMl-ASLCS, σε ελεγχόμενο περιβάλλον. Τα προβλήματα αυτά πρωτοεισήχθησαν στο [Mil11] και αποτελούν την επέκταση των τυπικών τεχνητών μονοκατηγορικών προβλημάτων που χρησιμοποιούνται για την αξιολόγηση ΜαΣΤ μίας κατηγορίας, στο πεδίο των πολλών ετικετών. Η χρήση των τριών πολυκατηγορικών προβλημάτων $mlposition_N$, $identity_N$ και $adder_N^k$ ενέχει τα πλεονεκτήματα α) της apriori γνώσης της βέλτιστης λύσης, καθιστώντας ευκολότερη τη διαπίστωση σημείων βελτίωσης και β) της συμπαγούς και κατανοητής αναπαράστασης των κανόνων που παράγονται από το ΜαΣΤ.

Περιγραφη των Τεχνητων Συνολων Δεδομενων

Τα τεχνητά σύνολα δεδομένων χρησιμοποιούνται με σκοπό να καλύψουν ένα ευρύ φάσμα πιθανών χαρακτηριστικών και καταστάσεων σε πραγματικά σύνολα δεδομένων και να δοκιμάσουν τις διάφορες ικανότητες των ΜαΣΤ ως προς το βαθμό γενίκευσης, την ακρίβειά τους κ.ο.κ. Στον Πίνακα 7.1 παρουσιάζονται συνοπτικά τα τρία προβλήματα που χρησιμοποιήθηκαν, ενώ αυτά περιγράφονται στις παρακάτω ενότητες.

Προτού αρχίσουμε την αναφορά μας σε αυτά, όμως, πρέπει να αναφερθούμε στην έννοια του *Χάρτη Βέλτιστων Αποφάσεων* (ΧΒΑ), όσο αυτός αφορά στα τεχνητά σύνολα δεδομένων. Ο ΧΒΑ αποτελεί το σύνολο των κανόνων που έχει την ικανότητα να συμπυκνώνει τη γνώση για το σύνολο των δειγμάτων εκπαίδευσης D με πλήρη ορθότητα και που το μέγεθος του είναι μικρότερο από αυτό του D.

Πίνακας 7.1: Συνοπτικά χαρακτηριστικά τεχνητών συνόλων δεδομένων.

	$\|L\|$	LC	LD	Μέγεθος ΧΒΑ	P_{DIST}	P_{maxF}
$mlPosition_N$	N	$(2^N-1)/2^N$	$1/N$	$N+1$	$N/2^N$	0.5
$mlIdentity_N$	N	$N/2$	$1/2$	$2N$	1	$1/2^N$
$adder_N^k$ $(k mod q = 0)$	N	$N/2$	$1/2$	$2^{N-q}+2q$	1	$1/2^N$

Δεδομένης της χρήσης αναπαράστασης με αδιαφορίες, για τα γνωρίσματα και τις ετικέτες, είναι δυνατόν να προσδιοριστούν διαφορετικά σύνολα κανόνων με τις παραπάνω ιδιότητες για ένα δεδομένο σύνολο δεδομένων και συνεπώς, εν γένει, ένας ΧΒΑ δεν αποτελεί απαραίτητα μία ένα-προς-ένα αντιστοιχία με το σύνολο δεδομένων που περιγράφει. Επιπρόσθετα, η ικανότητα ανάπτυξης του ΧΒΑ από ένα ΜαΣΤ είναι μετά βίας ενδεικτική της ικανότητας του για ορθή ταξινόμηση ή γενίκευση: είναι απλώς μία ικανή συνθήκη, αλλά όχι αναγκαία.

Το πρόβλημα $mlPosition_N$

Το πρόβλημα αυτό είναι αντίστοιχο του $position_N$ της απλής κατηγοριοποίησης. Αποτελείται από N δυαδικά γνωρίσματα και N ετικέτες. Κάθε δείγμα έχει ακριβώς μία ετικέτα ενεργοποιημένη (εκτός από έναν κανόνα), και αυτή είναι η ετικέτα που αντιστοιχεί στο πιο σημαντικό ψηφίο (MSB) του δυαδικού αριθμού που σχηματίζουν τα γνωρίσματα. Έτσι, αν

$$b_{N-1} \dots b_1 b_0$$

είναι η αναπαράσταση του δυαδικού αριθμού που αντιστοιχεί στα γνωρίσματα, τότε κάθε ετικέτα l_i με $i = 0..N-1$ παίρνει την τιμή που δίνεται από την ακόλουθη έκφραση Bool:

$$l_i = b_i \cdot (\bar{b}_{i+1} \cdot \bar{b}_{i+2} \dots \bar{b}_{N-1})$$

Ένα παράδειγμα, για $N = 4$, φαίνεται στον Πίνακα 7.2.

Πίνακας 7.2: Δείγματα Τεχνητού Συνόλου $mlPosition_4$.

$0000 \to 0000$	$1000 \to 1000$
$0001 \to 0001$	$1001 \to 1000$
$0010 \to 0010$	$1010 \to 1000$
$0011 \to 0010$	$1011 \to 1000$
$0100 \to 0100$	$1100 \to 1000$
$0101 \to 0100$	$1101 \to 1000$
$0110 \to 0100$	$1110 \to 1000$
$0111 \to 0100$	$1111 \to 1000$

Όπως είναι εμφανές, οι ετικέτες είναι πλήρως εξαρτημένες μεταξύ τους, καθώς είναι αμοιβαία αποκλειόμενες. Υπάρχει μεγάλη ανισορροπία μεταξύ των διάφορων συνδυασμών ετικετών, καθώς η ετικέτα l_0 ενεργοποιείται ακριβώς μία φορά, ενώ η ετικέτα l_{N-1} 2^{N-1} φορές. Η βέλτιστη χαρτογράφηση του προβλήματος εδώ είναι κάτι το αντικειμενικό: αποτελείται από $N+1$ κανόνες, με διαφορετικούς βαθμούς γενίκευσης σε κάθε τμήμα συνθήκης. Για $N=4$ ο Βέλτιστος Χάρτης Αποφάσεων φαίνεται στον Πίνακα 7.3.

Πίνακας 7.3: Βέλτιστος Χάρτης Αποφάσεων του συνόλου $mlPosition_4$.

$0000 \rightarrow 0000$
$0001 \rightarrow 0001$
$001\# \rightarrow 0010$
$01\#\# \rightarrow 0100$
$1\#\#\# \rightarrow 1000$

Το πρόβλημα $mlIdentity_N$

Το πρόβλημα $mlIdentity_N$ αποτελεί την επέκταση του προβλήματος $decoder_N$ της απλής ταξινόμησης στον πολυ-ετικετικό χώρο. Κάθε δείγμα αποτελείται από N δυαδικά γνωρίσματα και N ετικέτες. Αν b_i το i χαρακτηριστικό και l_i η i ετικέτα, για κάθε δείγμα ισχύει η σχέση

$$b_i = l_i, \forall i \in [0, N-1]$$

Ένα παράδειγμα για $N=4$ φαίνεται στον Πίνακα 7.4.

Πίνακας 7.4: Δείγματα Τεχνητού Συνόλου mlIdentity_4.

$0000 \rightarrow 0000$	$1000 \rightarrow 1000$
$0001 \rightarrow 0001$	$1001 \rightarrow 1001$
$0010 \rightarrow 0010$	$1010 \rightarrow 1010$
$0011 \rightarrow 0011$	$1011 \rightarrow 1011$
$0100 \rightarrow 0100$	$1100 \rightarrow 1100$
$0101 \rightarrow 0101$	$1101 \rightarrow 1101$
$0110 \rightarrow 0110$	$1110 \rightarrow 1110$
$0111 \rightarrow 0111$	$1111 \rightarrow 1111$

Στα προβλήματα της οικογένειας $mlIdentity_N$ οι ετικέτες είναι πλήρως ανεξάρτητες μεταξύ τους, καθώς η γνώση για την ενεργοποίηση μίας δεδομένης ετικέτας δε σημαίνει την εξαγωγή κάποιου συμπεράσματος για τις τιμές των υπόλοιπων ετικετών. Κάθε διαφορετική ετικέτα εμφανίζεται στο σύνολο δεδομένων ακριβώς 2^{N-1} φορές, κάθε συνδυασμός ετικετών ακριβώς μία φορά και, συνεπώς, δεν υπάρχει

καμία ανισορροπία μεταξύ τους. Σε αυτό το πρόβλημα, θα μπορούσαν να υπάρχουν διαφορετικές βέλτιστες χαρτογραφήσεις. Η μορφή του BXA που επιλέγουμε ως σημείο αναφοράς για τα πειράματα που ακολουθούν φαίνεται στον Πίνακα 7.5 για $N = 4$.

Πίνακας 7.5: Βέλτιστος Χάρτης Αποφάσεων του συνόλου $mlIdentity_4$.

1### → 1###
#1## → #1##
##1# → ##1#
###1 → ###1
0### → 0###
#0## → #0##
##0# → ##0#
###0 → ###0

Το παραπάνω σύνολο κανόνων παρατηρούμε ότι περιλαμβάνει έντονη χρήση αδιαφοριών στις ετικέτες, σε αντίθεση με το πρόβλημα $mlPosition_N$ όπου υπάρχει πλήρης απουσία αδιαφοριών. Δεδομένης της έκπτωσης στην καταλληλότητα (Εξ. 6.14 και 6.15) που υφίστανται οι κανόνες που περιλαμβάνουν αδιαφορίες στο τμήμα της απόφασής τους, αναμένουμε ότι η εύρεση του παραπάνω συνόλου κανόνων θα γίνει δυσχερέστερη έως αδύνατη για τον GMl-ASLCS σε σχέση με τον GMl-ASLCS$_0$[1], λόγω του μεγάλου ποσοστού ύπαρξης αδιαφοριών στον XBA σε σχέση με τον αριθμό των ετικετών.

Το πρόβλημα $adder_N^k$

Το πρόβλημα $adder_N^k$ αντιστοιχεί στο πρόβλημα $parity$ της απλής κατηγοριοποίησης. Κάθε δείγμα προκύπτει από τη πρόσθεση του αριθμού k στο δυαδικό αριθμό που αντιστοιχεί στα γνωρίσματα, δηλαδή

$$(l_{N-1}l_{N-2}\dots l_0)_2 \leftarrow (b_{N-1}b_{N-2}\dots b_0)_2 + (k)_2$$

Ο XBA εξαρτάται άμεσα από τον αριθμό k, ενώ η αναπαράσταση ετικετών με αδιαφορίες, ανά περίπτωση, ίσως να μη μπορεί να συμπιέσει το μέγεθός του. Ένα παράδειγμα τέτοιου συνόλου, για $N = 4$ και $k = 3$, φαίνεται στον Πίνακα 7.6.

Ο XBA του συνόλου δεδομένων $adder_4^3$ ταυτίζεται με τα στοιχεία του συνόλου. Προφανώς, λόγω και της αναπαράστασης ετικετών με χρήση αδιαφοριών, είναι δυνατόν να βρεθούν κανόνες που δεν ανήκουν στον παραπάνω χάρτη, αφού για να καλυφθούν όλες τις περιπτώσεις χρειάζονται περισσότεροι κανόνες. Η παραπάνω ιδιότητα των $adder_N^k$ αναμένεται να δυσκολέψει τον GMl-ASLCS. Ενδεικτικά, κάποιοι από αυτούς τους κανόνες φαίνονται στον Πίνακα 7.7.

[1]Θυμηθείτε ότι ο GMl-ASLCS$_0$ χειρίζεται ισοδύναμα, ως προς τη μεταβολή της ακρίβειας των κανόνων, την ύπαρξη αδιαφορίας και την πλήρη συμφωνία με μία ετικέτα (Εξ. 5.2).

Πίνακας 7.6: Δείγματα Τεχνητού Συνόλου $adder_4^3$.

$0000 \rightarrow 0011$	$1000 \rightarrow 1011$
$0001 \rightarrow 0100$	$1001 \rightarrow 1100$
$0010 \rightarrow 0101$	$1010 \rightarrow 1101$
$0011 \rightarrow 0110$	$1011 \rightarrow 1110$
$0100 \rightarrow 0111$	$1100 \rightarrow 1111$
$0101 \rightarrow 1000$	$1101 \rightarrow 0000$
$0110 \rightarrow 1001$	$1110 \rightarrow 0001$
$0111 \rightarrow 1010$	$1111 \rightarrow 0010$

Πίνακας 7.7: Μη βέλτιστοι γενικευμένοι κανόνες τεχνητού συνόλου $adder_4^3$.

$\#\#\#0 \rightarrow \#\#\#1$	$\#1\#1 \rightarrow \#0\#\#$
$\#\#\#1 \rightarrow \#\#\#0$	$\#11\# \rightarrow \#0\#\#$
$\#\#10 \rightarrow \#\#01$	$\#100 \rightarrow \#0\#\#$
$\#\#11 \rightarrow \#\#10$	$10\#\# \rightarrow 1\#\#\#$
$\#\#00 \rightarrow \#\#11$	$1\#0\# \rightarrow 1\#\#\#$
$\#\#01 \rightarrow \#\#00$	$01\#1 \rightarrow 1\#\#\#$
$\#000 \rightarrow \#0\#\#$	$011\# \rightarrow 1\#\#\#$
$\#0\#1 \rightarrow \#1\#\#$	$00\#\# \rightarrow 0\#\#\#$
$\#01\# \rightarrow \#1\#\#$	$0\#00 \rightarrow 0\#\#\#$

Για διαφορετικές τιμές του k, και ειδικά όταν αυτό είναι πολλαπλάσιο του 2, μπορεί να προκύψουν πιο συμπαγείς χάρτες, δημιουργώντας ένα πρόβλημα "ενδιάμεσο" των $mlPosition_N$ και $mlIdentity_N$. Έτσι, για παράδειγμα, για το πρόβλημα $adder_4^4$, ο ΧΒΑ φαίνεται στον Πίνακα 7.8.

Πίνακας 7.8: Χάρτης Βέλτιστων Αποφάσεων του συνόλου $adder_4^4$.

$\#\#\#0 \rightarrow \#\#\#1$	$00\#\# \rightarrow 01\#\#$
$\#\#\#1 \rightarrow \#\#\#0$	$01\#\# \rightarrow 10\#\#$
$\#\#1\# \rightarrow \#\#1\#$	$10\#\# \rightarrow 11\#\#$
$\#\#0\# \rightarrow \#\#0\#$	$11\#\# \rightarrow 00\#\#$

Δεδομένης, και εδώ, της ισχυρής παρουσίας αδιαφοριών στο τμήμα απόφασης των κανόνων των δύο παραπάνω συνόλων, αναμένουμε μικρή εκπροσώπησή τους στον τελικό πληθυσμό που εξελίσσει ο GMl-ASLCS.

Παραμετροι Πειραματων και Αρχικες Παρατηρησεις

Τα παραπάνω τεχνητά σύνολα δεδομένων μπορούν να χρησιμοποιηθούν για να διακρίνουν τα πολυκατηγορικά ΜαΣΤ ως προς την ικανότητά τους για ορθή ταξινόμηση και την ανάπτυξη του ΧΒΑ. Συγκεκριμένα, οι μετρικές αξιολόγησης που θα χρησιμοποιηθούν στα παρακάτω πειράματα είναι αυτές της Ακρίβειας (Accuracy), της Ακριβούς Ορθότητας (Exact Match), αλλά και το ποσοστό του ΧΒΑ που απεικονίστηκε στο σύνολο κανόνων του πληθυσμού στη διάρκεια της εκπαίδευσης. Ως σύνολο εκπαίδευσης και σύνολο ελέγχου σε κάθε πείραμα χρησιμοποιείται πάντοτε το ίδιο, ολικό, σύνολο δεδομένων. Για κάθε πολυκατηγορικό πρόβλημα πραγματοποιήθηκαν 10 εκπαιδεύσεις, βάσει των οποίων, στα διαγράμματα των ενοτήτων που ακολουθούν εμφανίζεται τόσο ο μέσος όρος, όσο το εύρος τιμών των παραπάνω μετρούμενων δεικτών.

Οι παράμετροι που χρησιμοποιήθηκαν είναι οι εξής: $\mu = 0.04$, $\chi = 0.8$, $\theta_{GA} = 300$, $\theta_{del} = 20$, $P_{\#A} = P_{\#L} = 0.33$, $\beta = 0.2$, $\nu = 10$, $|P| = 1000$, ενώ ο αριθμός των επαναλήψεων εκπαίδευσης ήταν $|I| = 1000 \cdot |D|$ για κάθε σύνολο δεδομένων D. Το διάστημα ενημέρωσης τέθηκε στις $0.1 \cdot |I|$ επαναλήψεις. Η ταξινόμηση που φαίνεται στα διαγράμματα που ακολουθούν έγινε με χρήση της στρατηγικής απόφασης των βέλτιστων κανόνων αντί της ψηφοφορίας, η οποία προϋποθέτει κάποια μέθοδο επιλογής κατωφλίου. Σε όλα τα πειράματα που περιγράφονται στις παρακάτω ενότητες, οι παράμετροι διατηρούνται σταθερές, ώστε να διευκολυνθεί η σύγκριση των αποτελεσμάτων.

Αποτιμηση των Τροποποιησεων του GMl-ASLCS$_0$

Σε αυτή την ενότητα θα αποτιμήσουμε την επίδραση των τεσσάρων βασικών τροποποιήσεων που επιφέραμε στον GMl-ASLCS$_0$, με βάση τις μετρικές αξιολόγησης που προαναφέραμε, πάνω στα τεχνητά σύνολα $mlposition_N$, $identity_N$ και $adder_N^k$, ως προς το αρχικό, βασικό σύστημα του GMl-ASLCS$_0$. Οι τέσσερις αυτές τροποποιήσεις αφορούν

1. στη λειτουργία διαγραφής (Παρ. 6.2.5)
2. στον τελεστή διασταύρωσης (Παρ. 6.2.1)
3. στη διεύρυνση της μέσης κάλυψης δειγμάτων από τους κανόνες του πληθυσμού μέσω της διαγραφής κανόνων από τα Match Sets (Παρ. 6.3.2), και
4. στην έκπτωση που υφίσταται η ακρίβεια κανόνων που αδιαφορούν για ετικέτες (Παρ. 6.5), με παραμέτρους $(\omega, \phi) \equiv (0.9, 1)$.

Οι αλγόριθμοι που προκύπτουν προσθέτοντας στον GMl-ASLCS$_0$ κάθε μία από τις παραπάνω τροποποιήσεις ονοματοδοτούνται ως GMl-ASLCS$_{0D}$, GMl-ASLCS$_{0GA}$, GMl-ASLCS$_{0M}$ και GMl-ASLCS$_{0C}$, αντίστοιχα.

Η επίδραση της λειτουργίας διαγραφής κανόνων μηδενικής κάλυψης δε μελετήθηκε ως πλεονασματική: όλα τα τεχνητά σύνολα δεδομένων είναι πλήρη δειγμάτων, συνεπώς όλοι οι κανόνες που δημιουργούνται καλύπτουν κάποιο μέρος του προβλήματος.

Ο GMl-ASLCS$_0$ ως σημείο αναφοράς

Σε αυτή την παράγραφο παραθέτουμε την πορεία των μετρικών αξιολόγησης της συμπεριφοράς του GMl-ASLCS$_0$ στο πέρασμα των επαναλήψεων εκπαίδευσης για τρεις οικογένειες πολυκατηγορικών προβλημάτων, ως ένα σημείο αναφοράς για τη μετέπειτα σύγκριση των τεσσάρων τροποποιήσεων που προαναφέραμε. Στα γραφήματα των Σχημάτων 7.1, 7.2, 7.3 και 7.4 παρατίθεται η εξέλιξη των μετρούμενων δεικτών αξιολόγησης για τα τέσσερα πολυκατηγορικά προβλήματα $mlposition_7$, $identity_7$, $adder_7^3$ και $adder_7^{24}$, αντίστοιχα. Τα γραφήματα πάρθηκαν απευθείας από το [Mil11].

Όσον αφορά στην οικογένεια προβλημάτων $adder_N^k$, μελετήθηκαν δύο είδη: το $adder_7^3$ και το $adder_7^{24}$. Και τα δύο προβλήματα διαθέτουν την ίδια πολυκατηγορική πληθικότητα. Όμως, το $adder_7^3$, στο βέλτιστο χάρτη του έχει όλους τους κανόνες ειδικευμένους, ενώ το $adder_7^{24}$ έχει έναν ανάμεικτο χάρτη από πλευράς γενίκευσης, με γενικούς κανόνες στα δυαδικά ψηφία χαμηλής σημασίας και ειδικούς στα υψηλής. Σε αυτά τα δύο προβλήματα δε διερευνήθηκε η ικανότητα προσέγγισης του βέλτιστου χάρτη, αφού εξετάζονται οι δύο ακραίες περιπτώσεις ανάπτυξής του με τα προβλήματα $mlPosition_7$ και $mlIdentity_7$, αλλά μόνο η προβλεπτική ικανότητα.

Οι τροποποιημένοι αλγόριθμοι GMl-ASLCS$_{0*}$ στο πρόβλημα $mlPosition_7$

Τα Σχήματα 7.5, 7.6, 7.7 και 7.8 παρουσιάζουν την εξέλιξη των μετρικών της Ακρίβειας, της Ακριβούς Ορθότητας και του ποσοστού κάλυψης του XBA για το σύνολο δεδομένων $mlPosition_7$, των ΜαΣΤ GMl-ASLCS$_{0D}$, GMl-ASLCS$_{0GA}$, GMl-ASLCS$_{0M}$ και GMl-ASLCS$_{0C}$, αντίστοιχα.

Οι GMl-ASLCS$_{0D}$ και GMl-ASLCS$_{0GA}$, αν και δεν καταφέρνουν να επιτύχουν την πλήρη σύγκλιση τόσο στο XBA, όσο και για τη μετρική της Ακρίβειας, επιπρόσθετα βελτιώνουν σε διαφορετικό βαθμό το ρυθμό και το χρόνο της σύγκλισης, σε σχέση με τον GMl-ASLCS$_0$, ενώ μειώνουν σε μικρό βαθμό το εύρος για τις μετρικές της Ακρίβειας, και σε μεγάλο, το εύρος ανάπτυξης του Βέλτιστου Χάρτη. Παρατηρούμε, δηλαδή μία βελτίωση ως προς την ικανότητα γενίκευσης, κάτι που απαιτείται από τη φύση του προβλήματος και της λύσης του, με ταυτόχρονη συγκράτηση της Ακρίβειας του τελικού μοντέλου. Λόγω της μικρής πολυκατηγορικής πυκνότητας του συνόλου $mlPosition_7$, η συμπεριφορά της Διασταύρωσης Δύο Τμημάτων προσεγγίζει αυτήν της Διασταύρωσης Ενός σημείου, καθώς δεν παρατηρούμε κάποια ραγδαία βελτίωση στο ρυθμό σύγκλισης του αλγορίθμου GMl-ASLCS$_{0GA}$.

Ο GMl-ASLCS$_{0M}$, παρ' όλη την ικανότητά του για γενίκευση, δε φαίνεται να μπορεί να αναπτύξει το σαφή βέλτιστο χάρτη του $mlPosition_7$ καλύτερα από τον GMl-ASLCS$_0$, και ακόμα χειρότερα δεν μπορεί να κρατήσει τους κανόνες που έχει βρει και που ανήκουν σε αυτόν, όπως φαίνεται από τις ταλαντώσεις της μέσης τιμής του ποσοστού εύρεσης του βέλτιστου χάρτη, αλλά και το εύρος των τιμών του. Ένα μέρος της ευθύνης, βέβαια, φέρει και το γεγονός ότι δεν τιμωρούνται οι κανόνες που φέρουν αδιαφορίες στο τμήμα της απόφασής τους, με αποτέλεσμα το σύστημα να μη μπορεί να διακρίνει αποτελεσματικά ανάμεσα σε έναν κανόνα του βέλτιστου χάρτη και έναν παρόμοιο, αλλά με αδιαφορίες στο τμήμα της απόφασής του, από τη στιγμή μάλιστα που ο βέλτιστος χάρτης περιλαμβάνει κανόνες από

τους οποίους κανένας δεν αδιαφορεί για ετικέτες. Η ταλάντωση του μέσου όρου επίτευξης του βέλτιστου χάρτη και το γεγονός ότι στις τελευταίες επαναλήψεις ο αλγόριθμος αποκλίνει ελάχιστα από τις παραπάνω λύσεις, μάς κάνει να υποθέσουμε ότι η διαγραφή κανόνων από τα Match Sets, ως σχετικιστική λειτουργία, ίσως λειτουργεί αποσταθεροποιητικά για τους κανόνες που διατηρεί το σύστημα. Επομένως, μελλοντικά, θα έπρεπε να εξεταστεί η εφαρμογή ορίων καταλληλότητας για τη διαγραφή των κανόνων, ιδιαίτερα εάν δεν εφαρμόζεται έκπτωση της ακρίβειας για τους κανόνες με αδιαφορίες στις ετικέτες τους.

Τέλος, ο GMl-ASLCS$_{0C}$ καταφέρνει να συγκλίνει ταχύτερα από όλους τους τροποποιημένους GMl-ASLCS$_{0*}$ στο ανώφλι της Ακρίβειας και της Ακριβούς Ορθότητας για το πρόβλημα $mlPosition_7$. Η έκπτωση που επιφέρει στην καταλληλότητα των κανόνων με αδιαφορίες στις ετικέτες τους ($\phi = 0.9$ αντί για $\phi = 1$ για κάθε ετικέτα για την οποία αδιαφορεί ένας κανόνας), σε συνδυασμό με τη χαμηλή πολυκατηγορική πληθικότητα του συνόλου (εν προκειμένω την κατηγοριοποίηση των δειγμάτων το πολύ σε μία ετικέτα), και το γεγονός πως οι κανόνες του XBA δεν περιλαμβάνουν αδιαφορίες στις αποφάσεις τους, βοηθάει το σύστημα να αναπτύξει κανόνες βέλτιστα γενικούς, με τη μέγιστη καταλληλότητα, αυξάνοντας και το ποσοστό εύρεσης των κανόνων του Βέλτιστου Χάρτη. Η αρνητική επίδραση του Γενετικού Αλγορίθμου με Διασταύρωση Ενός Σημείου φαίνεται ότι αποσβένεται λόγω της χαμηλής πολυκατηγορικής πληθικότητας και του γεγονότος ότι, πλέον, οι κανόνες που συμμετέχουν στα Correct Sets, από όπου επιλέγονται κανόνες προς αναπαραγωγή, διαθέτουν σε πολύ μικρότερο βαθμό αδιαφορίες στις ετικέτες τους σε σχέση με τον GMl-ASLCS$_0$. Συνεπώς, λόγω και της ανεξαρτησίας των ετικετών του συνόλου, οι αποφάσεις των κανόνων που συμμετέχουν στα διάφορα $[C]$ είναι σε μεγαλύτερο βαθμό ίδιες από ότι σε αυτά που σχηματίζονται στον GMl-ASLCS$_0$.

Οι τροποποιημένοι αλγόριθμοι GMl-ASLCS$_{0*}$ στο πρόβλημα $mlIdentity_7$

Τα Σχήματα 7.9, 7.10, 7.11 και 7.12 παρουσιάζουν την εξέλιξη των μετρικών της Ακρίβειας, της Ακριβούς Ορθότητας και το ποσοστό κάλυψης του XBA για το σύνολο δεδομένων $mlIdentity_7$, των ΜαΣΤ GMl-ASLCS$_{0D}$, GMl-ASLCS$_{0GA}$, GMl-ASLCS$_{0M}$ και GMl-ASLCS$_{0C}$, αντίστοιχα.

Οι GMl-ASLCS$_{0D}$ και GMl-ASLCS$_{0GA}$ κινούνται στα ίδια πλαίσια με τον GMl-ASLCS$_0$ ως προς τις τελικες τιμές των μετρικών αξιολόγησης, αλλά καταφέρνουν να συγκλίνουν γρηγορότερα, μειώνοντας ταυτόχρονα το εύρος τους. Όσον αφορά στην εύρεση του Βέλτιστου Χάρτη, παρατηρούμε πως και εδώ κινούνται στα βήματα του GMl-ASLCS$_0$, αλλά μέσω του μπορούμε να βγάλουμε το συμπέρασμα πως καταφέρνουν να γενικεύσουν καλύτερα από τον GMl-ASLCS$_0$ (αλλά χειρότερα από όσο στο σύνολο $mlPosition_7$, ίσως λόγω και της μεγαλύτερης πολυκατηγορικής πληθικότητας του συνόλου $mlIdentity_7$). Συνεπώς, και εδώ οι τροποποιήσεις της λειτουργίας διαγραφής και του τελεστή διασταύρωσης επιτυγχάνουν να διατηρήσουν την Ακρίβεια του τελικού μοντέλου που αναπτύσσουν, με ταυτόχρονη αύξηση του αριθμού δειγμάτων που καλύπτουν οι κανόνες τους.

Ο GMl-ASLCS$_{0M}$, εδώ, καταφέρνει να κάνει παρατηρήσιμη τη λειτουργία για την οποία επινοήσαμε τη διαγραφή κανόνων από τα Match Sets. Εκμεταλλευόμενος την ισοτροπία διαχείρισης των αδιαφοριών στο τμήμα της απόφασης των κανόνων

από τη συνιστώσα ενημέρωσης και την ύπαρξη πληθώρας αδιαφοριών για ετικέτες από τους κανόνες του βέλτιστου χάρτη για το σύνολο $mlIdentity_7$, καταφέρνει να αυξήσει τη μέση κάλυψη των κανόνων του πληθυμού, καταλήγοντας σε μία αύξηση κατά 10% του ποσοστού εύρεσης του βέλτιστου χάρτη, διατηρώντας ταυτόχρονα τις μετρικές της Ακρίβειας και της Ακριβούς Ορθότητας στα ίδια επίπεδα με αυτά του GMl-ASLCS$_0$ για το ίδιο πρόβλημα.

Λόγω της ισχυρής παρουσίας αδιαφοριών στις ετικέτες των κανόνων του Χάρτη Βέλτιστων Αποφάσεων και της τιμωρίας κανόνων που αδιαφορούν για ετικέτες μέσω της έκπτωσης της καταλληλότητας τους, ο GMl-ASLCS$_{0C}$ αποτυγχάνει πλήρως να αναπτύξει το Βέλτιστο Χάρτη. Ωστόσο, η ανάπτυξή του, όπως προαναφέραμε, είναι μία ικανή και όχι αναγκαία συνθήκη, και για αυτό δεν μπορούμε να συμπεράνουμε κάτι παραπάνω για τη συμπεριφορά του GMl-ASLCS$_{0C}$. Από τα γραφήματα της Ακρίβειας και της Ακριβούς Ορθότητας, όμως, παρατηρούμε την αδυναμία του GMl-ASLCS$_{0C}$ να εμφανίσει μία αύξουσα πορεία ως προς αυτές και, συνεπώς, την αδυναμία εύρεσης νέων, βελτιωμένων συνολικά λύσεων στο πολυκατηγορικό πρόβλημα. Η αδυναμία αυτή πηγάζει κυρίως από τη Διασταύρωση Ενός Σημείου που χρησιμοποιεί ο Γενετικός Αλγόριθμος του GMl-ASLCS$_0$ και ενισχύεται από την υψηλή κατηγορική πληθικότητα του συνόλου δεδομένων $mlIdentity_7$.

Οι τροποποιημένοι αλγόριθμοι GMl-ASLCS$_{0*}$ στο πρόβλημα $adder_7^3$

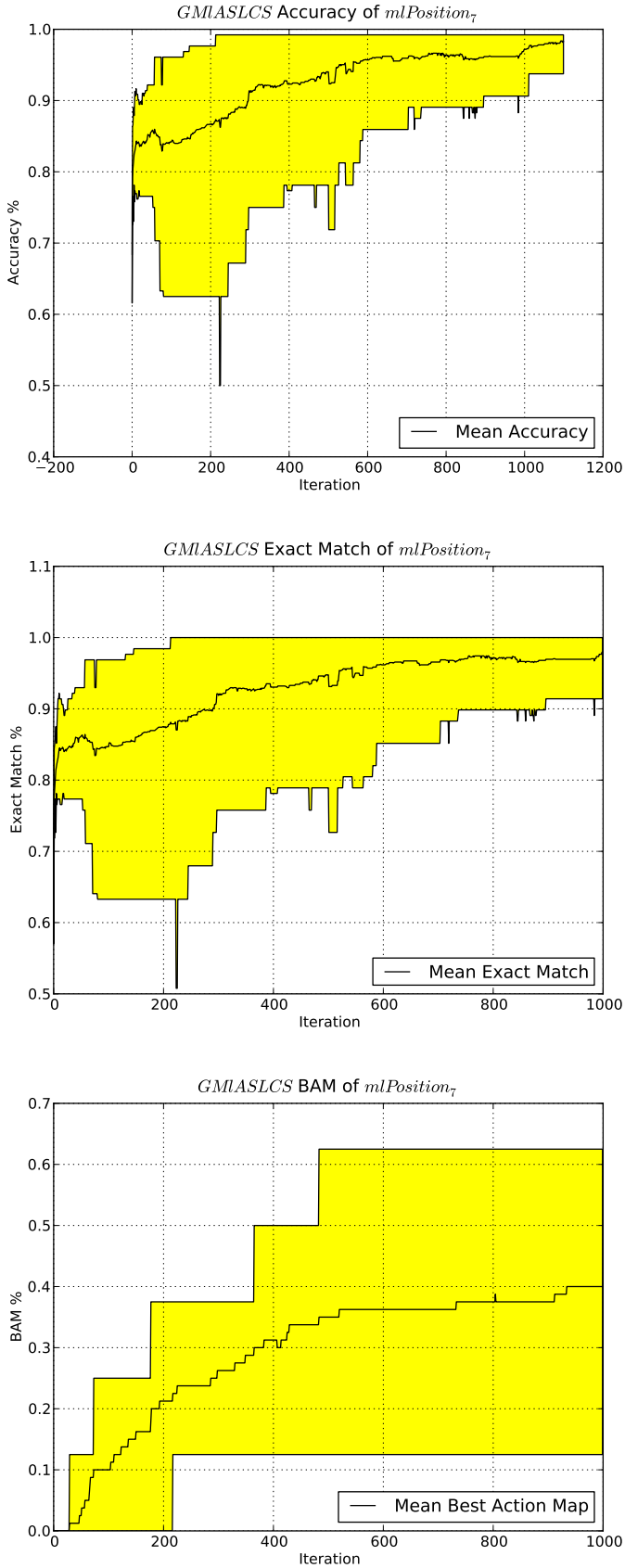
Τα Σχήματα 7.13, 7.14, 7.15 και 7.16 παρουσιάζουν την εξέλιξη των μετρικών της Ακρίβειας και της Ακριβούς Ορθότητας για το σύνολο δεδομένων $adder_7^3$, των ΜαΣΤ GMl-ASLCS$_{0D}$, GMl-ASLCS$_{0GA}$, GMl-ASLCS$_{0M}$ και GMl-ASLCS$_{0C}$, αντίστοιχα.

Οι GMl-ASLCS$_{0D}$ και GMl-ASLCS$_{0GA}$ καταφέρνουν και εδώ να βρουν γρηγορότερα το τελικό σύνολο λύσεων (δηλαδή τους κανόνες του ΧΒΑ), με ταυτόχρονη αύξηση της τελικής Ακριβούς Ορθότητας, και αύξηση της Ακρίβειας για τον GMl-ASLCS$_{0GA}$. Ωστόσο, ο GMl-ASLCS$_{0GA}$ φαίνεται ότι είναι εξαιρετικά συνεπής, καθώς μειώνει αισθητά το εύρος των μετρικών αξιολόγησης, περισσότερο από τον GMl-ASLCS$_{0D}$. Ένα σημείο ανησυχίας παρατηρείται στο διάστημα των τελευταίων επαναλήψεων, όπου το εύρος της Ακρίβειας και ακόμη περισσότερο της Ακριβούς Ορθότητας φαίνεται ότι αυξάνεται, φανερώνοντας πιθανά συμπτώματα υπερεκπαίδευσης.

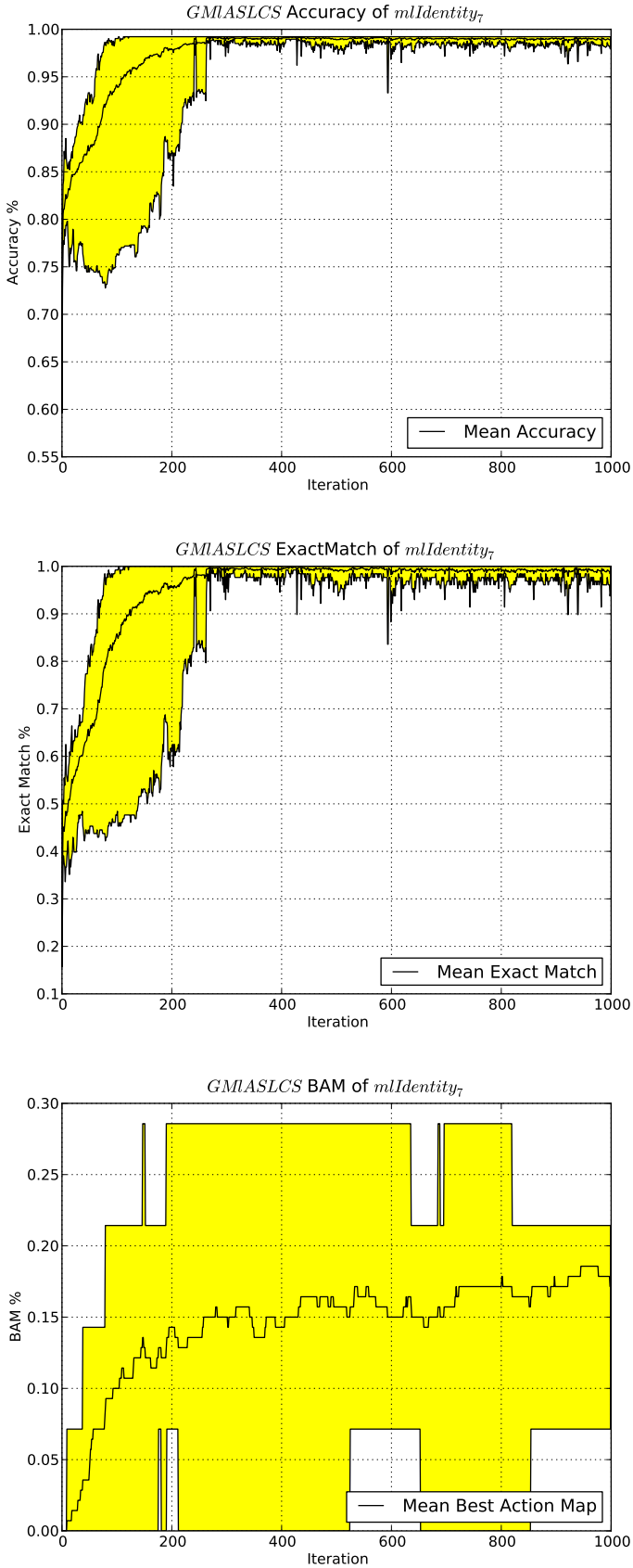
Ο GMl-ASLCS$_{0M}$ εμφανίζει ελαφρώς αυξημένα επίπεδα τιμών Ακρίβειας και σημαντικά αυξημένα επίπεδα Ακριβούς Ορθότητας, ενώ καταφέρνει να μειώσει τις ελάχιστες τιμές τους, ιδιαίτερα στις τελευταίες επαναλήψεις, σε αντίθεση με τον GMl-ASLCS$_0$. Το ποσοστό κάλυψης, αν και δε φαίνεται στα διαγράμματα, αυξάνει και αυτό και μάλιστα διπλασιάζεται.

Ο GMl-ASLCS$_{0C}$ εμφανίζει και εδώ το πρόβλημα της στασιμότητας των λύσεων που αναπτύσσει, όπως και στο πρόβλημα $mlIdentity_7$. Επιπλέον, και εδώ εντοπίζουμε την ίδια συμπεριφορά του Γενετικού Αλγορίθμου, καθώς τα δύο σύνολα διαθέτουν την ίδια τιμή πολυκατηγορικής πληθικότητας.

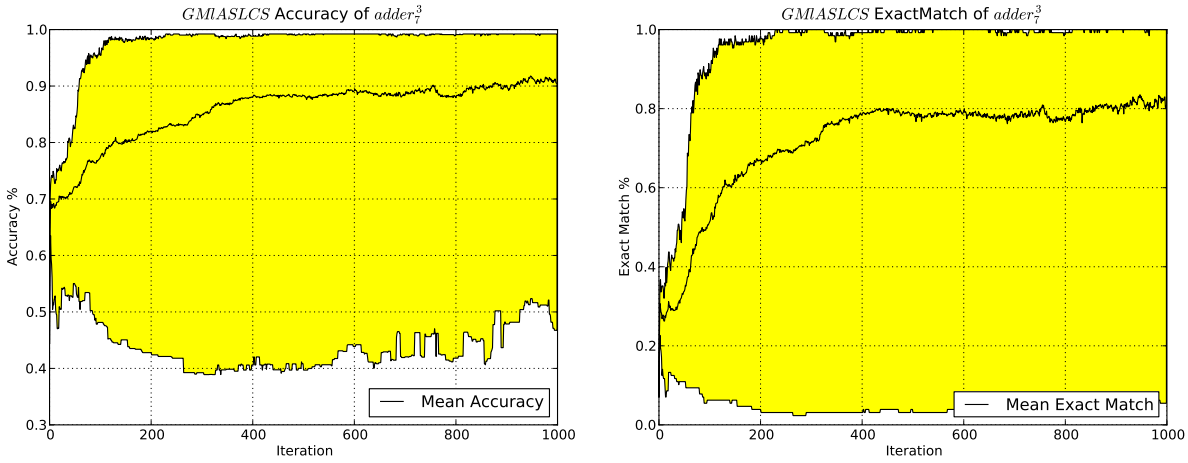
Σχήμα 7.1: Διαγράμματα χαρτογράφησης $mlPosition_7$ του GMl-ASLCS$_0$.



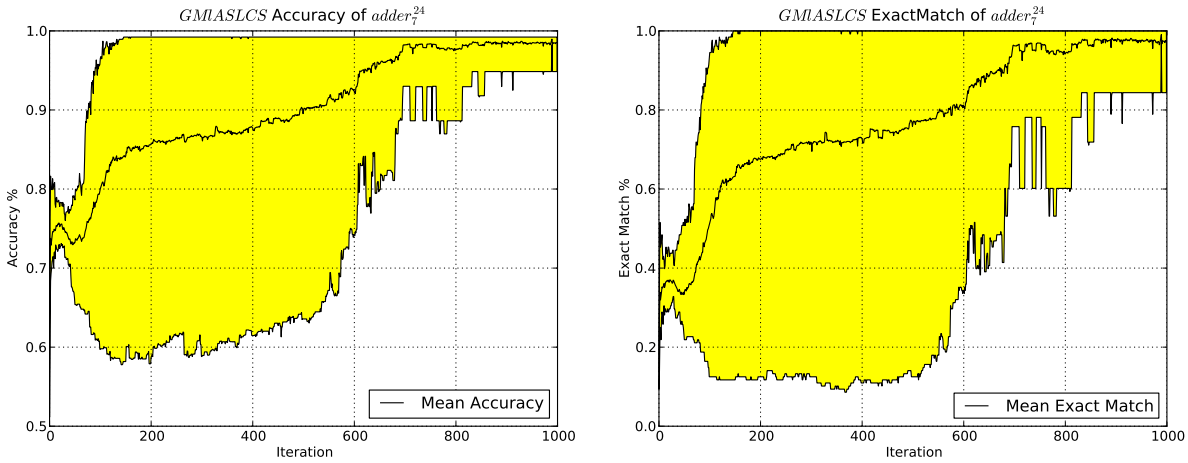
Σχήμα 7.2: Διαγράμματα χαρτογράφησης $mlIdentity_7$ του GMl-ASLCS$_0$.



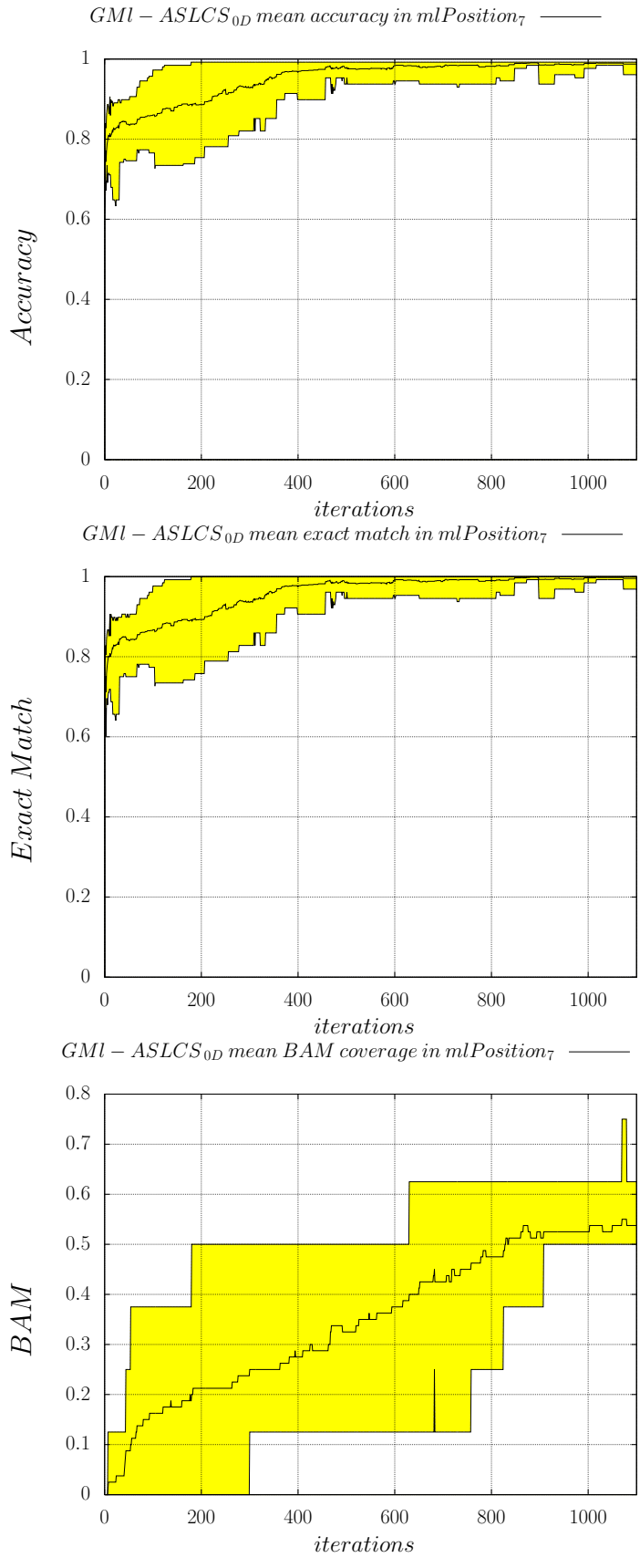
Σχήμα 7.3: Διαγράμματα χαρτογράφησης $adder_7^3$ του GMl-ASLCS$_0$.



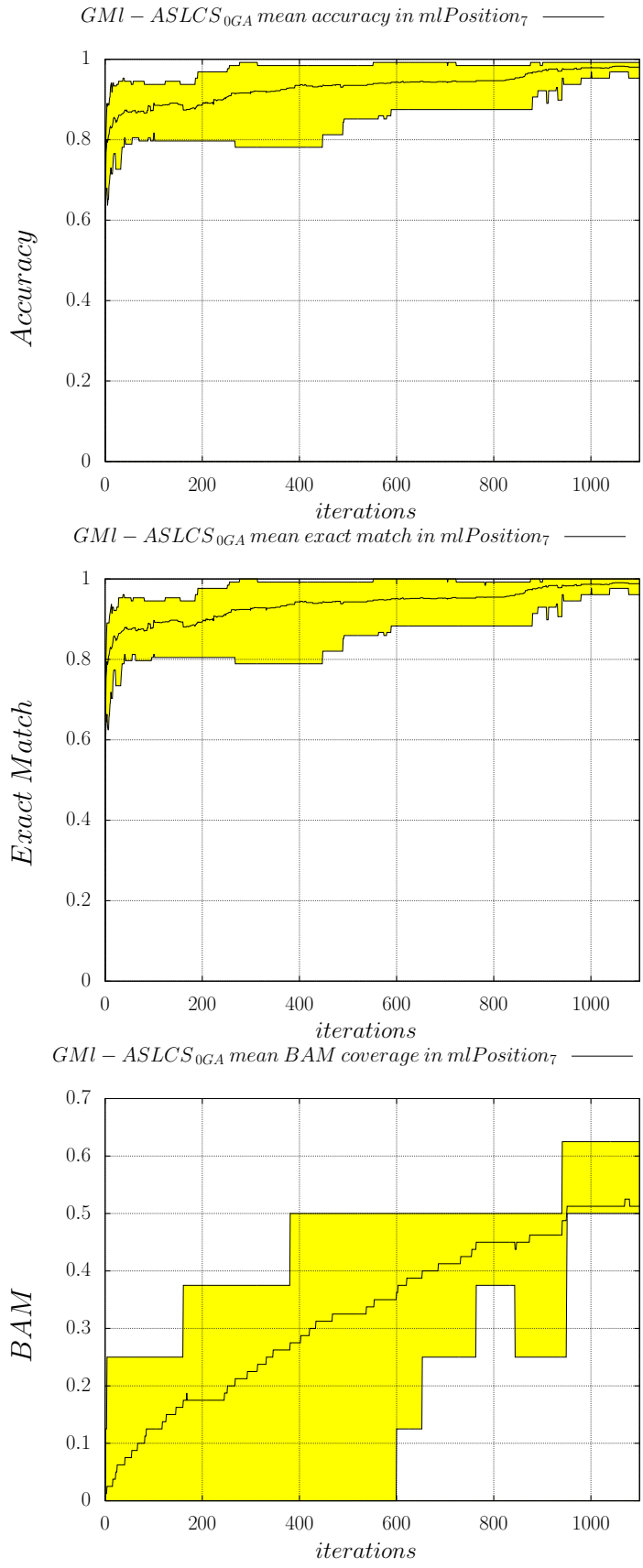
Σχήμα 7.4: Διαγράμματα χαρτογράφησης $adder_7^{24}$ του GMl-ASLCS$_0$.



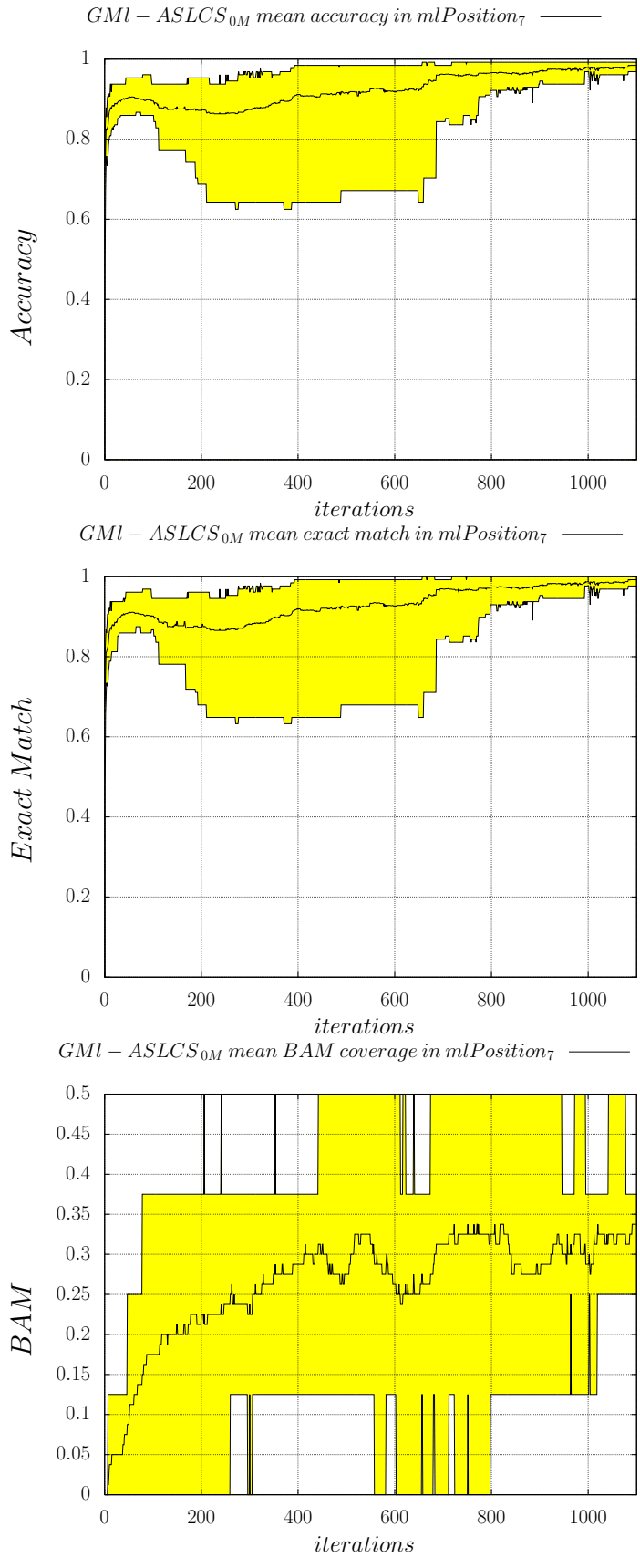
Σχήμα 7.5: Διαγράμματα χαρτογράφησης $mlPosition_7$ του GMl-ASLCS$_{0D}$.



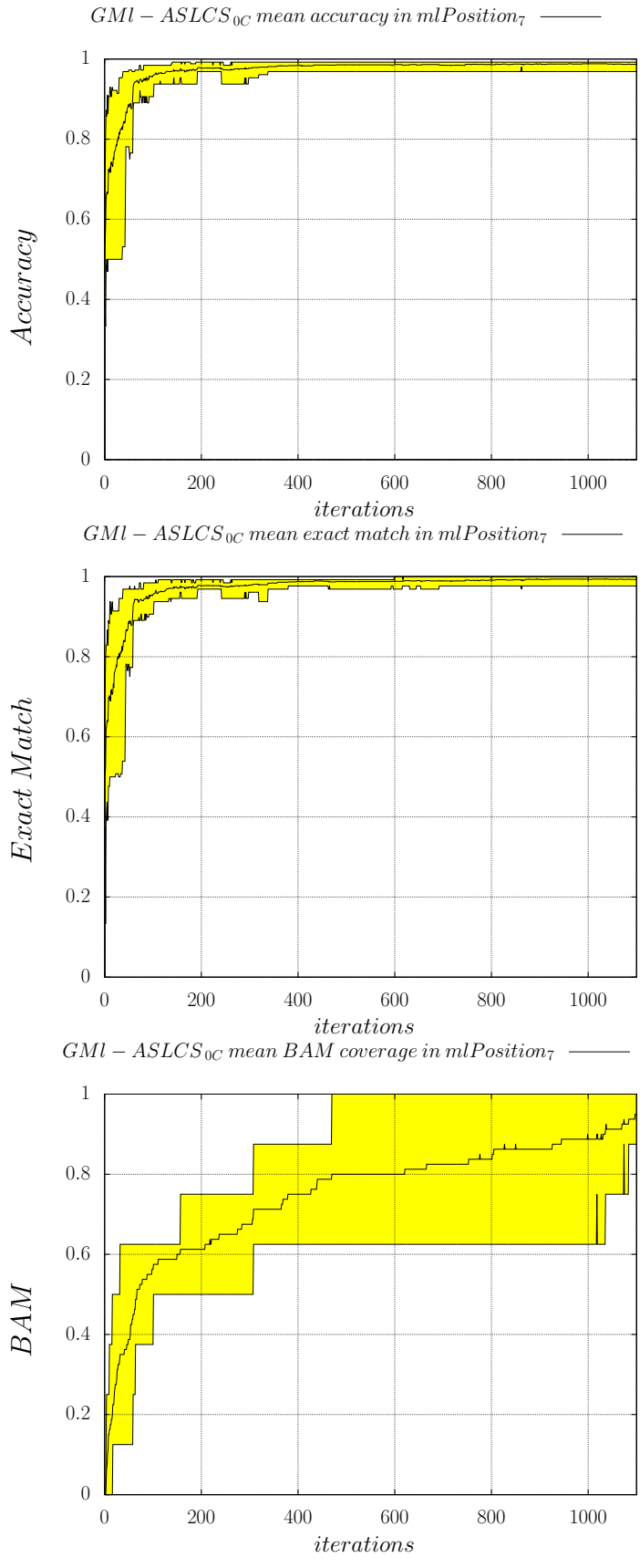
Σχήμα 7.6: Διαγράμματα χαρτογράφησης $mlPosition_7$ του GMl-ASLCS$_{0GA}$.



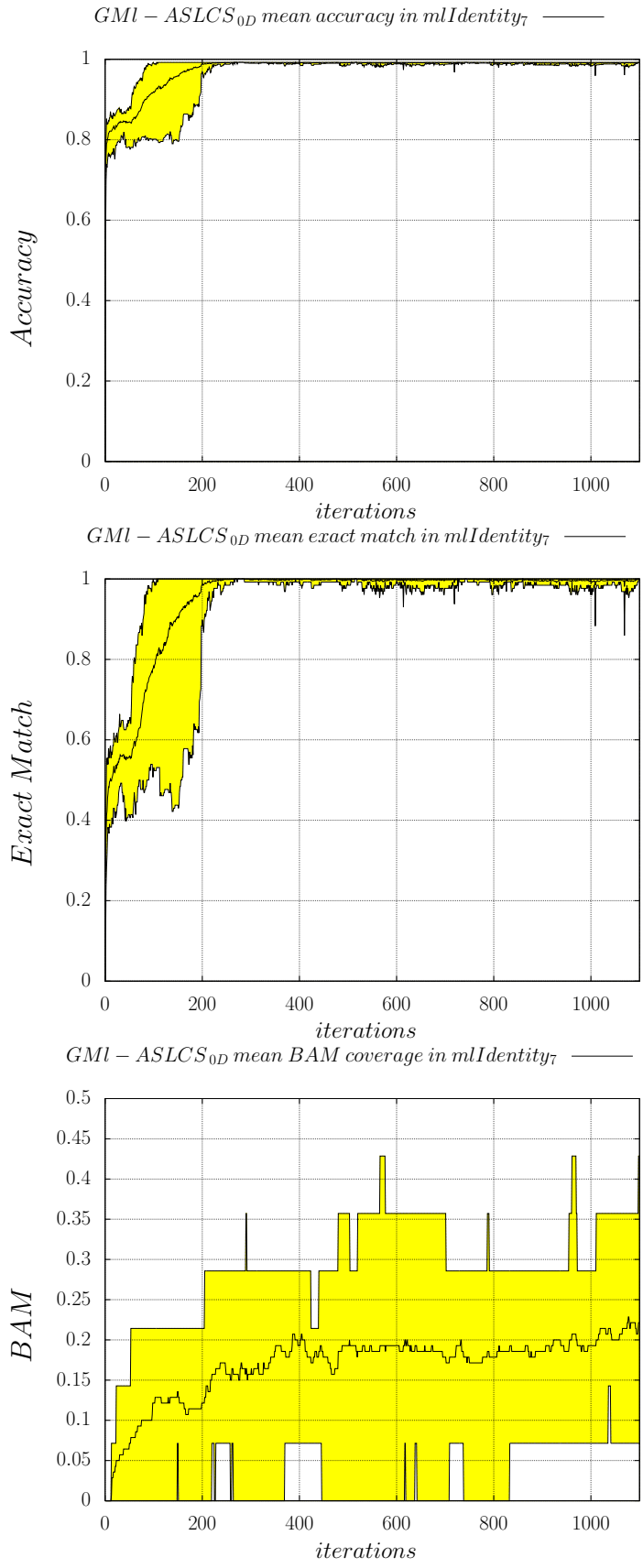
Σχήμα 7.7: Διαγράμματα χαρτογράφησης $mlPosition_7$ του GMl-ASLCS$_{0M}$.



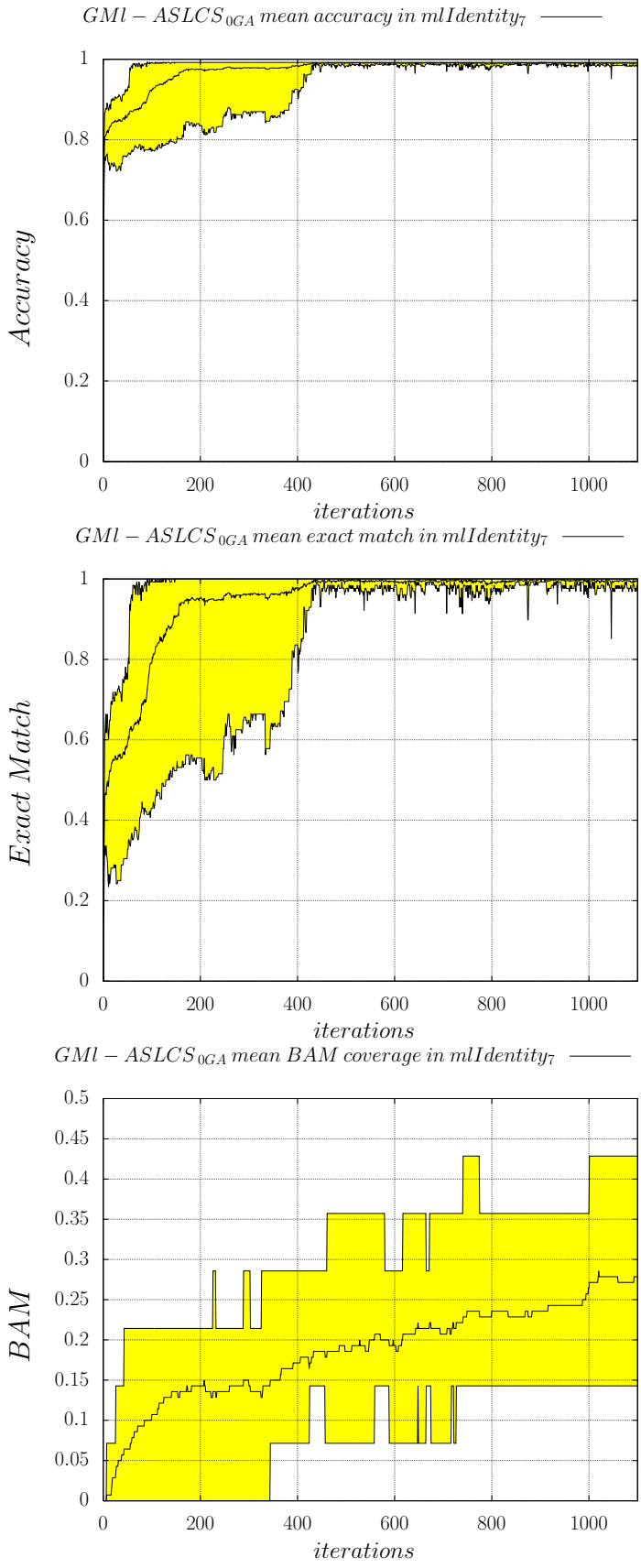
Σχήμα 7.8: Διαγράμματα χαρτογράφησης $mlPosition_7$ του GMl-ASLCS$_{0C}$.



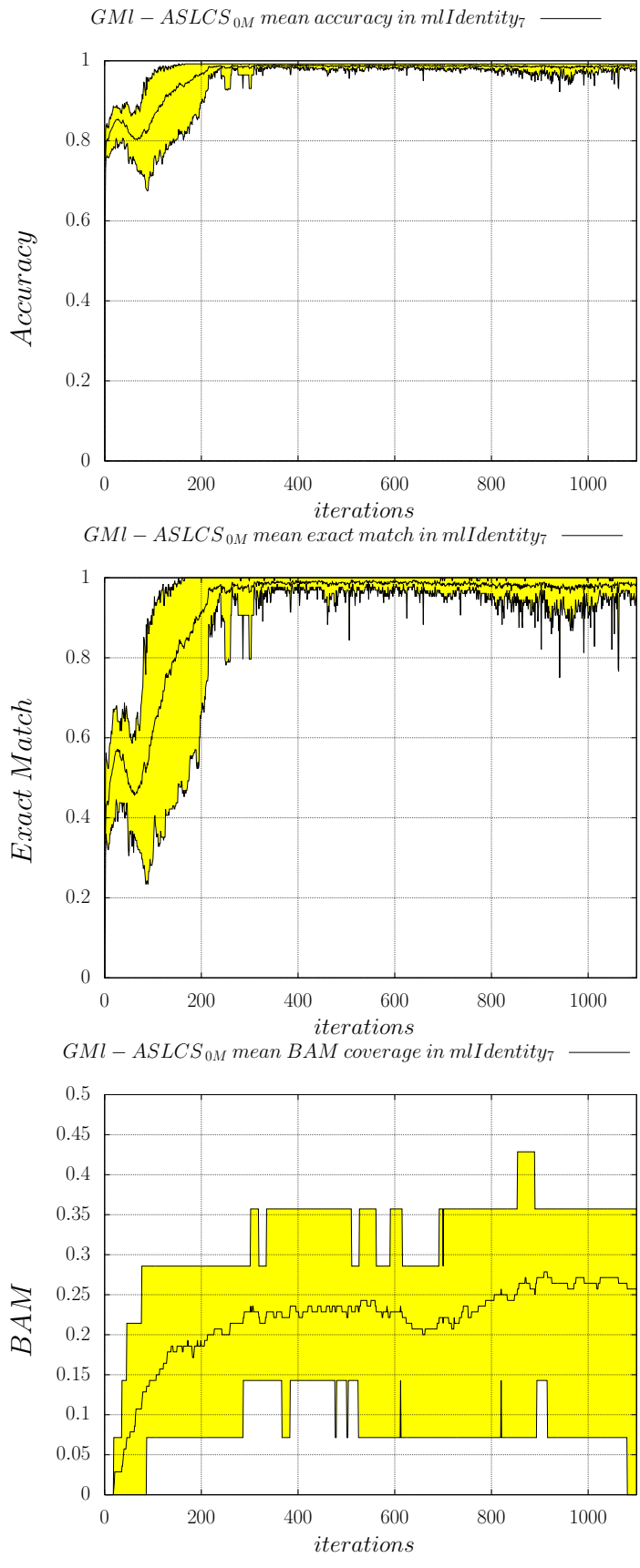
Σχήμα 7.9: Διαγράμματα χαρτογράφησης $mlIdentity_7$ του GMl-ASLCS$_{0D}$.



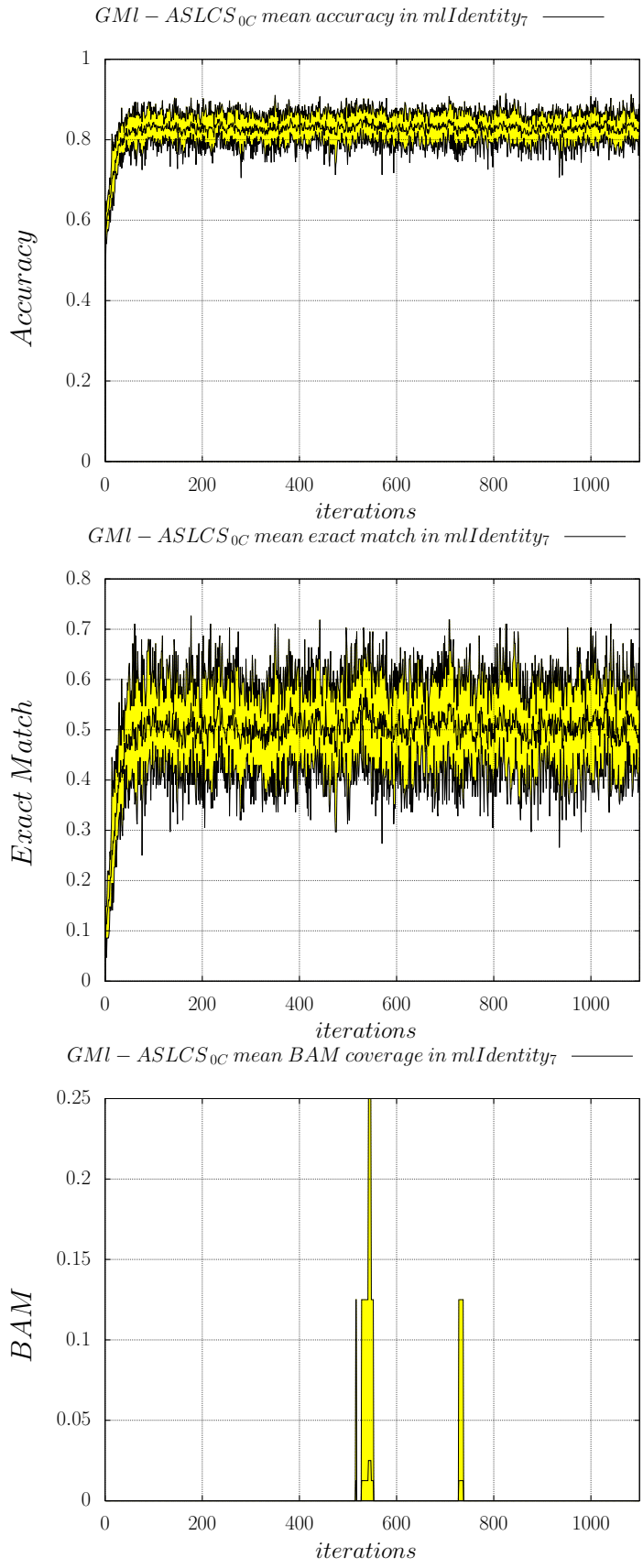
Σχήμα 7.10: Διαγράμματα χαρτογράφησης $mlIdentity_7$ του GMl-ASLCS$_{0GA}$.



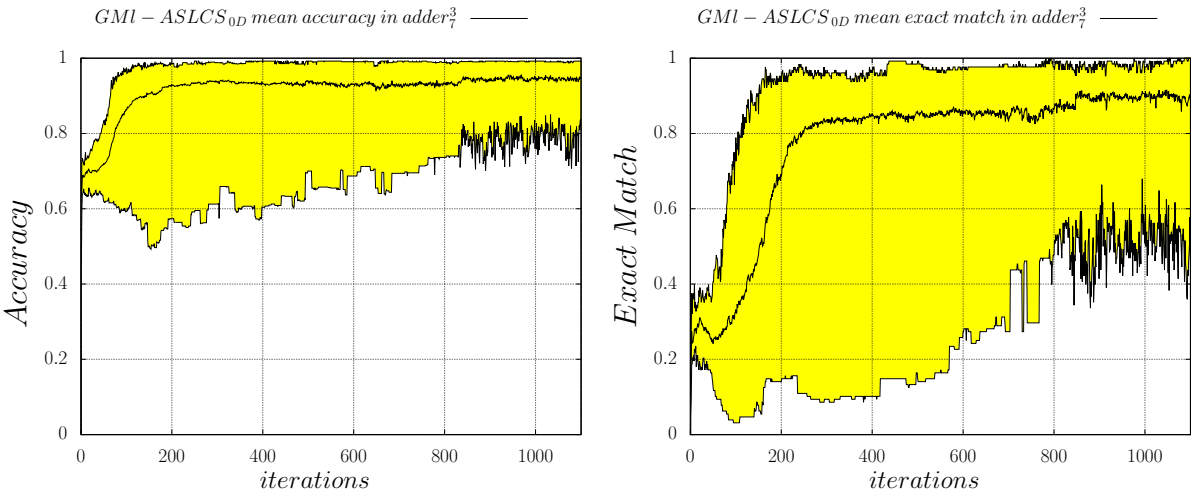
Σχήμα 7.11: Διαγράμματα χαρτογράφησης $mlIdentity_7$ του GMl-ASLCS$_{0M}$.



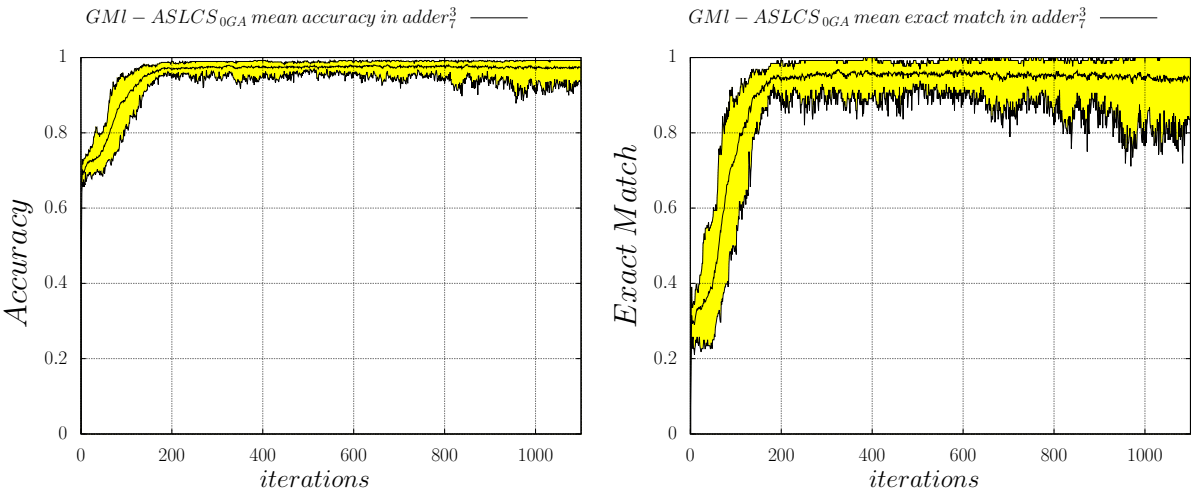
Σχήμα 7.12: Διαγράμματα χαρτογράφησης $mlIdentity_7$ του GMl-ASLCS$_{0C}$.



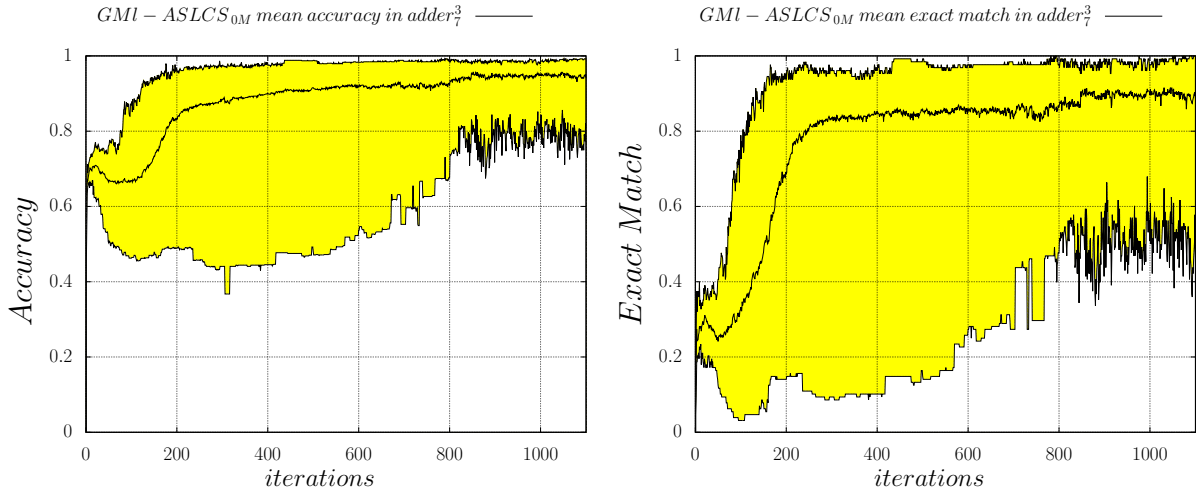
Σχήμα 7.13: Διαγράμματα χαρτογράφησης $adder_7^3$ του GMl-ASLCS$_{0D}$.



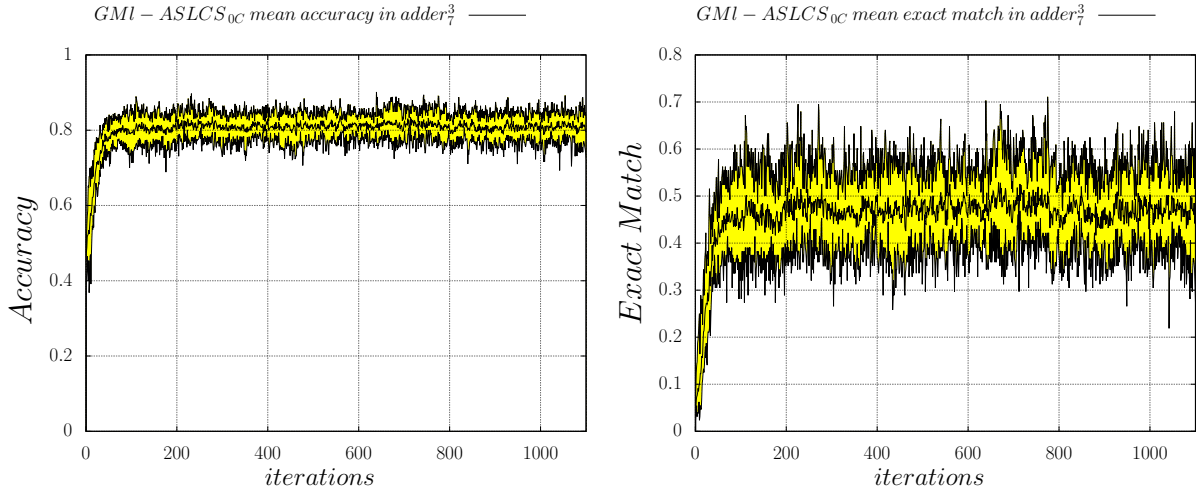
Σχήμα 7.14: Διαγράμματα χαρτογράφησης $adder_7^3$ του GMl-ASLCS$_{0GA}$.



Σχήμα 7.15: Διαγράμματα χαρτογράφησης $adder_7^3$ του GMl-ASLCS$_{0M}$.



Σχήμα 7.16: Διαγράμματα χαρτογράφησης $adder_7^3$ του GMl-ASLCS$_{0C}$.

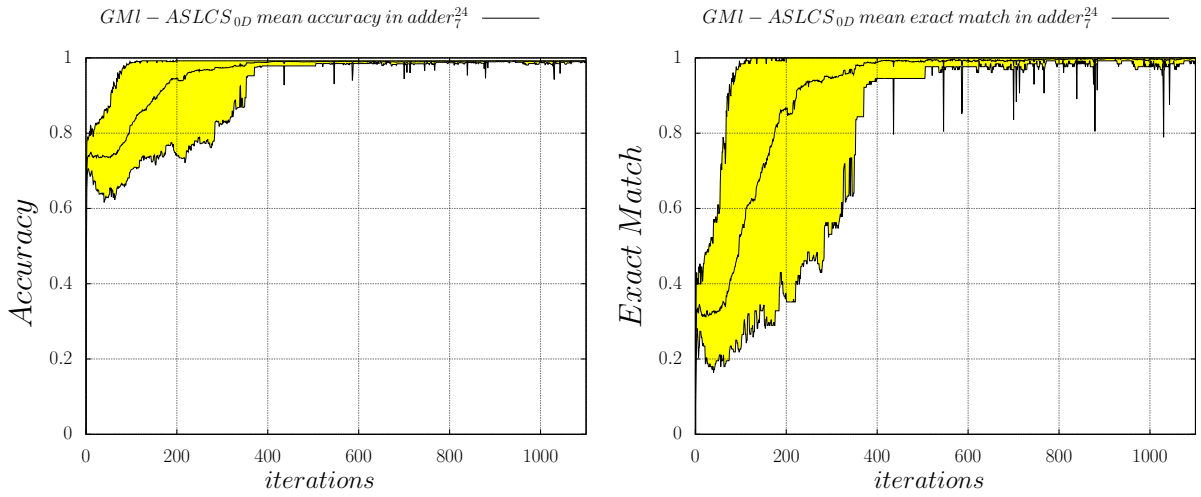


Οι τροποποιημένοι αλγόριθμοι GMl-ASLCS$_{0*}$ στο πρόβλημα $adder_7^{24}$

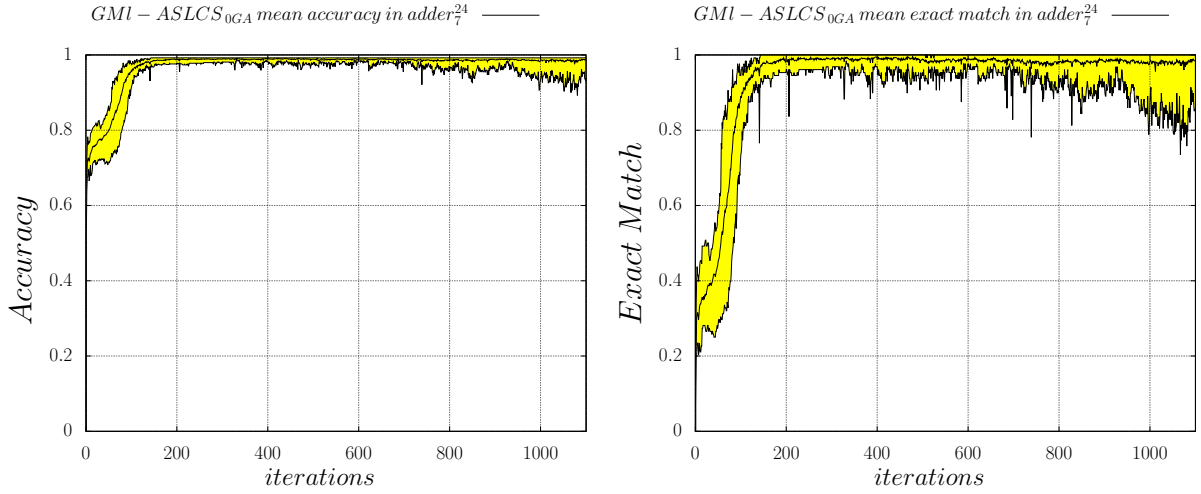
Τα Σχήματα 7.17, 7.18, 7.19 και 7.20 παρουσιάζουν την εξέλιξη των μετρικών της Ακρίβειας και της Ακριβούς Ορθότητας για το σύνολο δεδομένων $adder_7^{24}$, των ΜαΣΤ GMl-ASLCS$_{0D}$, GMl-ASLCS$_{0GA}$, GMl-ASLCS$_{0M}$ και GMl-ASLCS$_{0C}$, αντίστοιχα.

Λόγω της μεγάλης παρουσίας αδιαφοριών στους κανόνες του XBA και του γεγονότος ότι δεν τιμωρούνται οι αδιαφορίες στο τμήμα της απόφασης των κανόνων, οι τροποποιήσεις της λειτουργίας διαγραφής και του τελεστή διασταύρωσης κάνουν τους GMl-ASLCS$_{0D}$ και GMl-ASLCS$_{0GA}$ να καταφέρνουν όχι μόνο να συγκλίνουν προς τις βέλτιστες λύσεις, σε αντίθεση με τον GMl-ASLCS$_0$, αλλά και να το κάνουν σε λιγότερες από 600 επαναλήψεις. Και σε αυτό το σύνολο δεδομένων, όπως και στο $adder_7^3$, παρατηρούμε την πτώση (δηλαδή τη σημαντική βελτίωση) των ελάχιστων τιμών των μετρικών της Ακρίβειας και της Ακριβούς Ορθότητας στον GMl-ASLCS$_{0GA}$.

Σχήμα 7.17: Διαγράμματα χαρτογράφησης $adder_7^{24}$ του GMl-ASLCS$_{0D}$.



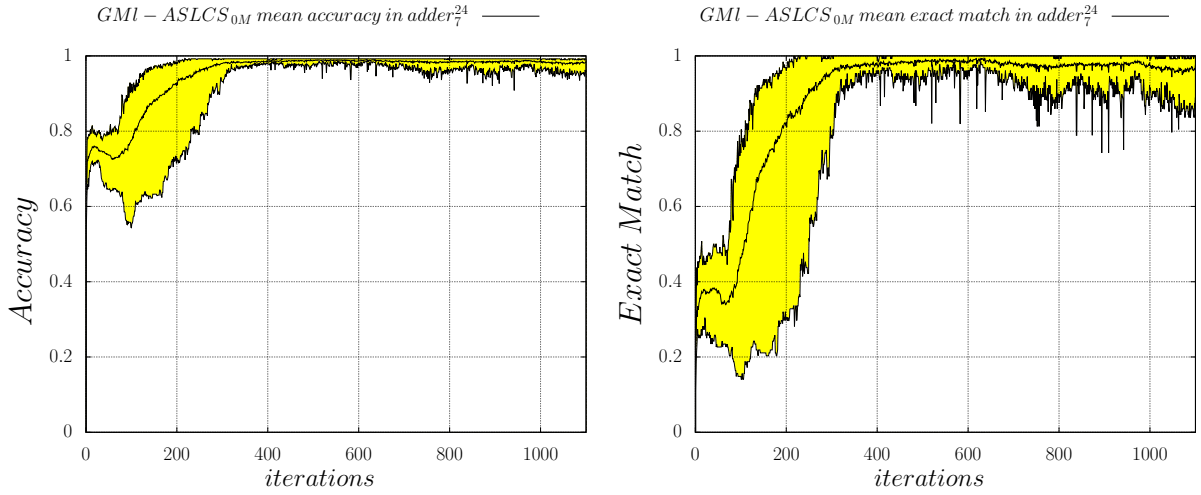
Σχήμα 7.18: Διαγράμματα χαρτογράφησης $adder_7^{24}$ του GMl-ASLCS$_{0GA}$.



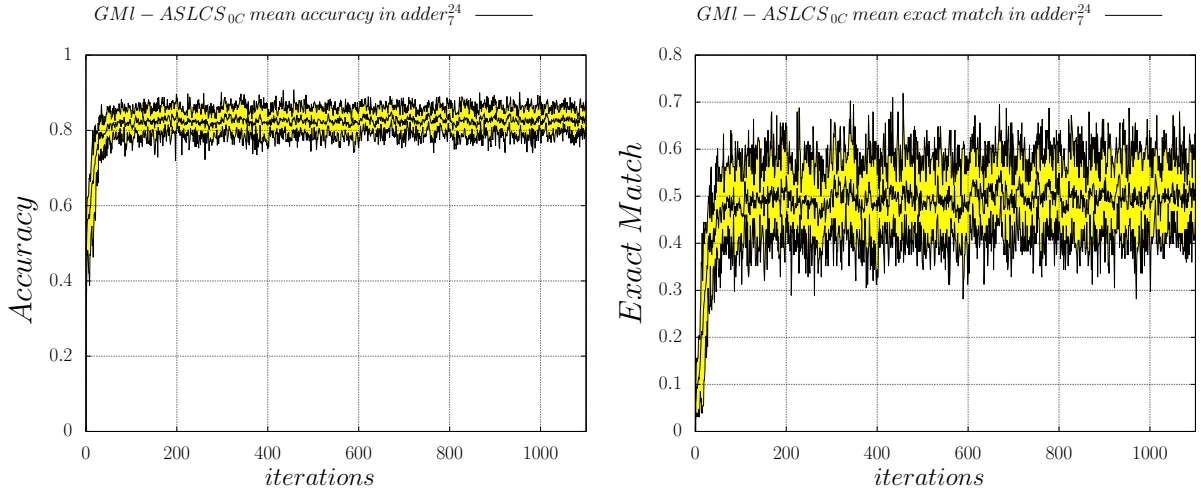
Ο GMl-ASLCS$_{0M}$ καταφέρνει να συγκλίνει και αυτός γρηγορότερα από τον GMl-ASLCS$_0$, και μάλιστα μέσα σε 400 επαναλήψεις, βελτιώνοντας ταυτόχρονα τις ελάχιστες τιμές των μετρικών αξιολόγησης που χρησιμοποιήθηκαν. Καταφέρνει επίσης να αυξήσει τη μέση κάλυψη δειγμάτων από τους κανόνες του πληθυσμού του, εκμεταλλευόμενος και εδώ την παρουσία αδιαφοριών στο τμήμα απόφασης των κανόνων του XBA.

Τέλος, ο GMl-ASLCS$_{0C}$ αποτυγχάνει και εδώ να βελτιώσει με το χρόνο τις λύσεις που εξελίσσει, καθώς παρατηρείται η ίδια συμπεριφορά όπως και στα σύνολα $mlIdentity_7$ και $adder_7^3$. Ενδιαφέρον είναι ότι η πολυκατηγορική πληθικότητα του συνόλου $adder_7^{24}$ έχει την ίδια τιμή με αυτή των δύο παραπάνω συνόλων.

Σχήμα 7.19: Διαγράμματα χαρτογράφησης $adder_7^{24}$ του GMl-ASLCS$_{0M}$.



Σχήμα 7.20: Διαγράμματα χαρτογράφησης $adder_7^{24}$ του GMl-ASLCS$_{0C}$.



Αποτίμηση της επίδρασης της τροποποιημένης λειτουργίας διαγραφής και του τελεστή διασταύρωσης Δύο Τμημάτων

Η συμπεριφορά των GMl-ASLCS$_{0D}$ και GMl-ASLCS$_{0GA}$, σε γενικές γραμμές, συνάδει με τα αποτελέσματα που αναμέναμε να επιφέρει η τροποποίηση της λειτουργίας διαγραφής και του τελεστή διασταύρωσης, αντίστοιχα, με βάση τους λόγους για την τροποποίησή τους. Γενικότερα, παρατηρούμε την ταχύτερη σύγκλιση σε όλα τα σύνολα δεδομένων για όλες τις μετρικές αξιολόγησης των μοντέλων που αναπτύσσουν οι GMl-ASLCS$_{0D}$ και GMl-ASLCS$_{0GA}$, την αύξηση των επιδόσεών τους, και την ταυτόχρονη μείωση του εύρους των χρησιμοποιούμενων μετρικών και στα δέκα πειράματα.

Όσον αφορά στην τροποποίηση της λειτουργίας διαγραφής, δεδομένης της ισοτιμίας αδιαφορίας και συμφωνίας για τις ετικέτες ως προς τη μεταβολή του αριθμού ορθών κατηγοριοποίησεων, το πρώτο σκέλος της Εξ. 6.12, που περιλαμβάνει τον τρόπο διαγραφής για τους κανόνες κάτω από το κατώφλι εμπειρίας θ_{del}, δείχνει

ότι μπορεί να επιτύχει αξιόλογο διαχωρισμό των κανόνων μέσα στο σύνολο νέων κανόνων με χαμηλή τιμή καταλληλότητας, αλλά και σε σχέση με το σύνολο των κανόνων με $experience \geq \theta_{del}$, λόγω της πιο άμεσης αφαίρεσης κανόνων χαμηλής ποιότητας από τον πληθυσμό, σε σχέση με τον GMl-ASLCS$_0$.

Όσον αφορά στην τροποποίηση του τελεστή διασταύρωσης, η Διασταύρωση Δύο Τμημάτων επιβεβαιώνεται ως μία αξιόπιστη και αποτελεσματικότερη εναλλακτική της Διασταύρωσης Ενός Σημείου, ιδιαίτερα μάλιστα στα πλαίσια της συνιστώσας ενημέρωσης του GMl-ASLCS$_0$, όπου η παρουσία αδιαφοριών στα τμήματα απόφασης των κανόνων είναι εντονότερη, λόγω της ισότιμης αντιμετώπισης αδιαφοριών και σαφών αποφάσεων υπέρ ετικετών. Εκτός της ορθότερης εναλλαγής αποφάσεων ανάμεσα στους κανόνες που διασταυρώνονται, και συνεπώς της εξέλιξης ακριβέστερων κανόνων, η μέθοδος Διασταύρωσης Δύο Τμημάτων είναι και ανεξάρτητη από την πολυκατηγορική πυκνότητα των συνόλων δεδομένων.

Αποτίμηση της επίδρασης της διαγραφής κανόνων με κριτήρια πάνω στα Match Sets

Από τη συμπεριφορά του GMl-ASLCS$_{0M}$ στα τέσσερα τεχνητά σύνολα δεδομένων συμπεραίνουμε πως η λειτουργία της διαγραφής κανόνων από τα Match Set είναι εύρωστη και αποτελεσματική ως προς την αύξηση του μέσου ποσοστού κάλυψης δειγμάτων, ενώ ταυτόχρονα δεν επηρεάζει με αρνητικό τρόπο την Ακρίβεια. Σε πιθανό συνδυασμό με την υλοποίηση της τιμωρίας κανόνων που αδιαφορούν για ετικέτες, θα περιμέναμε την απόσβεση όποιου φαινομένου αποσταθεροποίησης του πληθυσμού και των χαρακτηριστικών των κανόνων του, όπως φάνηκε από το γράφημα του Σχήματος 7.7, του μέσου ποσοστού εύρεσης του XBA του προβλήματος $mlPosition_7$ και τη συνολική βελτίωση της ευρωστίας της μεθόδου, αλλά και του αλγορίθμου.

Αποτίμηση της επίδρασης της έκπτωσης του αριθμού ορθών κατηγοριοποιήσεων για τους κανόνες που αδιαφορούν για ετικέτες

Το εύρος της Ακρίβειας των λύσεων που εξελίσσει ο GMl-ASLCS$_{0C}$ μειώνεται σε σχέση με αυτό των λύσεων του GMl-ASLCS$_0$, κάνοντας τον GMl-ASLCS$_{0C}$ να υπολείπεται σε επιδόσεις, εμφανίζοντας στάσιμη συμπεριφορά όταν η πολυκατηγορική πυκνότητα του συνόλου εκπαίδευσης εμφανίζει υψηλές τιμές. Αυτό είναι κάτι αναμενόμενο, καθώς, μειώνονται μεν με το πέρασμα των επαναλήψεων οι αδιαφορίες στα τμήματα απόφασης των κανόνων, λόγω της τιμωρίας της καταλληλότητάς τους, η εναλλαγή όμως ολόκληρων τμημάτων των αποφάσεων τους μέσω της Διασταύρωσης Ενός Σημείου λειτουργεί αποτρεπτικά προς την εύρεση νέων, βελτιωμένων λύσεων και, συνεπώς, προς την αύξηση των επιδόσεων του ΜαΣΤ. Εν τέλει, το ΜαΣΤ αναλώνεται στην "εύρεση" των ίδιων λύσεων, ενώ παράλληλα ο συνδυασμός του παραπάνω τελεστή διασταύρωσης και της έκπτωσης της ακρίβειας των κανόνων είναι δυνατόν να λειτουργήσει με τέτοιο τρόπο που να μην αυξάνει στον επιθυμητό βαθμό τη μέση κάλυψη των δειγμάτων από τους κανόνες του πληθυσμού. Παρ' όλα

αυτά, ο συνδυασμός αυτής της λειτουργίας, με τη Διασταύρωση Δύο Τμημάτων που είναι ανεξάρτητη από την πολυκατηγορική πυκνότητα των συνόλων εκπαίδευσης, αλλά και τη διαγραφή κανόνων από τα Match Sets που διαγράφει τους κανόνες χαμηλής καταλληλότητας στα χαμηλότερα επίπεδα κάλυψης, αναμένουμε ότι θα συνηγορήσουν υπέρ της αποτελμάτωσης των λύσεων που εξελίσσει ο GMl-ASLCS.

Πολυκατηγορικη Ταξινομηση με τον GMl-ASLCS

Ο GMl-ASLCS καταφέρνει να συγκεράσει και τις τέσσερις παραπάνω δομικές τροποποιήσεις και να τις συνδυάσει, ώστε να βελτιώσει ποσοτικά και ποιοτικά τη συμπεριφορά και τις επιδόσεις του αλγορίθμου, όχι μόνο ως προς τις μετρικές της Ακρίβειας και της Ακριβούς Ορθότητας, αλλά και της Μέσης Κάλυψης δειγμάτων από τους κανόνες που εξελίσσει. Στα Σχήματα 7.21, 7.22, 7.23 και 7.24 παρατίθεται η εξέλιξη του μετρικών αξιολόγησης για τα σύνολα $mlPosition_7$, $mlIdentity7$, $adder_7^3$ και $adder_7^{24}$, αντίστοιχα.

Στο σύνολο δεδομένων $mlPosition_7$, ο GMl-ASLCS συγκλίνει στη βέλτιστη λύση, και μάλιστα μέσα σε 600 επαναλήψεις, σε αντίθεση με τον GMl-ASLCS$_0$ που δεν καταφέρνει να συγκλίνει στο σύνολο των 1000 επαναλήψεων μάθησης. Από το διάγραμμα του ποσοστού εύρεσης του XBA, συμπεραίνουμε πως ο GMl-ASLCS καταφέρνει σε αρκετές περιπτώσεις να βρει ακόμα και όλους του κανόνες του, εμφανίζοντας συνεπέστερη συμπεριφορά, όσον αφορά στο εύρος του ποσοστού ανάμεσα στα 10 πειράματα. Η καμπύλη του ίδιου διαγράμματος έχει μεγαλύτερη κλίση από την αντίστοιχη του GMl-ASLCS$_0$, δηλαδή ο GMl-ASLCS καταφέρνει να βρει τους κανόνες του Βέλτιστου Χάρτη με ταχύτερο ρυθμό. Αυτό σημαίνει ότι ο GMl-ASLCS χρησιμοποιεί τη διαγραφή από τα Match Sets και την έκπτωση της ακρίβειας για κανόνες που αδιαφορούν για ετικέτες με βέλτιστο τρόπο, ώστε να παράξει και να εξελίξει κανόνες ταυτόχρονα μέγιστα γενικούς και ακριβείς.

Στο σύνολο $mlIdentity_7$, ο GMl-ASLCS αποτυγχάνει πλήρως να βρει τους κανόνες του Βέλτιστου Χάρτη, όπως ήταν αναμενόμενο, λόγω της τιμωρίας των κανόνων που αδιαφορούν για ετικέτες (για τον ίδιο λόγο που αποτυγχάνει και ο GMl-ASLCS$_{0C}$). Καθώς το ποσοστό εύρεσής των κανόνων του Βέλτιστου Χάρτη αποτελεί ένα είδος διαγνωστικού εργαλείου και μία ικανή αλλά όχι αναγκαία συνθήκη για εξέλιξη βέλτιστων λύσεων, μεταφέρουμε την προσοχή μας στα διαγράμματα της Ακρίβειας και της Ακριβούς Ορθότητας. Ο GMl-ASLCS, σε αντίθεση με τον GMl-ASLCS$_0$ καταφέρνει να συγκλίνει, να μειώσει το εύρος των μετρικών των λύσεων που εξελίσσει, αλλά και να είναι συνεπής ως προς την ποιότητά τους. ακριβώς λόγω της συνέπειας της ποιότητας των λύσεων που παράγει ο GMl-ASLCS, η όποια αποσταθεροποιητική συμβολή της λειτουργίας διαγραφής από τα Match Sets φαίνεται ότι αποσβένεται, ενώ, αν και δε φαίνεται στα διαγράμματα, η μέση κάλυψη των δειγμάτων αυξάνεται.

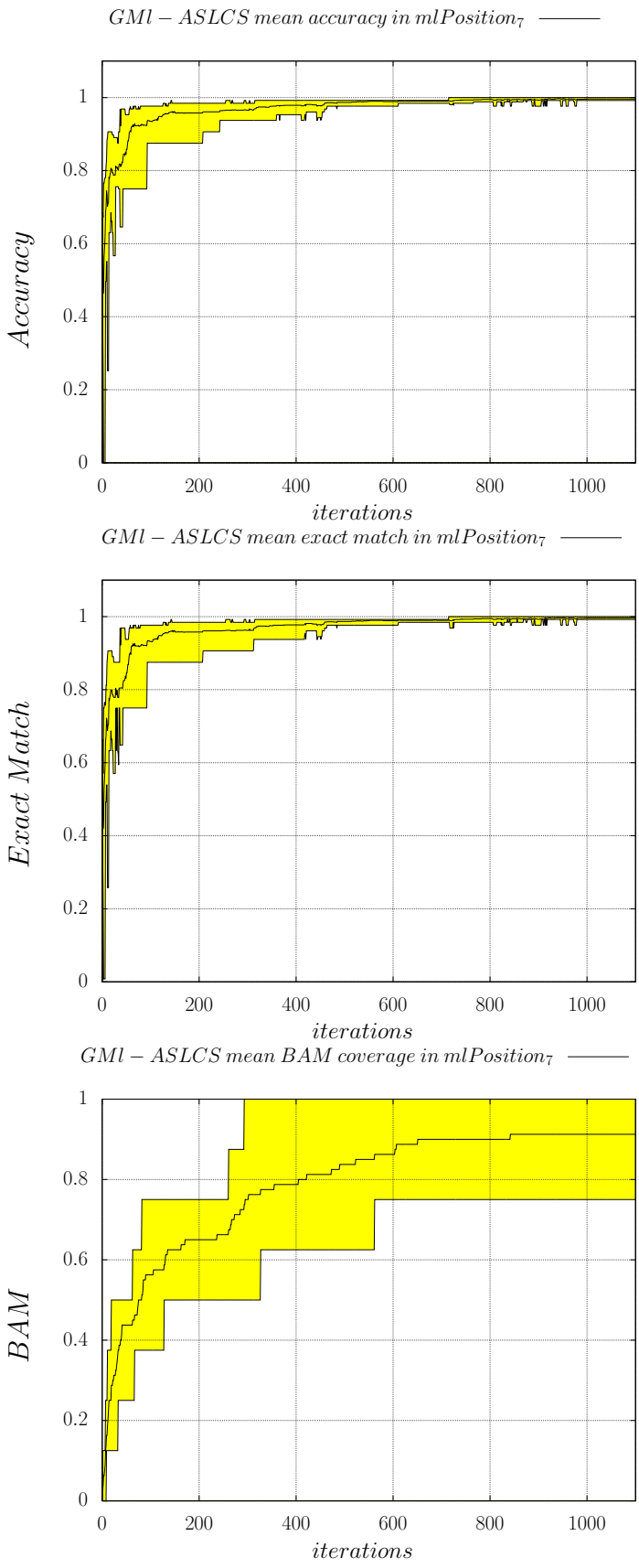
Όσον αφορά στο σύνολο $adder_7^3$, τα διαγράμματα της πορείας των μετρικών αξιολόγησης φαίνεται ότι ακολουθούν αυτές των διαγραμμάτων του αλγορίθμου GMl-ASLCS$_{0C}$, αλλά με μία μετατόπιση προς τα πάνω. Σε σχέση με τον GMl-ASLCS$_0$, ο GMl-ASLCS εμφανίζει και εδώ μία πιο συνεπή συμπεριφορά, που καταδεικνύεται από τη μείωση του εύρους των τιμών των χρησιμοποιούμενων μετρικών. Ο

GMl-ASLCS συγκλίνει γρηγορότερα στις ίδιες τιμές όσον αφορά στη μετρική της Ακρίβειας, αλλά μειώνει τις τιμές της Ακριβούς Ορθότητας, λόγω της έκπτωσης της ακρίβειας κανόνων που αδιαφορούν για ετικέτες. Λόγω των πολλών και ειδικών κανόνων που πρέπει να εξελίξει ο GMl-ASLCS για την αποτελεσματική λύση του προβλήματος, η έκπτωση της ακρίβειας σε συνδυασμό με τη διαγραφή κανόνων στο χαμηλότερο επίπεδο κάλυψης από τα Match Sets, φαίνεται ότι εξελίσσει κανόνες περισσότερο γενικούς από ότι χρειάζονται για τη λύση του προβλήματος, με ανάλογη πτώση της συνολικής προβλεπτικής τους ικανότητας.

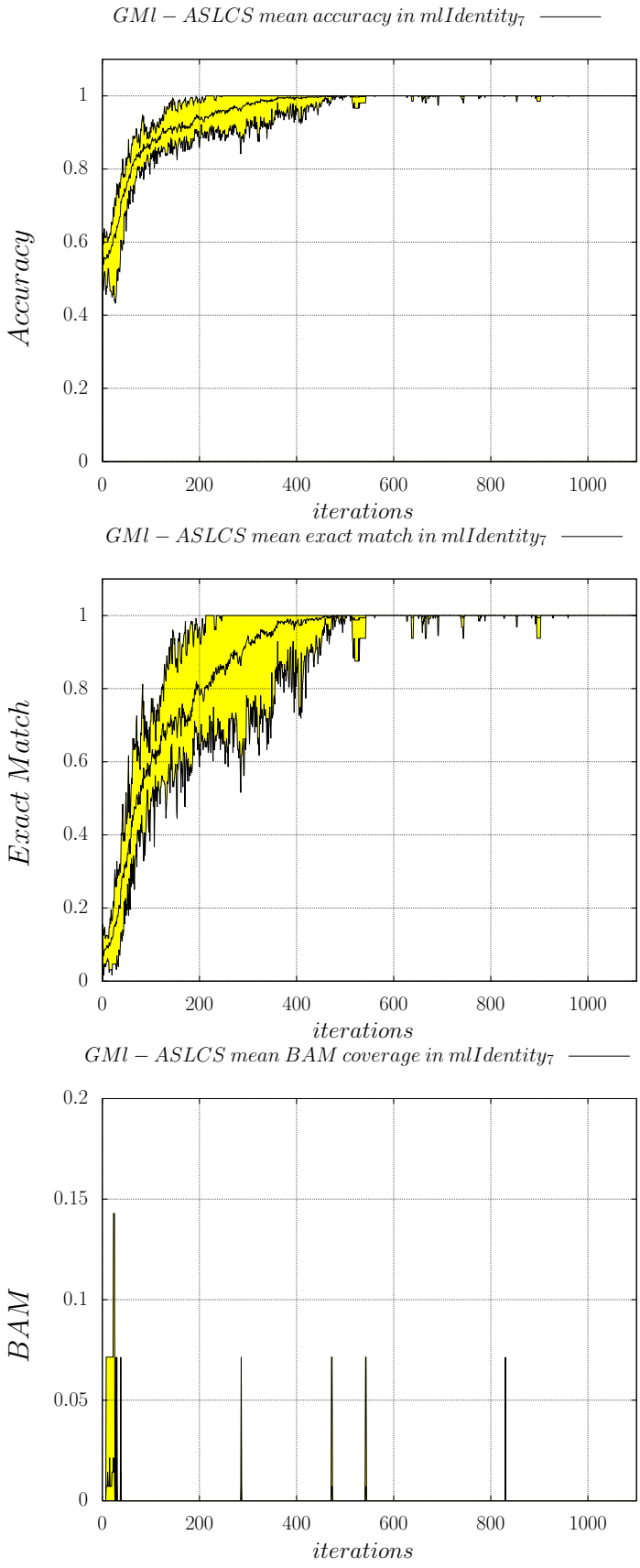
Αντίθετα, στο σύνολο $adder_7^{24}$, το οποίο απαιτεί μία πιο συμπαγή αναπαράσταση γνώσης, παρατηρούμε πως ο GMl-ASLCS γενικεύει στο βέλτιστο βαθμό με ταυτόχρονη διατήρηση των μέγιστων επιπέδων προβλεπτικής ικανότητας, τόσο από την πλευρά της Ακρίβειας, όσο και από αυτήν της Ακριβούς Ορθότητας. Επιπλέον, καταφέρνει όχι μόνο να συγκλίνει στις βέλτιστες τιμές για τις δύο προηγούμενες μετρικές, σε αντίθεση με τον GMl-ASLCS$_0$, αλλά και να το κάνει μέσα σε λιγότερο από 400 επαναλήψεις.

Συμπερασματικά, η αποτίμηση των τεσσάρων τροποποιήσεων κρίνεται από θετική έως εξαιρετικά θετική και για αυτό το λόγο αποφασίσαμε να τις διατηρήσουμε στο βασικό ορισμό του GMl-ASLCS.

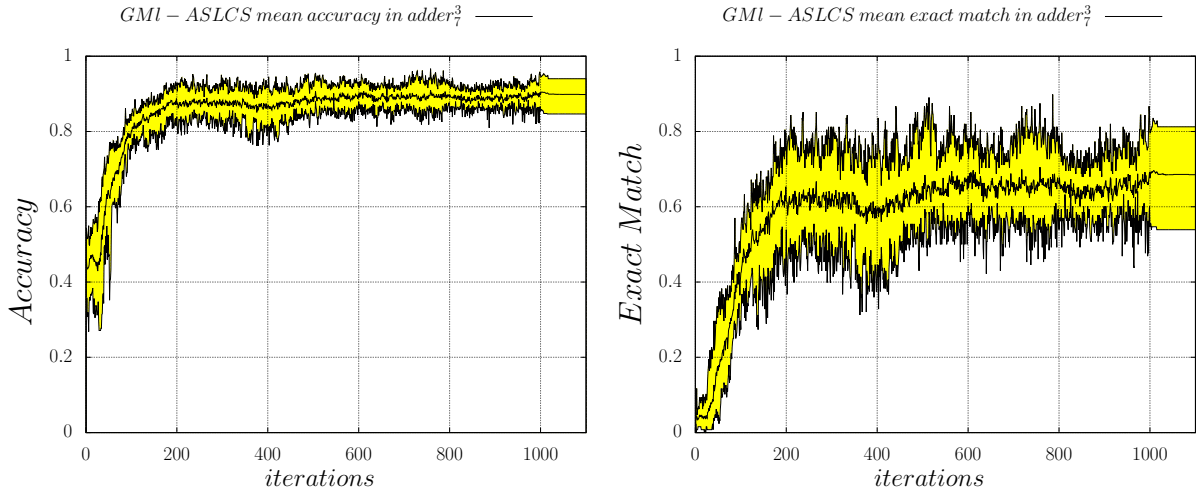
Σχήμα 7.21: Διαγράμματα χαρτογράφησης $mlPosition_7$ του GMl-ASLCS.



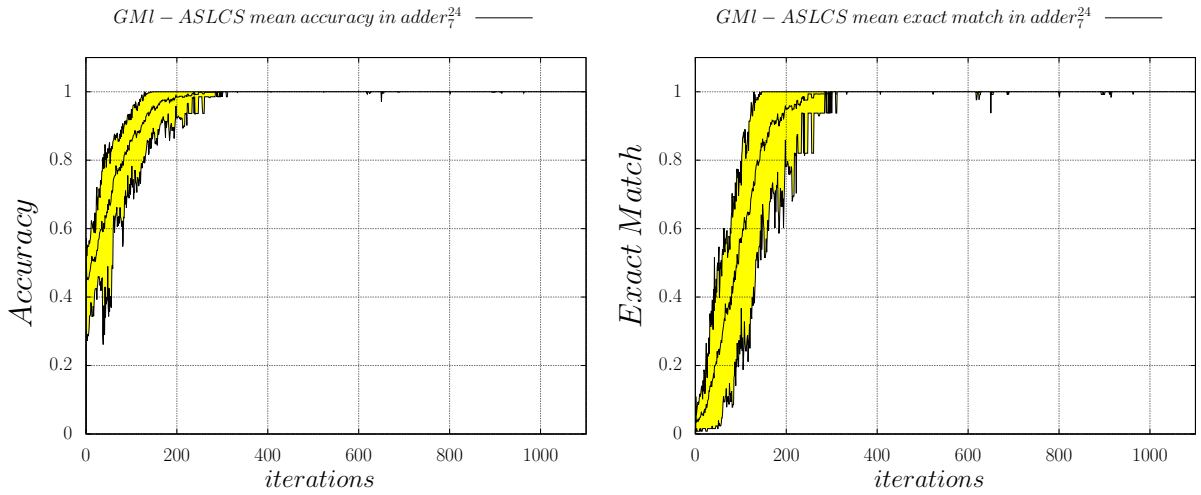
Σχήμα 7.22: Διαγράμματα χαρτογράφησης $mlIdentity_7$ του GMl-ASLCS.



Σχήμα 7.23: Διαγράμματα χαρτογράφησης $adder_7^3$ του GMl-ASLCS.



Σχήμα 7.24: Διαγράμματα χαρτογράφησης $adder_7^{24}$ του GMl-ASLCS.



8

Πειράματα Πραγματικών Συνόλων Δεδομένων

Στη διεθνή βιβλιογραφία υπάρχει πληθώρα διαφορετικών συνόλων πολυκατηγορικών δεδομένων που χρησιμοποιούνται ως μέσα δοκιμής των αλγορίθμων πολυκατηγορικής ταξινόμησης. Επιπλέον, στο [Mil11] παρουσιάζονται οι πρώτες επιδόσεις του GMl-ASLCS$_0$, σε σχέση με έξι πολυκατηγορικά σύνολα δεδομένων.

Το παρόν κεφάλαιο χωρίζεται σε δύο μέρη: στο πρώτο μέρος εστιάζουμε στα ίδια έξι σύνολα δεδομένων, παρουσιάζουμε τις διαφορές στις επιδόσεις ανάμεσα στον GMl-ASLCS$_0$ και τον GMl-ASLCS για τα σύνολα αυτά και συγκρίνουμε τις επιδόσεις του GMl-ASLCS με τρεις γνωστούς αλγορίθμους από τη βιβλιογραφία της πολυκατηγορικής ταξινόμησης. Στο δεύτερο μέρος διενεργούμε πειράματα α) για να εξακριβωθούν οι επιδόσεις των διαφόρων τροποποιήσεων που μπορούμε να επιφέρουμε στον GMl-ASLCS, που παρουσιάστηκαν στην Παρ. 6.5 και β) για να προσδιοριστεί η επίδραση των δύο πρωτοτυπιών που εισάγει αυτή η εργασία, του τελεστή Διασταύρωσης Δύο Τμημάτων και της λειτουργίας διαγραφής κανόνων με κριτήρια κάλυψης και καταλληλότητας από τα Match Sets, στην απόδοση του GMl-ASLCS.

Υπο Μελετη Συνολα Δεδομενων

Τα χαρακτηριστικά των πραγματικών συνόλων δεδομένων που χρησιμοποιήθηκαν για την πειραματική αξιολόγηση των GMl-ASLCS$_0$, GMl-ASLCS και των τροποποιήσεών του τελευταίου παρουσιάζονται στον Πίνακα 8.1, όπου $|D|$ είναι ο αριθμός δειγμάτων του συνόλου δεδομένων, $|L|$ ο αριθμός των ετικετών των δειγμάτων του και M ο αριθμός των γνωρισμάτων του, με το b να σηματοδοτεί τα δυαδικά γνωρίσματα και το n αριθμητικά. Οι στήλες LC, LD, P_{DIST} και P_{MaxF}

Πίνακας 8.1: Συνοπτικά χαρακτηριστικά των πραγματικών συνόλων δεδομένων.

Dataset	$\|D\|$	$\|L\|$	M	LC	LD	P_{DIST}	P_{MaxF}
Music	593	6	$72n$	1.87	0.31	0.046	0.137
Yeast	2417	14	$103n$	4.24	0.30	0.082	0.098
Genbase	661	27	$1185b$	1.25	0.05	0.048	0.257
Scene	2407	6	$294n$	1.07	0.18	0.006	0.168
Medical	978	45	$1449b$	1.25	0.03	0.096	0.158
Enron	1702	53	$1001b$	3.38	0.06	0.442	0.096

αναφέρονται αντίστοιχα στην πολυκατηγορική πληθικότητα, στην πολυκατηγορική πυκνότητα, στο ποσοστό μοναδικότητας των συνδυασμών ετικετών των δειγμάτων και στο ποσοστό συχνότερου συνδυασμού ετικετών (βλ. Παρ. 2.5.4).

Τα υπό μελέτη έξι σύνολα δεδομένων είναι διαθέσιμα στο δικτυακό χώρο της εφαρμογής πολυκατηγορικής ταξινόμησης Mulan[1] και είναι τα εξής:

Music Είναι ένα μικρό σύνολο δεδομένων [TTKV08] όπου κατηγοριοποιούνται κομμάτια μουσικής σε έξι πιθανά συναισθήματα (sad-lonely, angry-aggressive, amazed-surprised, relaxing-calm, quiet-still, happy-pleased).

Yeast Είναι ένα ευρέως χρησιμοποιούμενο σύνολο δεδομένων [EW02], όπου γονίδια συσχετίζονται με 14 βιολογικές λειτουργίες-ετικέτες.

Genbase Είναι ακόμη ένα βιολογικό σύνολο δεδομένων [DTMV05], παρόμοιο με το yeast, το οποίο καταγράφει τη συσχέτιση γονιδίων με 27 λειτουργίες-ετικέτες.

Scene Αναφέρεται στην πολυκατηγορική ταξινόμηση τοπίων (scenes) σε έξι πιθανές κατηγορίες (beach, sunset, field, fall-foliage, mountain, urban) [BLSB04].

Medical Αποτελεί ένα σύνολο δεδομένων ιατρικών κειμένων που δημιουργήθηκε για το Computational Medicine Centers 2007 Medical Natural Language Processing Challenge [PBM+07]. Κάθε έγγραφο αποτελείται από μία περιγραφή των συμπτωμάτων ενός ασθενή, ενώ οι ετικέτες αντιστοιχούν σε κωδικούς ασφάλισης.

Enron Αποτελεί το σύνολο δεδομένων που συλλέχθηκαν από τα ηλεκτρονικά μηνύματα της εταιρίας Enron στα πλαίσια του Enron Email Analysis[2], μετά το σκάνδαλο Enron το 2001. Αν και στο αρχικό σύνολο δεδομένων έχει γίνει ιεραρχική κατηγοριοποίηση των δεδομένων, στο τελικό σύνολο χρησιμοποιούνται μόνο οι ετικέτες-φύλλα.

[1] http://mulan.sourceforge.net/datasets.html
[2] http://bailando.sims.berkeley.edu/enron_email.html

Πειραματα σε Πραγματικα Συνολα Δεδομενων

Τα αναλυτικά αποτελέσματα από την αξιολόγηση του GMl-ASLCS πάνω στα έξι παραπάνω πραγματικά σύνολα δεδομένων παρατίθεται στους Πίνακες 8.3 έως 8.8 για κάθε σύνολο δεδομένων. Κάθε πίνακας αναφέρεται σε ένα σύνολο δεδομένων και περιέχει τις τιμές των μετρικών της Ακρίβειας, Ανάκλησης, Απώλειας Hamming και Ακριβούς Ορθότητας. Για κάθε μετρική αξιολόγησης εμφανίζονται τρία αποτελέσματα, ένα για μία από τις στρατηγικές συμπερασμού, IVal, PCut και Best, εφόσον αυτές είναι εφαρμόσιμες. Για την πληρέστερη αξιολόγηση του GMl-ASLCS, εκτός από τον GMl-ASLCS$_0$, παρατίθενται και οι επιδόσεις των state-of-art αλγορίθμων πολυκατηγορικής ταξινόμησης:

BR-J48 Αποτελεί το μετασχηματισμό *BR* των προβλημάτων με χρήση $|L|$ δυαδικών ταξινομητών J48.

RAkEL-J48 Αποτελεί έναν αλγόριθμο μετασχηματισμού σε k διαφορετικά σύνολα ετικετών [TV07]. Και εδώ ως ταξινομητής βάσης χρησιμοποιείται ο J48.

MlkNN Αποτελεί μία τροποποίηση του κλασικού ταξινομητή kNN (k Nearest Neighbors) που μπορεί να προβλέπει απευθείας δείγματα που ανήκουν σε περισσότερες από μία κατηγορίες.

Για τα πειράματα στα σύνολα δεδομένων music, yeast και genbase χρησιμοποιήθηκε αξιολόγηση 10-πλης διασταυρωμένης επικύρωσης (10-fold cross validation), ενώ για τα scene, medical και enron χρησιμοποιήθηκε ο προϋπάρχων (στο διαδικτυακό χώρο της εφαρμογής Mulan) χωρισμός τους σε σύνολα εκπαίδευσης και ελέγχου. Για τα τρία τελευταία σύνολα δεδομένων, τα σύνολα εκπαίδευσης και ελέγχου πάρθηκαν από το διαδικτυακό χώρο της εφαρμογής Mulan.

Πειραματική Μεθοδολογία

Όσον αφορά στην επιλογή των τιμών των παραμέτρων των επιμέρους λειτουργιών του GMl-ASLCS, εστιάσαμε την προσοχή μας σε ένα υποσύνολο των διαθέσιμων παραμέτρων. Οι παράμετροι αυτές ρυθμίστηκαν ανά πρόβλημα, σε πρώτη φάση με βάση τους εμπειρικούς κανόνες που αναφέρονται σε παρατηρήσιμα χαρακτηριστικά των υπό μελέτη συνόλων δεδομένων και, σε δεύτερη, μέσω μιας διαδικασίας δοκιμής και σφάλματος. Οι εμπειρικοί κανόνες που χρησιμοποιήθηκαν προκύπτουν από τις εξής παρατηρήσεις:

Αριθμός και είδος Γνωρισμάτων Ο αριθμός και το είδος των γνωρισμάτων των δειγμάτων ενός συνόλου δεδομένων επηρεάζει τον απαιτούμενο αριθμό κανόνων, ώστε να εξασφαλιστεί η αποτελεσματικότητα της εξελικτικής διαδικασίας. Επιπλέον, η ύπαρξη αριθμητικών γνωρισμάτων επιδεινώνει την κατάσταση, αφού δημιουργεί μεγαλύτερους χώρους αναζήτησης.

Πολυπλοκότητα Προβλήματος Αν και δε διαθέτουμε κάποιο σαφές μέσο υπολογισμού αυτού του μεγέθους, προβλήματα με αυξημένη πολυπλοκότητα, δηλαδή με μεγάλο αριθμό δειγμάτων, γνωρισμάτων, ετικετών, πολυκατηγορικής

πυκνότητας, ή/και διακριτών συνδυασμών ετικετών, απαιτούν συνήθως μεγαλύτερα μεγέθη πληθυσμών κανόνων και μεγαλύτερες τιμές πιθανότητας γενίκευσης (γνωρισμάτων ή/και ετικετών).

Ανισορροπία ετικετών Η ανισορροπία στη συχνότητα εμφάνισης των διακριτών συνδυασμών ετικετών επηρεάζει τη συμπεριφορά των ΜαΣΤ, αφού δημιουργούνται κανόνες με μεγάλες διαφορές ως προς την εμπειρία και την κάλυψη δειγμάτων και, επομένως, στην ποιότητα προσέγγισης των παραμέτρων τους. Για την προστασία εκείνων των κανόνων που καλύπτουν τα υπο-αντιπροσωπευόμενα δείγματα του χώρου αναζήτησης των ετικετών είναι αναγκαία α) η ακριβής ρύθμιση της πιθανότητας διαγραφής των κανόνων με βάση την εμπειρία και του κατωφλίου της εμπειρίας θ_{del}, β) η ρύθμιση του κατωφλίου εμπειρίας θ_{exp} όσον αφορά στην έκπτωση με βάση την εμπειρία στη συνιστώσα εξερεύνησης και γ) η ρύθμιση του ρυθμού ενεργοποίησης του Γενετικού Αλγορίθμου θ_{GA}, ώστε να εφαρμόζεται σε σύνολα κανόνων με ικανοποιητικές προσεγγίσεις των παραμέτρων ποιότητάς τους.

Βαθμός πλήρωσης του συνόλου δεδομένων Σύνολα δεδομένων με μεγάλο αριθμό γνωρισμάτων αλλά μικρό αριθμό δειγμάτων, σε σχέση με το μέγιστο αριθμό συνδυασμών των τιμών των γνωρισμάτων, απαιτούν προσεκτική ρύθμιση του μεγέθους του πληθυσμού και του ρυθμού ενεργοποίησης του Γενετικού Αλγορίθμου, ενώ χρειάζονται μεγάλες τιμές της πιθανότητας γενίκευσης γνωρισμάτων και μικρές της πιθανότητας γενίκευσης ετικετών (χαρακτηριστικά παραδείγματα αποτελούν τα σύνολα δεδομένων enron, medical και scene).

Οι συγκεκριμένες παράμετροι που χρησιμοποιήθηκαν ανά πείραμα παρατίθενται στον Πίνακα 8.2. Κάθε δείγμα παρουσιάζεται στο σύστημα I φορές, με συνολικό αριθμό επαναλήψεων μάθησης $\mathrm{I} \cdot |D|$. $P_{\#A}$ και $P_{\#L}$ είναι οι πιθανότητες γενίκευσης γνωρισμάτων και ετικετών, αντίστοιχα, που χρησιμοποιούνται από τη λειτουργία κάλυψης, ενώ $|P|$ είναι ο συνολικός αριθμός (άνω όριο) μικρο-κανόνων που επιθυμούμε να συγκρατήσει το ΜαΣΤ ως πληθυσμό.

Πίνακας 8.2: Παράμετροι πειραμάτων του GMl-ASLCS.

Σύνολο δεδομένων	I	$\lvert P\rvert$	θ_{GA}	$P_{\#A}$	$P_{\#L}$
music	500	5000	2000	0.80	0.01
yeast	500	18000	4000	0.85	0.01
genbase	500	12000	2000	0.40	0.10
scene	2500	9000	300	0.99	0.10
medical	4000	2500	2000	0.99	0.10
enron	600	25000	2000	0.99	0.10

Τέλος, οι κοινές παράμετροι που χρησιμοποιήθηκαν είχαν τις εξής τιμές: $\mu = 0.04$, $\chi = 0.8$, $\nu = 10$, $\beta = 0.2$, $\omega = 0.9$ και $\phi = 1$.

Πίνακας 8.3: Αποτελέσματα του GMl-ASLCS, του προκατόχου του, GMl-ASLCS$_0$, και των αλγορίθμων πολυκατηγορικής ταξινόμησης της βιβλιογραφίας οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Απώλειας Hamming και της Ακριβούς Ορθότητας στο σύνολο δεδομένων music.

	Ακρίβεια (%)			Ανάκληση (%)			Απώλεια Hamming (%)			Ακριβής Ορθότητα (%)		
	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best
GMl-ASLCS$_0$	49.57	50.05	43.44	61.14	63.51	52.19	23.6	24.05	26.48	23.78	23.11	22.42
GMl-ASLCS	**58.68**	**60.47**	**48.20**	**70.24**	**74.42**	**60.80**	**19.07**	**18.77**	**25.11**	**33.75**	**34.39**	**24.31**
BR-J48	-	46.23	-	-	59.94	-	-	24.74	-	-	18.38	-
RAkEL-J48	-	50.91	-	-	62.73	-	-	21.81	-	-	24.78	-
MlkNN	-	53.26	-	-	60.50	-	-	19.51	-	-	28.31	-

Πίνακας 8.4: Αποτελέσματα του GMl-ASLCS, του προκατόχου του, GMl-ASLCS$_0$, και των αλγορίθμων πολυκατηγορικής ταξινόμησης της βιβλιογραφίας οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Απώλειας Hamming και της Ακριβούς Ορθότητας στο σύνολο δεδομένων yeast.

	Ακρίβεια (%)			Ανάκληση (%)			Απώλεια Hamming (%)			Ακριβής Ορθότητα (%)		
	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best
GMl-ASLCS$_0$	46.09	45.51	39.01	60.83	65.69	44.28	23.91	25.16	**23.90**	6.33	5.01	5.79
GMl-ASLCS	**51.92**	**51.67**	**41.49**	**64.20**	**69.06**	**55.15**	**21.23**	22.30	26.88	**13.07**	11.36	**8.80**
BR-J48	-	43.95	-	-	57.84	-	-	24.54	-	-	6.83	-
RAkEL-J48	-	48.74	-	-	62.89	-	-	22.58	-	-	11.71	-
MlkNN	-	51.62	-	-	59.13	-	-	**19.33**	-	-	**18.74**	-

Πίνακας 8.5: Αποτελέσματα του GMl-ASLCS, του προκατόχου του, GMl-ASLCS$_0$, και των αλγορίθμων πολυκατηγορικής ταξινόμησης της βιβλιογραφίας οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Απώλειας Hamming και της Ακριβούς Ορθότητας στο σύνολο δεδομένων genbase.

	Ακρίβεια (%)			Ανάκληση (%)			Απώλεια Hamming (%)			Ακριβής Ορθότητα (%)		
	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best
GMl-ASLCS$_0$	86.71	87.90	89.45	91.04	90.86	90.04	1.07	0.99	**0.75**	84.43	84.16	**88.38**
GMl-ASLCS	**97.47**	**98.63**	**91.01**	**98.81**	98.81	**97.00**	**0.21**	0.12	0.88	**94.72**	96.99	83.68
BR-J48	-	98.62	-	-	**99.14**	-	-	**0.11**	-	-	**97.13**	-
RAkEL-J48	-	98.62	-	-	**99.14**	-	-	**0.11**	-	-	**97.13**	-
MlkNN	-	94.11	-	-	94.28	-	-	0.50	-	-	90.64	-

Πίνακας 8.6: Αποτελέσματα του GMl-ASLCS, του προκατόχου του, GMl-ASLCS$_0$, και των αλγορίθμων πολυκατηγορικής ταξινόμησης της βιβλιογραφίας οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Απώλειας Hamming και της Ακριβούς Ορθότητας στο σύνολο δεδομένων scene.

	Ακρίβεια (%)			Ανάκληση (%)			Απώλεια Hamming (%)			Ακριβής Ορθότητα (%)		
	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best
GMl-ASLCS$_0$	41.14	42.39	44.88	44.98	51.84	45.90	19.30	21.46	17.66	34.44	30.27	41.38
GMl-ASLCS	**62.82**	63.15	**52.59**	**69.23**	70.23	**53.30**	**12.04**	12.11	**15.82**	**53.43**	53.18	**48.83**
BR-J48	-	51.34	-	-	61.12	-	-	13.89	-	-	40.13	-
RAkEL-J48	-	57.76	-	-	62.17	-	-	11.50	-	-	50.75	-
MlkNN	-	**66.14**	-	-	**77.24**	-	-	**9.53**	-	-	**60.12**	-

Πίνακας 8.7: Αποτελέσματα του GMl-ASLCS, του προκατόχου του, GMl-ASLCS$_0$, και των αλγορίθμων πολυκατηγορικής ταξινόμησης της βιβλιογραφίας οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Απώλειας Hamming και της Ακριβούς Ορθότητας στο σύνολο δεδομένων medical.

	Ακρίβεια (%)			Ανάκληση (%)			Απώλεια Hamming (%)			Ακριβής Ορθότητα (%)		
	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best
GMl-ASLCS$_0$	39.49	40.14	**44.92**	46.20	48.30	47.55	2.98	3.10	**2.78**	28.53	27.93	**36.34**
GMl-ASLCS	**52.34**	51.58	38.73	**60.03**	56.80	**49.72**	**2.06**	1.95	3.20	**39.84**	41.40	22.79
BR-J48	-	**74.26**	-	-	**79.41**	-	-	**1.06**	-	-	**65.12**	-
RAkEL-J48	-	72.84	-	-	78.22	-	-	1.13	-	-	64.03	-
MlkNN	-	41.77	-	-	43.31	-	-	1.88	-	-	35.19	-

Πίνακας 8.8: Αποτελέσματα του GMl-ASLCS, του προκατόχου του, GMl-ASLCS$_0$, και των αλγορίθμων πολυκατηγορικής ταξινόμησης της βιβλιογραφίας οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Απώλειας Hamming και της Ακριβούς Ορθότητας στο σύνολο δεδομένων enron.

	Ακρίβεια (%)			Ανάκληση (%)			Απώλεια Hamming (%)			Ακριβής Ορθότητα (%)		
	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best	PCut	IVal	Best
GMl-ASLCS$_0$	38.93	39.40	25.04	53.67	49.81	29.81	6.26	6.00	**6.77**	**7.43**	**11.92**	**9.15**
GMl-ASLCS	**40.36**	40.35	**27.33**	**54.34**	**54.50**	**43.74**	**5.88**	5.88	8.10	7.08	6.91	2.59
BR-J48	-	36.71	-	-	44.81	-	-	5.40	-	-	8.64	-
RAkEL-J48	-	**41.04**	-	-	49.46	-	-	**5.09**	-	-	10.71	-
MlkNN	-	31.85	-	-	35.80	-	-	5.14	-	-	6.22	-

Συγκριτικη Αναλυση Αποτελεσματων

Τα αποτελέσματα που προέκυψαν από την πειραματική διαδικασία σχολιάζονται στις επόμενες παραγράφους.

Σύγκριση του GMl-ASLCS με τον προκάτοχό του

Ο GMl-ASLCS εμφανίζει καλύτερα αποτελέσματα[3] ως προς όλες τις μετρικές σε σχέση με τον GMl-ASLCS, για όλα τα σύνολα δεδομένων, εκτός από τη μετρική της ακρίβειας για το σύνολο δεδομένων medical για τη στρατηγική συμπερασμού best, για τη μετρική της Απώλειας Hamming στα σύνολα δεδομένων yeast, genbase, medical και enron για τη στρατηγική συμπερασμού best και της Ακριβούς Ορθότητας για τα σύνολα δεδομένων genbase, medical και enron για όλες τις στρατηγικές συμπερασμού. Όσον αφορά στη βασική μετρική αξιολόγησης, την ακρίβεια, η προσέγγισή μας τη βελτιώνει κατά 20.82% στο σύνολο music, 13.54% στο σύνολο yeast, 12.21% στο genbase, 48.97% στο scene, 28.50% στο medical και κατά 2.41% στο σύνολο enron, με βάση τη μέθοδο επιλογής κατωφλίου IVal.

Στους Πίνακες 8.9 και 8.10 παρουσιάζονται, για κάθε σύνολο δεδομένων, το ποσοστό μέσης κάλυψης δειγμάτων (coverage %), το άνω όριο του αριθμού των μικρο-κανόνων του πληθυσμού $|P|$, ο αριθμός των μικρο-κανόνων του τελικού μοντέλου των δύο ΜαΣΤ $|P|_{final}$ και το ποσοστό της απόκλισης του τελικού αριθμού των μικρο-κανόνων του πληθυσμού από το μέγιστο (population lost %). Σε αυτούς τους πίνακες αυτό που πρέπει να παρατηρήσουμε, κατ' αρχάς, είναι η αύξηση του ποσοστού της μέσης κάλυψης δειγμάτων των κανόνων του GMl-ASLCS σε σχέση με αυτό του GMl-ASLCS$_0$. Αυτή οδηγεί σε απόλυτη αύξηση σε κάλυψη 8 δειγμάτων για το σύνολο music, 10 για το yeast, 5.4 για το genbase, 6.7 για το scene και 2.89 δειγμάτων για το σύνολο enron. Αν και στο σύνολο medical παρατηρείται πτώση του ποσοστού κάλυψης, αυτή είναι πλασματική: λόγω της υπερ-παραγωγής κανόνων μηδενικής κάλυψης, η ποιότητα των χρήσιμων κανόνων του πληθυσμού έχει συμπιεστεί σε τέτοιο βαθμό ώστε ο GMl-ASLCS$_0$ εξελίσσει μόνο πολύ γενικούς κανόνες, περισσότερο δηλαδή από όσο θα έπρεπε ώστε να μην διακινδυνεύει την ακρίβειά του. Το φαινόμενο παραγωγής κανόνων μηδενικής κάλυψης στον GMl-ASLCS$_0$ έχει τέτοια ένταση που το τελικό σύνολο κανόνων αποτελείται μόλις από το 14.69% του αριθμού των κανόνων που θα έπρεπε να συγκρατεί ο πληθυσμός.

Η στήλη $|P|_{final}$ στον Πίνακα 8.9, σε αντιπαραβολή με τα στοιχεία της στήλης $|P|$, χρησιμοποιείται για να παρουσιάσει το αποτέλεσμα της συγκράτησης στον πληθυσμό κανόνων μηδενικής κάλυψης στον GMl-ASLCS$_0$, όσον αφορά στα πραγματικά σύνολα δεδομένων, δεδομένου πως στην περίοδο ενημέρωσης $0.1 \cdot |I|$ επαναλήψεων που ακολουθεί το ουσιώδες μέρος της διαδικασίας εκπαίδευσης, η μόνη λειτουργία αφαίρεσης κανόνων από τον πληθυσμό είναι αυτή της αφαίρεσης κανόνων λόγω της διαπίστωσης πως δεν καλύπτουν κανένα δείγμα του συνόλου δεδομένων.

[3] Για τις μετρικές της Ακρίβειας, της Ανάκλησης, της Ακριβούς Ορθότητας και της Μέσης Κάλυψης, μεγαλύτερες τιμές είναι καλύτερες. Για την Απώλεια Hamming ισχύει το αντίθετο.

Στον Πίνακα 8.10 η στήλη population lost % και, κατά συνέπεια, και η στήλη $|P|_{final}$, χρησιμοποιούνται για να προσδώσουν μία εικόνα του μεγέθους του αριθμού των κανόνων που διαγράφονται από τη λειτουργία διαγραφής κανόνων από τα Match Sets, κατά το διάστημα ενημέρωσης στον GMl-ASLCS. Και στα δύο ΜαΣΤ, η λειτουργία διαγραφής κανόνων μέσω επιλογής ρουλέτας είναι απενεργοποιημένη λόγω της απενεργοποίησης του Γενετικού Αλγορίθμου.

Εν κατακλείδι, ο GMl-ASLCS επιτυγχάνει τους δύο στόχους που είχαμε θέσει: α) τη βελτίωση της ακρίβειάς του, που αποδεικνύεται από τις επιδόσεις του στα έξι πραγματικά σύνολα πολυκατηγορικών δεδομένων που δοκιμάσαμε, και β) την αύξηση των δειγμάτων που καλύπτουν οι κανόνες του πληθυσμού του, μέσω της λειτουργίας διαγραφής κανόνων από τα Match Sets. Συγκεκριμένα, όσον αφορά στην αύξηση της μέσης κάλυψης των κανόνων, φαίνεται ότι αυτή δεν ακυρώνεται και ότι δε δημιουργείται κάποια επιπρόσθετη πίεση προς τη συνολικότερη ειδίκευση των κανόνων, μέσω της τροποποίησης συνιστωσών ή λειτουργιών.

Πίνακας 8.9: Μέση Κάλυψη δειγμάτων στα πειράματα του GMl-ASLCS$_0$.

dataset	coverage %	$\|P\|$	$\|P\|_{final}$	population lost %
music	0.2827	8000	6618	17.23
yeast	0.1361	14000	6673	52.33
genbase	2.1677	6000	5023	16.28
scene	0.1628	10000	4043	59.57
medical	4.7601	8000	1175	85.31
enron	0.3545	10000	6133	38.67

Πίνακας 8.10: Μέση Κάλυψη δειγμάτων στα πειράματα του GMl-ASLCS.

dataset	coverage %	$\|P\|$	$\|P\|_{final}$	population lost %
music	1.7896	5000	4872	2.56
yeast	0.4733	18000	17721	1.55
genbase	3.0724	12000	10980	8.50
scene	0.4411	9000	8857	1.59
medical	3.2497	2500	2498	0.08
enron	0.5242	25000	22841	8.64

Σύγκριση των Αλγορίθμων με βάση την Ακρίβεια

Ο Πίνακας 8.11 συνοψίζει τα ποσοστά της ακριβείας των υπό μελέτη αλγορίθμων για όλα τα χρησιμοποιούμενα σύνολα δεδομένων, για τη στρατηγική συμπερασμού IVal. Μαζί με τις μετρικές επίδοσης του καθενός, αναφέρουμε και τις αντίστοιχες κατατάξεις ανά σύνολο δεδομένων ως εκθέτες στην τιμή της μετρικής, τη μέση κατά-

ταξη κάθε αλγορίθμου όπως αυτή προκύπτει από την εφαρμογή του τεστ Friedman στη στήλη με τίτλο "Κατάταξη", καθώς και την απόλυτη θέση των αλγορίθμων στην τελική κατάταξη ως εκθέτη της τιμής της μέσης κατάταξης.

Πίνακας 8.11: Σύγκριση αλγορίθμων πολυκατηγορικής ταξινόμησης με βάση την ακρίβεια, με στρατηγική συμπερασμού IVal, στο σύνολο των υπό μελέτη προβλημάτων ταξινόμησης. Οι εκθέτες αναφέρονται στην κατάταξη του κάθε αλγορίθμου ανά σύνολο δεδομένων, σύμφωνα με το στατιστικό τεστ Friedman. Η στήλη με τίτλο "Κατάταξη" περιέχει τη συνολική κατάταξη του αλγορίθμου στην αντίστοιχη γραμμή, ενώ ο εκθέτης σημαίνει τη θέση του στην (απόλυτη) τελική κατάταξη.

Αλγόριθμοι	music	yeast	genbase	scene	medical	enron	Κατάταξη
GMl-ASLCS$_0$	50.05^{4}	45.51^{4}	87.90^{5}	42.39^{5}	40.14^{5}	39.40^{3}	4.33^{5}
GMl-ASLCS	$\mathbf{60.47}^{1}$	$\mathbf{51.67}^{1}$	$\mathbf{98.63}^{1}$	63.15^{2}	51.58^{3}	40.35^{2}	$\mathbf{1.67}^{1}$
BR-J48	46.23^{5}	43.95^{5}	$98.62^{2.5}$	51.34^{4}	$\mathbf{74.26}^{1}$	36.71^{4}	3.58^{4}
RAkEL-J48	50.91^{3}	48.74^{3}	$98.62^{2.5}$	57.76^{3}	72.84^{2}	$\mathbf{41.04}^{1}$	2.42^{2}
MlkNN	53.26^{2}	51.62^{2}	94.11^{4}	$\mathbf{66.14}^{1}$	41.77^{4}	31.84^{5}	3.00^{3}

Σύμφωνα με τη μέση κατάταξη των αλγορίθμων, ο GMl-ASLCS κατατάσσεται πρώτος, ενώ αμέσως μετά ακολουθούν οι state-of-the-art αλγόριθμοι RAkEL-J48, MlkNN, και BR-J48. Στην τελευταία θέση βρίσκεται ο προκάτοχος του GMl-ASLCS, ο GMl-ASLCS$_0$.

Για τη διερεύνηση της στατιστικής σημαντικότητας των μετρούμενων διαφορών στην κατάταξη των αλγόριθμων, πραγματοποιήθηκε το μη παραμετρικό στατιστικό τεστ Friedman [Fri40], με παραμέτρους $k = 5$ και $N = 6$, το οποίο *απέρριψε τη μηδενική υπόθεση (null hypothesis-H_0)*, δηλαδή την υπόθεση ότι όλοι οι αλγόριθμοι έχουν ισοδύναμη απόδοση, για επίπεδο εμπιστοσύνης $\alpha = 0.05$. Για τον εντοπισμό των μεθόδων μεταξύ των οποίων υπάρχει στατιστικά σημαντική διαφορά απόδοσης, εκτελέστηκε η post-hoc δοκιμή Nemenyi [Nem63] σε επίπεδο εμπιστοσύνης $\alpha = 0.05$, η οποία αποκάλυψε ότι:

- *δεν υπάρχει στατιστικά σημαντική διαφορά στην απόδοση μεταξύ του GMl-ASLCS και των αντιπάλων του αλγορίθμων που δεν ανήκουν στην οικογένεια των ΜαΣΤ (RAkEL-J48, MlkNN, BR-J48)*

- *υπάρχει στατιστικά σημαντική διαφορά στην απόδοση του GMl-ASLCS σε σχέση με τον GMl-ASLCS$_0$.*

Σύγκριση των Αλγορίθμων με βάση την Ακριβή Ορθότητα

Ο Πίνακας 8.12 συνοψίζει τα ποσοστά της ακριβούς ορθότητας των υπό μελέτη αλγορίθμων για όλα τα χρησιμοποιούμενα σύνολα δεδομένων, για τη στρατηγική συμπερασμού IVal. Οι τιμές αντιστοιχούν σε αυτές των πειραμάτων που διενεργήθηκαν και απέδωσαν τις τιμές ακρίβειας της προηγούμενης παραγράφου. Μαζί με τις μετρικές επίδοσης του καθενός, αναφέρουμε και τις αντίστοιχες κατατάξεις ανά

σύνολο δεδομένων ως εκθέτες στην τιμή της μετρικής, τη μέση κατάταξη κάθε αλγορίθμου όπως αυτή προκύπτει από την εφαρμογή του τεστ Friedman στη στήλη με τίτλο "Κατάταξη", καθώς και την απόλυτη θέση των αλγορίθμων στην τελική κατάταξη, ως εκθέτη της τιμής της μέσης κατάταξης.

Πίνακας 8.12: Σύγκριση αλγορίθμων πολυκατηγορικής ταξινόμησης με βάση την ακριβή ορθότητα, με στρατηγική συμπερασμού IVal, στο σύνολο των υπό μελέτη προβλημάτων ταξινόμησης. Οι εκθέτες αναφέρονται στην κατάταξη του κάθε αλγορίθμου ανά σύνολο δεδομένων, σύμφωνα με το στατιστικό τεστ Friedman. Η στήλη με τίτλο "Κατάταξη" περιέχει τη συνολική κατάταξη του αλγορίθμου στην αντίστοιχη γραμμή, ενώ ο εκθέτης σημαίνει τη θέση του στην (απόλυτη) τελική κατάταξη.

Αλγόριθμοι	music	yeast	genbase	scene	medical	enron	Κατάταξη
GMl-ASLCS$_0$	23.11^{2}	5.01^{5}	84.16^{5}	30.27^{5}	27.93^{5}	$\mathbf{11.92}^{1}$	3.83^{5}
GMl-ASLCS	$\mathbf{34.39}^{1}$	11.36^{3}	96.99^{3}	53.18^{2}	41.40^{3}	6.91^{4}	2.67^{2}
BR-J48	6.83^{5}	6.83^{4}	$\mathbf{97.13}^{1.5}$	40.13^{4}	$\mathbf{65.12}^{1}$	8.64^{3}	3.08^{4}
RAkEL-J48	11.71^{4}	11.71^{2}	$\mathbf{97.13}^{1.5}$	50.75^{3}	64.03^{2}	10.71^{2}	$\mathbf{2.42}^{1}$
MlkNN	18.74^{3}	$\mathbf{18.74}^{1}$	90.64^{4}	$\mathbf{60.12}^{1}$	35.19^{4}	6.22^{5}	3.00^{3}

Σύμφωνα με τη μέση κατάταξη των αλγορίθμων, ο RAkEL-J48 κατατάσσεται πρώτος, ενώ ακολουθεί ο GMl-ASLCS στη δεύτερη θέση, με τους MlkNN και BR-J48 να ακολουθούν στην τρίτη και τέταρτη θέση αντίστοιχα. Ο GMl-ASLCS$_0$ κατατάσσεται και εδώ στην τελευταία θέση.

Για τη διερεύνηση της στατιστικής σημαντικότητας των μετρούμενων διαφορών στην κατάταξη των αλγορίθμων για τη μετρική της ακριβούς ορθότητας, πραγματοποιήθηκε το μη παραμετρικό στατιστικό τεστ Friedman, με παραμέτρους $k = 5$ και $N = 6$, όπως και στην προηγούμενη περίπτωση, το οποίο *δεν απέρριψε τη μηδενική υπόθεση*, για όλα τα επίπεδα εμπιστοσύνης. Συνεπώς, αυτό που συμπεραίνουμε είναι ότι δεν υπάρχει στατιστικά σημαντική διαφορά στις επιδόσεις των πέντε αλγορίθμων, όσον αφορά στη μετρική της ακριβούς ορθότητας.

Επίδραση των Διαφοροποιησεων στην Επίδοση του GMl-ASLCS

Σε αυτή την Ενότητα διερευνούμε δύο ειδών επιδράσεις:

1. αυτές των βασικών λειτουργιών που αντικαταστήσαμε ή εισηγάγαμε στον ορισμό του GMl-ASLCS ως πρωτότυπες λειτουργίες, ώστε να εξακριβωθεί η διαφορά που αυτές επιφέρουν στις επιδόσεις του GMl-ASLCS (Παρ. 6.2.1 και 6.3.2), και

2. αυτές των τροποποιήσεων που προτείναμε στην Εν. 6.5.

Στην πρώτη υπό μελέτη ομάδα επιδράσεων ανήκει η εξέταση των επιδόσεων του GMl-ASLCS α) με χρήση του Γενετικού Αλγορίθμου με Διασταύρωση Ενός Σημείου και β) απουσία της διαγραφής κανόνων από τα Match Sets, η οποία αναμένεται

ότι θα μειώσει τα επίπεδα μέσης κάλυψης δειγμάτων σε αυτά του GMl-ASLCS$_0$, με απροσδιόριστες όμως τιμές της τελικής ακρίβειας. Θα ονομάσουμε το πρώτο ΜαΣΤ GMl-ASLCS$_{sp\chi}$ και το δεύτερο GMl-ASLCS$_{!M}$.

Στην περίπτωση του GMl-ASLCS$_{!M}$, λόγω της περαιτέρω διαγραφής κανόνων, υπάρχουν δύο δυνατότητες: α) να θέσουμε το μέγεθος του πληθυσμού ίσο με αυτό των βασικών πειραμάτων (GMl-ASLCS$_{!Ms}$) ή β) να αυξήσουμε το μέγεθος του πληθυσμού ανάλογα με τον αριθμό των κανόνων που διαγράφηκαν από τη λειτουργία διαγραφής κανόνων από τα Match Sets (GMl-ASLCS$_{!Ma}$). Σε αυτή την περίπτωση το μέγεθος του πληθυσμού υπολογίζεται από τη σχέση

$$|P| = (\frac{M}{M + d + |P_0|} + 1) \cdot |P_0| \tag{8.1}$$

όπου M ο αριθμός των κανόνων που διαγράφηκαν από τη λειτουργία διαγραφής κανόνων από τα Match Sets, d ο αριθμός των κανόνων που διαγράφηκαν από την τυπική λειτουργία διαγραφής, με επιλογή κανόνων μέσω επιλογής ρουλέτας, και $|P_0|$ το μέγεθος του πληθυσμού που υπαγορεύεται από τη στήλη $|P|$ του Πίνακα 8.2. Το άθροισμα $S = M+d+|P_0|$ είναι ο συνολικός αριθμός κανόνων που παρήχθησαν από το ΜαΣΤ στο σύνολο της εκπαίδευσης, συνεπώς ο λόγος M/S εκφράζει το ποσοστό των κανόνων που διαγράφηκαν από τη λειτουργία διαγραφής κανόνων στα Match Sets. Κατά το ποσοστό αυτό αυξάνουμε το μέγεθος του πληθυσμού, ώστε εν τέλει να έχει παραχθεί ο ίδιος αριθμός κανόνων, χωρίς να έχει αφαιρεθεί κάποιος από τα Match Sets.

Στη δεύτερη υπό μελέτη ομάδα επιδράσεων ανήκει α) η τροποποίηση της μεθόδου υπολογισμού της καταλληλότητας στη συνιστώσα ενίσχυσης, η οποία περιγράφηκε στην Παρ. 6.5.1 (GMl-ASLCS$_f$), β) αυτή της μεθόδου υπολογισμού ανάθεσης πιθανότητας διαγραφής κανόνων, η οποία παρουσιάστηκε στην Παρ. 6.5.2 (GMl-ASLCS$_d$), γ) η μεταβολή των μεταβλητών ω, ϕ από $(\omega, \phi) \equiv (0.9, 1)$ σε $(\omega, \phi) \equiv (0, 0)$ (GMl-ASLCS$_{!\#}$) και δ) ο GMl-ASLCS που κάνει χρήση της λειτουργίας ομαδοποίησης για την αρχικοποίηση του πληθυσμού (Παρ. 6.5.4). Στην τελευταία περίπτωση μελετήθηκαν τρεις μέθοδοι ομαδοποίησης:

1. Η ομαδοποίηση με παραμέτρους $\gamma = 0.01$ και $Pcl_{\#A} = Pcl_{\#L} = 0$ (GMl-ASLCS$_{Cl.sp}$), ώστε σε προβλήματα με ανισορροπία συνδυασμών ετικετών, που απαιτούν την εύρεση συγκεκριμένων, ειδικών κανόνων, αυτοί να τους παρασχεθούν από τη λειτουργία ομαδοποίησης. Η τιμή $\gamma = 0.01$ θεωρείται, και είναι, εξαιρετικά μικρή σε σχέση με τις συνήθεις τιμές της παραμέτρου γ. Η χρήση μίας μικρής τιμής για την παράμετρο γ, θα έχει ως αποτέλεσμα το σχηματισμό ενός μικρού αριθμού συστάδων για κάθε διακριτό συνδυασμό ετικετών του συνόλου δεδομένων. Πιο συγκεκριμένα, ο αριθμός των συστάδων θα είναι ένα, για κάθε partition με λιγότερα από 100 δείγματα. Σε συνδυασμό με μηδενικές πιθανότητες γενίκευσης γνωρισμάτων και ετικετών για τους κανόνες που θα δημιουργηθούν μέσω ομαδοποίησης, για τα σύνολα δεδομένων που μελετάμε σε αυτή την εργασία, αναμένουμε την εύρεση ειδικών κανόνων, τόσο στο τμήμα συνθήκης όσο και στο τμήμα απόφασης, που θα περιγράφουν με περιεκτικό τρόπο το τμήμα του χώρου αναζήτησης που ενδεχομένως δε θα μπορούσε να εξερευνήσει ο GMl-ASLCS λόγω της ανισορροπίας συνδυασμού ετικετών.

2. Η ομαδοποίηση με παραμέτρους $\gamma = 0.2$ και $Pcl_{\#A} = Pcl_{\#L} = 0$, (GMl-ASLCS$_{Cl.lsp}$) για την αρχικοποίηση με περισσότερους, ειδικούς κανόνες που συμπυκνώνουν τη γνώση περισσότερων τμημάτων του χώρου αναζήτησης.

3. Η ομαδοποίηση με παράμετρο $\gamma = 0.2$ και πιθανότητες γενίκευσης αυτές που χρησιμοποιήθηκαν στα βασικά πειράματα του GMl-ASLCS (GMl-ASLCS$_{Cl.lge}$), σύμφωνα με τον Πίνακα 8.2, ώστε η αρχικοποίηση των κανόνων να γίνει σε μεγαλύτερο βαθμό από τη λειτουργία ομαδοποίησης και σε μικρότερο από τη λειτουργία κάλυψης. Έτσι, οι αρχικοί κανόνες θα βασίζονται σε μία καλύτερη περίληψη του συνόλου δεδομένων, αντί να εξαρτώνται από τη σειρά εισαγωγής των δειγμάτων του και την τυχαιότητα που εισάγει το Τμήμα Κάλυψης.

Συνοπτικά, η ονοματοδοσία των τροποποιήσεων παρουσιάζεται στον Πίνακα 8.13.

Πίνακας 8.13: Ονοματοδοσία των τροποποιήσεων του GMl-ASLCS.

Αλγόριθμοι	Αναφερόμενη Τροποποίηση του GMl-ASLCS
GMl-ASLCS$_{sp\chi}$	Αντικατάσταση του τελεστή Διασταύρωσης Δύο Τμημάτων από τον τελεστή Διασταύρωσης Ενός Σημείου.
GMl-ASLCS$_{!Ms}$	Αφαίρεση της λειτουργίας διαγραφής κανόνων από τα Match Sets, κρατώντας το μέγεθος του πληθυσμού ίδιο με αυτό που χρησιμοποιήθηκε στα πειράματα της Παρ. 8.2.1 (Πίνακας 8.2).
GMl-ASLCS$_{!Ma}$	Αφαίρεση της λειτουργίας διαγραφής κανόνων από τα Match Sets, αυξάνοντας το μέγεθος του πληθυσμού σύμφωνα με την Εξ. 8.1, σε αντιστοιχία με το μέγεθος του πληθυσμού (Πίνακας 8.2) για το κάθε σύνολο δεδομένων.
GMl-ASLCS$_{f}$	Αντικατάσταση του τρόπου υπολογισμού της καταλληλότητας ενός κανόνα, σύμφωνα με την Παρ. 6.5.1
GMl-ASLCS$_{d}$	Αντικατάσταση της μεθοδολογίας υπολογισμού πιθανοτήτων διαγραφής, σύμφωνα με την Παρ. 6.5.2.
GMl-ASLCS$_{!\#}$	Αντικατάσταση των παραμέτρων ϕ, ω με τις τιμές $(\phi, \omega) \equiv (0, 0)$ (Παρ. 6.5.3).
GMl-ASLCS$_{Cl.sp}$	Αρχικοποίηση του πληθυσμού μέσω ομαδοποίησης, με παραμέτρους $(\gamma, Pcl_{\#A}, Pcl_{\#L}) \equiv (0.01, 0, 0)$.
GMl-ASLCS$_{Cl.lsp}$	Αρχικοποίηση του πληθυσμού μέσω ομαδοποίησης, με παραμέτρους $(\gamma, Pcl_{\#A}, Pcl_{\#L}) \equiv (0.2, 0, 0)$.
GMl-ASLCS$_{Cl.lge}$	Αρχικοποίηση του πληθυσμού μέσω ομαδοποίησης, με παραμέτρους $\gamma = 0.2$ και πιθανότητες γενίκευσης που προσδιορίζονται στον Πίνακα 8.2.

Σύγκριση των Αλγορίθμων με βάση την ακρίβεια

Ο Πίνακας 8.14 συνοψίζει τα ποσοστά της ακρίβειας των υπό μελέτη αλγορίθμων για όλα τα χρησιμοποιούμενα σύνολα δεδομένων, για τη στρατηγική συμπερασμού IVal. Μαζί με τις μετρικές επίδοσης του καθενός, αναφέρουμε και τις αντίστοιχες κατατάξεις ανά σύνολο δεδομένων ως εκθέτες στην τιμή της μετρικής, τη μέση κατάταξη κάθε αλγορίθμου όπως αυτή προκύπτει από την εφαρμογή του τεστ Friedman στη στήλη με τίτλο "Κατάταξη", καθώς και την απόλυτη θέση των αλγορίθμων στην τελική κατάταξη ως εκθέτη της τιμής της μέσης κατάταξης.

Πίνακας 8.14: Σύγκριση του GMl-ASLCS και των τροποποιημένων αλγορίθμων πολυκατηγορικής ταξινόμησης με βάση την ακρίβεια, με στρατηγική συμπερασμού IVal, στο σύνολο των υπό μελέτη προβλημάτων ταξινόμησης. Οι εκθέτες αναφέρονται στην κατάταξη του κάθε αλγορίθμου ανά σύνολο δεδομένων, σύμφωνα με το στατιστικό τεστ Friedman. Η στήλη με τίτλο "Κατάταξη" περιέχει τη συνολική κατάταξη του αλγορίθμου στην αντίστοιχη γραμμή, ενώ ο εκθέτης σημαίνει τη θέση του στην (απόλυτη) τελική κατάταξη.

Αλγόριθμοι	music	yeast	genbase	scene	medical	enron	Κατάταξη
GMl-ASLCS	60.47^{2}	$\mathbf{51.67}^{1}$	$\mathbf{98.63}^{1}$	63.15^{3}	$\mathbf{51.58}^{1}$	$\mathbf{40.35}^{1}$	1.50^{1}
GMl-ASLCS$_{sp\chi}$	57.30^{9}	51.26^{4}	97.85^{9}	62.07^{4}	49.65^{2}	37.51^{4}	5.33^{4}
GMl-ASLCS$_{!Ms}$	54.80^{10}	48.35^{9}	98.28^{4}	55.42^{10}	45.68^{8}	34.04^{10}	8.00^{9}
GMl-ASLCS$_{!Ma}$	57.31^{8}	48.46^{8}	97.81^{10}	58.67^{9}	46.71^{6}	36.33^{9}	8.33^{10}
GMl-ASLCS$_{f}$	58.09^{5}	50.46^{7}	98.03^{6}	61.69^{6}	46.69^{7}	36.89^{6}	6.17^{6}
GMl-ASLCS$_{d}$	$57.73^{6.5}$	47.04^{10}	98.39^{3}	64.33^{2}	46.82^{5}	36.56^{7}	5.58^{5}
GMl-ASLCS$_{!\#}$	59.15^{4}	51.29^{3}	98.08^{5}	$\mathbf{65.31}^{1}$	47.42^{4}	38.63^{2}	3.17^{2}
GMl-ASLCS$_{Cl.sp}$	59.91^{3}	50.76^{6}	98.02^{7}	60.08^{8}	45.23^{9}	37.16^{5}	6.33^{7}
GMl-ASLCS$_{Cl.lsp}$	$\mathbf{60.51}^{1}$	51.41^{2}	98.46^{2}	61.91^{5}	48.15^{3}	36.49^{8}	3.50^{3}
GMl-ASLCS$_{Cl.lge}$	$57.73^{6.5}$	50.94^{5}	97.99^{8}	61.51^{7}	42.74^{10}	37.92^{3}	6.58^{8}

Σύμφωνα με τη μέση κατάταξη των αλγορίθμων, ο GMl-ASLCS κατατάσσεται πρώτος, ενώ αμέσως μετά ακολουθεί η τροποποίηση του με παραμέτρους $(\omega, \phi) \equiv (0, 0)$ και τρίτος κατατάσσεται ο GMl-ASLCS που χρησιμοποιεί ομαδοποίηση με παραμέτρους $(\gamma, Pcl_{\#A}, Pcl_{\#L}) \equiv (0.2, 0, 0)$.

Για τη διερεύνηση της στατιστικής σημαντικότητας των μετρούμενων διαφορών στην κατάταξη των αλγόριθμων, πραγματοποιήθηκε το μη παραμετρικό στατιστικό τεστ Friedman, με παραμέτρους $k = 10$ και $N = 6$, το οποίο *απέρριψε τη μηδενική υπόθεση* για επίπεδο εμπιστοσύνης $\alpha = 0.01$. Για τον εντοπισμό των μεθόδων μεταξύ των οποίων υπάρχει στατιστικά σημαντική διαφορά απόδοσης, εκτελέστηκε η post-hoc δοκιμή Nemenyi. Σε επίπεδο εμπιστοσύνης $\alpha = 0.05$, η δοκιμή αποκάλυψε ότι υπάρχει στατιστικά σημαντική διαφορά απόδοσης ανάμεσα α) στον GMl-ASLCS και τους GMl-ASLCS$_{Cl.lge}$, GMl-ASLCS$_{!Ms}$ και GMl-ASLCS$_{!Ma}$, και β) ανάμεσα στον GMl-ASLCS$_{!\#}$ και τον GMl-ASLCS$_{!Ma}$. Σε επίπεδο εμπιστοσύνης $\alpha = 0.10$, υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στον GMl-ASLCS και τους GMl-

ASLCS$_f$, GMl-ASLCS$_{Cl.sp}$, GMl-ASLCS$_{Cl.lge}$, GMl-ASLCS$_{!Ms}$ και GMl-ASLCS$_{!Ma}$. Παράλληλα υπάρχει στατιστικά σημαντική διαφορά ανάμεσα στον GMl-ASLCS$_{!\#}$ και τους GMl-ASLCS$_{!Ms}$ και GMl-ASLCS$_{!Ma}$.

Σύγκριση των Αλγορίθμων με βάση την Ακριβή Ορθότητα

Πίνακας 8.15: Σύγκριση του GMl-ASLCS και των τροποποιημένων αλγορίθμων πολυκατηγορικής ταξινόμησης με βάση την ακριβή ορθότητα, με στρατηγική συμπερασμού IVal, στο σύνολο των υπό μελέτη προβλημάτων ταξινόμησης. Οι εκθέτες αναφέρονται στην κατάταξη του κάθε αλγορίθμου ανά σύνολο δεδομένων, σύμφωνα με το στατιστικό τεστ Friedman. Η στήλη με τίτλο "Κατάταξη" περιέχει τη συνολική κατάταξη του αλγορίθμου στην αντίστοιχη γραμμή, ενώ ο εκθέτης σημαίνει τη θέση των αλγορίθμων στην (απόλυτη) τελική κατάταξη.

Αλγόριθμος	music	yeast	genbase	scene	medical	enron	Κατάταξη
GMl-ASLCS	34.39^{2}	11.36^{4}	$\mathbf{96.99}^{1}$	$53.18^{4.5}$	$\mathbf{41.40}^{1}$	$6.91^{3.5}$	2.67^{1}
GMl-ASLCS$_{sp\chi}$	30.35^{8}	10.93^{5}	95.48^{9}	50.42^{8}	36.59^{5}	5.35^{7}	7.00^{9}
GMl-ASLCS$_{!Ms}$	29.70^{10}	9.84^{9}	96.08^{4}	45.82^{10}	30.85^{8}	2.76^{10}	8.50^{10}
GMl-ASLCS$_{!Ma}$	30.39^{7}	10.59^{6}	95.33^{10}	47.58^{9}	37.21^{3}	7.08^{2}	$6.17^{7.5}$
GMl-ASLCS$_f$	32.08^{5}	10.46^{7}	95.62^{8}	52.76^{6}	31.94^{7}	3.63^{8}	6.83^{8}
GMl-ASLCS$_d$	30.89^{6}	6.14^{10}	96.53^{3}	54.35^{2}	35.50^{6}	3.11^{9}	6.00^{6}
GMl-ASLCS$_{!\#}$	32.65^{4}	$\mathbf{12.28}^{1}$	$95.92^{5.5}$	$\mathbf{56.27}^{1}$	37.67^{2}	$6.91^{3.5}$	2.83^{2}
GMl-ASLCS$_{Cl.sp}$	33.75^{3}	10.30^{8}	$95.92^{5.5}$	51.00^{7}	29.77^{9}	$\mathbf{7.25}^{1}$	5.58^{4}
GMl-ASLCS$_{Cl.lsp}$	$\mathbf{35.24}^{1}$	11.51^{2}	96.68^{2}	$53.18^{4.5}$	36.90^{4}	5.87^{6}	3.25^{3}
GMl-ASLCS$_{Cl.lge}$	30.26^{9}	11.56^{3}	95.78^{7}	53.68^{3}	24.50^{10}	6.74^{5}	$6.17^{7.5}$

Σύμφωνα με τη μέση κατάταξη των αλγορίθμων, ο GMl-ASLCS κατατάσσεται πρώτος, ενώ και εδώ αμέσως μετά ακολουθεί η τροποποίηση του με παραμέτρους $(\omega, \phi) \equiv (0, 0)$ και τρίτος κατατάσσεται ο GMl-ASLCS που χρησιμοποιεί ομαδοποίηση με παραμέτρους $(\gamma, Pcl_{\#A}, Pcl_{\#L}) \equiv (0.2, 0, 0)$.

Για τη διερεύνηση της στατιστικής σημαντικότητας των μετρούμενων διαφορών στην κατάταξη των αλγόριθμων, πραγματοποιήθηκε το μη παραμετρικό στατιστικό τεστ Friedman, με παραμέτρους $k = 10$ και $N = 6$, το οποίο *απέρριψε τη μηδενική υπόθεση* για επίπεδο εμπιστοσύνης $\alpha = 0.01$. Για τον εντοπισμό των μεθόδων μεταξύ των οποίων υπάρχει στατιστικά σημαντική διαφορά απόδοσης, εκτελέστηκε η post-hoc δοκιμή Nemenyi, σε επίπεδο εμπιστοσύνης $\alpha = 0.05$, η οποία αποκάλυψε ότι υπάρχει στατιστικά σημαντική διαφορά στην απόδοση α) ανάμεσα στον GMl-ASLCS και τον GMl-ASLCS$_{!Ms}$, β) ανάμεσα στον GMl-ASLCS$_{!\#}$ και τον GMl-ASLCS$_{!Ms}$ και γ) ανάμεσα στον GMl-ASLCS$_{Cl.lsp}$ και τον GMl-ASLCS$_{!Ms}$.

Σχόλια πάνω στα αποτελέσματα

Από τα παραπάνω συγκριτικά αποτελέσματα και τη στατιστική ανάλυσή τους συνάγουμε τα εξής συμπεράσματα:

- Ο GMl-ASLCS που χρησιμοποιεί τις τιμές παραμέτρων $\omega = 0$ και $\phi = 0$, και ο GMl-ASLCS που κάνει χρήση της ομαδοποίησης με παραμέτρους $\gamma = 0.2$, $Pcl_{\#A} = 0$ και $Pcl_{\#L} = 0$, επιδεικνύουν συνεπή και ακριβή συμπεριφορά, στα πρότυπα των επιδόσεων του θεμελιώδους GMl-ASLCS, κατατασσόμενοι συνολικά στη δεύτερη και τρίτη θέση πίσω από αυτόν, αντίστοιχα, και για τις δύο μετρικές ορθότητας.

- Ο GMl-ASLCS που χρησιμοποιεί τη Διασταύρωση Ενός Σημείου υπολείπεται σε επιδόσεις σε σχέση με τον GMl-ASLCS που χρησιμοποιεί τη Διασταύρωση Δύο Τμημάτων, κυρίως ως προς τη μετρική της ακριβούς ορθότητας, αλλά δεν εμφανίζει στατιστικά σημαντική διαφορά στις μετρικές της ακρίβειας ως προς αυτόν. Συνεπώς, η Διασταύρωση Δύο Τμημάτων φαίνεται ότι είναι αποτελεσματικότερη και υπερτερεί σε σχέση με τη Διασταύρωση Ενός Σημείου και για αυτό είναι προτιμότερη στην ενσωμάτωσή της στον GMl-ASLCS, αλλά και εν γένει σε πολυκατηγορικά ΜαΣΤ.

- Η αφαίρεση κανόνων με κριτήρια που τίθενται στα Match Sets είναι κρίσιμης σημασίας, όχι μόνο ως προς τις μετρικές της ακρίβειας και της ακριβούς ορθότητας, για τις οποίες τα δύο τροποποιημένα ΜαΣΤ, GMl-ASLCS$_{!Ms}$ και GMl-ASLCS$_{!Ma}$, εμφανίζουν στατιστικά σημαντικές διαφορές σε σχέση με τον GMl-ASLCS, αλλά και ως προς τη μετρική της μέσης κάλυψης δειγμάτων, όπως φαίνεται στον Πίνακα 8.16. Με άλλα λόγια, η λειτουργία διαγραφής κανόνων από τα Match Sets, στο πλαίσιο λειτουργίας του GMl-ASLCS είναι απολύτως απαραίτητη, τόσο για την αύξηση των επιδόσεων ως προς τις μετρικές της ακρίβειας και της ακριβούς ορθότητας, όσο και για την αύξηση των δειγμάτων που καλύπτουν οι κανόνες του τελικού μοντέλου του GMl-ASLCS.

- Η αρχικοποίηση του πληθυσμού μέσω ομαδοποίησης απαιτεί προσεκτικό χειρισμό των υπό μεταβολή παραμέτρων. Τα πειράματά μας δείχνουν ότι η καλύτερη συμπεριφορά του GMl-ASLCS επιτυγχάνεται για έναν μικρό ($\gamma = 0.2$), αλλά όχι πολύ μικρό ($\gamma = 0.01$), αριθμό συστάδων (διαφορά του GMl-ASLCS$_{Cl.lsp}$ με τον GMl-ASLCS$_{Cl.sp}$) και για μικρότερες τιμές πιθανοτήτων γενίκευσης γνωρισμάτων και ετικετών, στη συγκεκριμένη περίπτωση για $(P_{\#A}, P_{\#L}) \equiv (0, 0)$ (διαφορά του GMl-ASLCS$_{Cl.lsp}$ με τον GMl-ASLCS$_{Cl.lge}$).

- Ο GMl-ASLCS με τροποποιημένη μέθοδο υπολογισμού ανάθεσης πιθανοτήτων διαγραφής υπό συνθήκες μπορεί να οδηγήσει τον πληθυσμό των κανόνων σε μεγαλύτερα επίπεδα γενίκευσης (Πίνακας 8.16) από αυτά που απαιτούνται ώστε να υπάρχει ευσταθής ισορροπία ανάμεσα στη γενίκευση και την ακρίβεια του τελικού μοντέλου, υποσκάπτοντας τα επίπεδα ακρίβειας και Ακριβούς Ορθότητας, όπως φαίνεται στα αποτελέσματα του GMl-ASLCS$_d$ για το σύνολο yeast.

Στον Πίνακα 8.16 παρατίθενται ενδεικτικές τιμές της μέσης κάλυψης δειγμάτων (ως ποσοστού επί τοις εκατό) από τον GMl-ASLCS και τις τροποποιήσεις του GMl-ASLCS$_*$. Αξιοσημείωτο είναι το γεγονός ότι οι αλγόριθμοι που δεν κάνουν χρήση της λειτουργίας διαγραφής κανόνων από τα Match Sets (GMl-ASLCS$_{!M*}$), εμφανίζουν

σημαντική διαφορά, όσον αφορά στα επίπεδα μέσης κάλυψης δειγμάτων, σε σχέση με τους υπόλοιπους αλγορίθμους, οι οποίοι χρησιμοποιούν στο σύνολό τους την παραπάνω λειτουργία.

Πίνακας 8.16: Ενδεικτικές τιμές της μετρικής Μέσης Κάλυψης δειγμάτων (ποσοστό επί τοις εκατό) για τον GMl-ASLCS και τους τροποποιημένους αλγορίθμους πολυκατηγορικής ταξινόμησης GMl-ASLCS$_*$, στο σύνολο των υπό μελέτη προβλημάτων ταξινόμησης.

Αλγόριθμος	music	yeast	genbase	scene	medical	enron
GMl-ASLCS	1.7896	0.47329	3.0724	0.4411	3.2497	**0.5242**
GMl-ASLCS$_{sp\chi}$	1.6628	0.4664	2.9716	0.4276	3.0542	0.5129
GMl-ASLCS$_{!Ms}$	0.4036	0.1209	1.9306	0.1653	1.8301	0.2787
GMl-ASLCS$_{!Ma}$	0.5393	0.1295	1.9516	0.17659	1.5916	0.2470
GMl-ASLCS$_f$	1.7515	0.5357	3.0587	**0.4620**	**3.7674**	0.5141
GMl-ASLCS$_d$	1.7746	**1.0167**	**3.2365**	0.4425	2.9760	0.4945
GMl-ASLCS$_{!\#}$	1.7297	0.5080	3.0827	0.4140	3.0057	0.5041
GMl-ASLCS$_{Cl.sp}$	**1.9291**	0.5083	3.0375	0.4351	2.9339	0.4977
GMl-ASLCS$_{Cl.lsp}$	1.8003	0.4648	3.0854	0.3955	3.1086	0.5098
GMl-ASLCS$_{Cl.lge}$	1.6784	0.4546	3.10178	0.4459	3.3272	0.4983

Σύνοψη

Στο παρόν κεφάλαιο αξιολογήσαμε την ικανότητα ταξινόμησης του GMl-ASLCS σε έξι διαδεδομένα σύνολα πολυκατηγορικών δεδομένων, τα music, yeast, genbase, scene, medical και enron. Αρχικά, συγκρίναμε τις επιδόσεις του σε σχέση με τον GMl-ASLCS$_0$ και βρήκαμε πως για τη μετρική της ακρίβειας, ο GMl-ASLCS εμφανίζει στατιστικά σημαντική διαφορά σε σχέση με τον προκάτοχό του, βελτιώνοντας τις τιμές της παραπάνω μετρικής σε όλα τα πραγματικά σύνολα δεδομένων που εξετάσαμε (Παρ. 8.3.2). Αντίθετα, για τη μετρική της ακριβούς ορθότητας, οι GMl-ASLCS και GMl-ASLCS$_0$ δεν εμφανίζουν στατιστικά σημαντικές διαφορές (Παρ. 8.3.3).

Στη συνέχεια, συγκρίναμε τις επιδόσεις του GMl-ASLCS με αυτές των διαδεδομένων στη βιβλιογραφία αλγορίθμων πολυκατηγορικής ταξινόμησης, οι οποίοι μάλιστα δεν ανήκουν στην οικογένεια των ΜαΣΤ. Βρήκαμε πως αν και ο GMl-ASLCS κατατάσσεται πρώτος ανάμεσά τους για τη μετρική της ακρίβειας και δεύτερος για αυτή της ακριβούς ορθότητας, συνολικά, δεν διακρίνεται κάποια στατιστικά σημαντική διαφορά στις επιδόσεις τους για τις δύο αυτές μετρικές (Παρ. 8.3.2 και 8.3.3).

Όσον αφορά στην επίδραση του τροποποιημένου τελεστή διασταύρωσης, τα αποτελέσματα δείχνουν πως η πρωτύτερη υλοποίηση, η Διασταύρωση Ενός Σημείου, εμφανίζει χειρότερη συμπεριφορά από την υλοποίηση που επινοήθηκε ώστε να προσιδιάζει στη φύση της πολυκατηγορικής ταξινόμησης, τη Διασταύρωση Δύο

Τμημάτων, σε κάθε πολυκατηγορικό σύνολο δεδομένων που χρησιμοποιήθηκε. Η βελτιωμένη συμπεριφορά του νέου τελεστή διασταύρωσης αφορά στις μετρικές της ακρίβειας (Πίνακας 8.14), της ακριβούς ορθότητας (Πίνακας 8.15), αλλά και της μέσης κάλυψης δειγμάτων (Πίνακας 8.16).

Διαπιστώσαμε, επιπλέον, πως ο GMl-ASLCS βελτιώνει συνολικά, όχι μόνο την προβλεπτική του ικανότητα, αλλά και την ικανότητά του να γενικεύει με ακρίβεια πάνω στα δείγματα των συνόλων δεδομένων, σε μεγαλύτερο βαθμό από ότι ο GMl-ASLCS$_0$, λόγω της λειτουργίας διαγραφής κανόνων από τα Match Sets που επινοήσαμε (Πίνακες 8.9 και 8.10).

Παρατηρήσαμε πως η λειτουργία διαγραφής κανόνων από τα Match Sets είναι ένα αναντικατάστατο κομμάτι του αλγορίθμου GMl-ASLCS, όχι μόνο αυξάνοντας τον αριθμό δειγμάτων που καλύπτουν κατά μέσο όρο οι κανόνες ενός ΜαΣΤ (πιν. 8.16), όπως αναφέραμε παραπάνω, αλλά βελτιώνοντας τη συνολική προβλεπτική ικανότητα του μοντέλου που κατασκευάζει ο GMl-ASLCS σε κάθε πολυκατηγορικό πρόβλημα.

Στους Πίνακες 8.9 και 8.10 φαίνεται η σημαντική βελτίωση που επέφερε η απαγόρευση συμμετοχής κανόνων μηδενικής κάλυψης στον πληθυσμό ενός ΜαΣΤ, καθώς πλέον η εξελικτική διαδικασία καθίσταται σε μεγαλύτερο βαθμό ελέγξιμη και αδιάλειπτη και, συνεπώς, η εργασία του χρήστη ή/και ερευνητή τού επιτρέπει να εξάγει ακριβέστερα συμπεράσματα για την εν μέρει και εν συνόλω συμπεριφορά ενός ΜαΣΤ.

Τέλος, εξάγαμε ενδιαφέροντα συμπεράσματα συγκρίνοντας τον GMl-ASLCS με τις διάφορες τροποποιήσεις που εισάγαμε στην Εν. 6.5. Οι αλγόριθμοι που βασίζονται στον GMl-ASLCS και περιλαμβάνουν α) τη μεταβολή των παραμέτρων ω, ϕ από $(\omega, \phi) \equiv (0.9, 1)$ σε $(\omega, \phi) \equiv (0, 0)$ και β) την αρχικοποίηση του πληθυσμού ενός ΜαΣΤ μέσω της ομαδοποίησης, με $\gamma = 0.2$ και μηδενικές πιθανότητες γενίκευσης των γνωρισμάτων και ετικετών των κεντροειδών, αποδεικνύονται συνεπείς και, εν γένει, εύρωστοι (Πίνακας 8.14 και 8.14). Αν και οι επιδόσεις τους υπολείπονται αυτών του GMl-ASLCS για τα έξι πραγματικά σύνολα δεδομένων, περαιτέρω πιο στοχευμένα πειράματα ίσως φέρουν κάποιον από τους δύο παραπάνω αλγορίθμους σε μεγαλύτερα επίπεδα ακρίβειας από αυτόν.

ΜΕΡΟΣ ΙΙΙ

Συμπεράσματα & Μελλοντικές Επεκτάσεις

9
Συμπεράσματα

Στην παρούσα διπλωματική εργασία παρουσιάστηκε η δεύτερη και βελτιστοποιημένη μορφή του Μανθάνοντος Συστήματος Πολυκατηγορικής Ταξινόμησης GMl-ASLCS. Αρχικά περιγράψαμε τα τμήματα τα οποία είναι κοινά ανάμεσα στον GMl-ASLCS και τον GMl-ASLCS$_0$ και αφορούν:

- στην αναπαράσταση του τμήματος συνθήκης και απόφασης των κανόνων που χρησιμοποιούν τα δύο ΜαΣΤ
- στη βασισμένη-στην-εμπειρία έκπτωση της καταλληλότητας των κανόνων και τη λειτουργία του τμήματος Κάλυψης στη συνιστώσα Εξερεύνησης
- στη μεθοδολογία υπολογισμού του μέσου μεγέθους των Correct Sets στα οποία συμμετέχουν οι κανόνες των ΜαΣΤ
- στο Γενετικό Αλγόριθμο με Επιλογή Ρουλέτας και
- στη συνιστώσα Επίδοσης, η οποία διατηρείται αμετάβλητη.

Παρουσιάσαμε, σε ένα πρώτο επίπεδο, τον κύκλο εκπαίδευσης του GMl-ASLCS$_0$ και τις τροποποιήσεις που ήταν αναγκαίες, ώστε αυτός να είναι ικανός για πολυκατηγορική ταξινόμηση. Η παραπάνω περιγραφή και ανάλυση ήταν αναγκαία, ώστε ο αναγνώστης να λάβει τις απαραίτητες γνώσεις για τον τρόπο λειτουργίας ενός πολυκατηγορικού ΜαΣΤ, αλλά και για την κατανόηση των λόγων για τις τροποποιήσεις που επιφέραμε στον αρχικό GMl-ASLCS (GMl-ASLCS$_0$).

Στη συνέχεια, παρουσιάσαμε τον κύκλο εκπαίδευσης του GMl-ASLCS, του τροποποιημένου αλγορίθμου ΜαΣΤ που αποτελεί το κεντρικό θέμα αυτής της εργασίας και εμβαθύναμε στα επιμέρους τμήματα και συνιστώσες του και στις εσωτερικές διεργασίες τους.

Συγκεκριμένα, στη συνιστώσα Εξερεύνησης ανακαλύψαμε πως ο τελεστής Διασταύρωσης Ενός Σημείου επιβραδύνει σε ένα βαθμό την εκπαιδευτική διαδικασία στα πραγματικά σύνολα πολυκατηγορικών δεδομένων, λόγω της, σχεδόν νομοτελειακής, μεταφοράς του συνόλου των αποφάσεων για ετικέτες από τους κανόνες-γονείς προς τους απογόνους τους. Για αυτό το λόγο προτείναμε την εφαρμογή μίας πρωτότυπης μεθόδου διασταύρωσης, του τελεστή Διασταύρωσης Δύο Τμημάτων, ο οποίος προσιδιάζει στη φύση των πολυκατηγορικών προβλημάτων. Για κάθε σχηματιζόμενο ανά δείγμα i και ετικέτα l Correct Set, η Διασταύρωση Δύο Τμημάτων δρα μεταφέροντας από τους γονείς στους απογόνους μόνο την συγκεκριμένη απόφαση για την l, η οποία λόγω της φύσης του συνόλου Correct Set, συμφωνεί με την απόφαση του i για την l. Με αυτό τον τρόπο, επαναληπτικά, μειώνουμε τη συμμετοχή λανθασμένων αποφάσεων για ετικέτες στο τμήμα απόφασης των κανόνων και επιτυγχάνουμε τη γρηγορότερη σύγκλιση του Γενετικού Αλγορίθμου και τη συνολική αύξηση της προβλεπτικής ικανότητας του μοντέλου που κατασκευάζει ο GMl-ASLCS.

Ωστόσο, η προβλεπτική ικανότητα των κανόνων του ΜαΣΤ δεν είναι από μόνη της μία ικανή συνθήκη για αποτελεσματική ταξινόμηση από το προβλεπτικό μοντέλο που κατασκευάζει ο GMl-ASLCS. Ταυτόχρονα, το μοντέλο θα πρέπει να μπορεί να ταξινομεί με ακρίβεια δείγματα με τα οποία δεν έχει εκπαιδευτεί. Με άλλα λόγια, οι κανόνες ενός ΜαΣΤ θα πρέπει να είναι ικανοί να γενικεύουν με ακρίβεια βάσει των δειγμάτων με τα οποία εκπαιδεύεται το ΜαΣΤ, ώστε να μπορούν να καλύπτουν αταξινόμητα δείγματα από το σύνολο ελέγχου. Για αυτό το σκοπό, προτείναμε την εφαρμογή μίας μεθόδου που αφαιρεί από τον πληθυσμό τους κανόνες εκείνους που καλύπτουν το μικρότερο αριθμό δειγμάτων μέσα σε κάθε Match Set και που έχουν τη χαμηλότερη καταλληλότητα ανάμεσα στους κανόνες του Match Set που καλύπτουν τον ίδιο αριθμό δειγμάτων. Για να καταστεί δικαιότερο το παραπάνω σχήμα διαγραφής, απαιτήσαμε από τους κανόνες να έχουν εξετάσει τουλάχιστον μία φορά κάθε δείγμα του συνόλου εκπαίδευσης ως προς την ικανότητα κάλυψής του.

Ακόμα, παρατηρήσαμε πως στον υπολογισμό της καταλληλότητας ενός κανόνα, ο GMl-ASLCS$_0$ συμπεριφέρεται με ισότιμο τρόπο στις σαφείς συμφωνίες και τις αδιαφορίες ως προς την απόφαση για μία ετικέτα, κάτι που ενέχει κινδύνους τόσο για την εξελικτική διαδικασία, όσο και για την τελική ταξινόμηση των δειγμάτων του συνόλου ελέγχου. Για να αποθαρρύνουμε το ΜαΣΤ από το να συσσωρεύουν οι κανόνες του αδιαφορίες στο τμήμα απόφασής τους, εισήγαμε ένα σχήμα έκπτωσης (ή τιμωρίας) για κάθε ετικέτα για την οποία ένας κανόνας δεν αποφασίζει με σαφήνεια.

Η τελευταία ριζοσπαστική μας κίνηση είχε ως αφετηρία την παρατήρηση του φαινομένου της υπερ-παραγωγής κανόνων μηδενικής κάλυψης, δηλαδή κανόνων που είναι ανίκανοι να ταξινομήσουν έστω και ένα του συνόλου εκπαίδευσης, στα σύνολα *enron* και *medical*. Παρατηρήσαμε τη συμπιεστική δράση του παραπάνω φαινομένου, όσον αφορά στη διαφορά του αριθμού των κανόνων που επιθυμούμε να συγκρατεί το ΜαΣΤ ως πληθυσμό και την πραγματική του τιμή, αλλά και την επίδρασή του στις επιδόσεις του ΜαΣΤ. Η κίνησή μας ήταν να εξαλείψουμε την παρουσία κανόνων μηδενικής κάλυψης στον πληθυσμό του ΜαΣΤ, με τον άμεσο έλεγχο κάλυψης δειγμάτων για κάθε απόγονο που δημιουργείται από το Γενετικό

Αλγόριθμο, ώστε να παρέχουμε ένα έρεισμα για την ομαλότητα της εξελικτικής διαδικασίας και να μπορούμε να προσδιορίσουμε με ακρίβεια τις παραμέτρους της εκπαίδευσης, ώστε στα πειράματα που διενεργούμε ο GMl-ASLCS να έχει ελέγξιμη συμπεριφορά και βέλτιστες επιδόσεις.

Γενικότερα, η εργασία στον πολυκατηγορικό χώρο με τη χρήση ΜαΣΤ, εκτός από την εισαγωγή των παραπάνω λειτουργιών, κυοφόρησε μία σειρά από παρατηρήσεις και συμπεράσματα, όπως το γιατί δεν πρέπει κανόνες που αδιαφορούν για μία ετικέτα να συμμετέχουν στο αντίστοιχο Correct Set το οποίο σχηματίζεται για αυτήν, το πώς επηρεάζει τη συμπεριφορά ενός ΜαΣΤ η μεταβολή του κατωφλίου εμπειρίας θ_{exp} και των βασικών παραμέτρων του $(|I|, \theta_{GA}, maxPopulationSize, P_{\#A})$, ή ακόμα το πώς ο υπολογισμός της παραμέτρου cs προκαλεί την υπερ-εκτίμηση και υπο-εκτίμησή της πραγματικής της τιμής για κάθε κανόνα, ανάλογα με την εμπειρία του και το χρονικό σημείο δημιουργίας του.

Στη συνέχεια, προτείναμε μερικές τροποποιήσεις σε επιμέρους τμήματα του GMl-ASLCS που θεωρήσαμε ως σκόπιμες και άξιες προς διερεύνηση της ικανότητάς τους να τον βοηθήσουν στο έργο της ταξινόμησης. Μεταξύ αυτών βρίσκεται η μεταβολή του τρόπου υπολογισμού της καταλληλότητας ενός κανόνα, η τροποποίηση της μεθόδου υπολογισμού των πιθανοτήτων διαγραφής των κανόνων μέσω επιλογής ρουλέτας και η αρχικοποίηση του πληθυσμού μέσω ομαδοποίησης.

Στο πειραματικό μέρος της εργασίας, πραγματοποιήσαμε μία πρώτη αξιολόγηση της συμπεριφοράς του GMl-ASLCS αλλά και της συνεισφοράς των τεσσάρων αρθρωτών τροποποιήσεων που επιφέραμε στον αρχικό GMl-ASLCS, σε τέσσερα τεχνητά σύνολα δεδομένων. Οι τροποποιήσεις αυτές αφορούν στη μεταβολή του τρόπου υπολογισμού των πιθανοτήτων που ανατίθενται στους κανόνες του πληθυσμού, στον τελεστή Διασταύρωσης Δύο Τμημάτων, στην αφαίρεση κανόνων του πληθυσμού μέσω της επιλεκτικής διαγραφής κανόνων από τα Match Sets και στην έκπτωση καταλληλότητας των κανόνων για κάθε παρουσία αδιαφορίας στο τμήμα απόφασής τους. Σε γενικές γραμμές, η συνεισφορά κάθε επιμέρους λειτουργίας που εισάγαμε ή τροποποιήσαμε αποτιμάται θετικά πάνω στα τέσσερα τεχνητά σύνολα δεδομένων ($mlPosition_7$, $mlIdentity_7$, $adder_7^3$ και $adder_7^{24}$), με την κάθε μία να βελτιώνει τις επιδόσεις του GMl-ASLCS$_0$. Εξαίρεση αποτελεί η υλοποίηση της τιμωρίας κανόνων για την παρουσία αδιαφοριών στο τμήμα απόφασης των κανόνων, όταν η πολυκατηγορική πυκνότητα ενός συνόλου δεδομένων εμφανίζει υψηλές τιμές.

Παρ' όλα αυτά, η άθροιση των τεσσάρων παραπάνω συνιστωσών στον ορισμό του GMl-ASLCS τον κάνουν να εμφανίζει συνολικά βελτιωμένες επιδόσεις σε σχέση με αυτές του GMl-ASLCS$_0$ και στα τέσσερα παραπάνω τεχνητά σύνολα, για τη μετρικής της ακρίβειας. Επιπρόσθετα, η ικανότητα ανάπτυξης του Χάρτη Βέλτιστων Αποφάσεων (ΧΒΑ) από τον GMl-ASLCS βελτιώνεται, μόνο για τα σύνολα εκείνα στα οποία δεν υπάρχει παρουσία αδιαφοριών για ετικέτες στους κανόνες του. Λόγω της τιμωρίας της καταλληλότητας των κανόνων που διατηρούν ετικέτες με αδιαφορίες στο τμήμα απόφασής τους, ο GMl-ASLCS εμφανίζει μεγαλύτερη ικανότητα εύρεσης του ΧΒΑ για το σύνολο $mlPosition_7$, του οποίου ο ΧΒΑ περιλαμβάνει κανόνες που αποφασίζουν με σαφήνεια για τις ετικέτες του προβλήματος, αλλά αποτυγχάνει πλήρως να αναπτύξει τον ΧΒΑ του προβλήματος $mlIdentity_7$. Βεβαίως, αυτό είναι

κάτι το αναμενόμενο και δεν επηρεάζει σε κανένα βαθμό τη συνολική προβλεπτική ικανότητα του GMl-ASLCS, καθώς η ικανότητα εύρεσης ενός XBA είναι μόνο μία ικανή συνθήκη για αποτελεσματική ταξινόμηση, αλλά όχι αναγκαία.

Στη συνέχεια πραγματοποιήσαμε μία περισσότερο ρεαλιστική αξιολόγηση του GMl-ASLCS σε έξι, διαδεδομένα στη βιβλιογραφία, πραγματικά σύνολα πολυκατηγορικών δεδομένων, τα *music*, *yeast*, *genbase*, *scene*, *medical* και *enron*. Συγκρίναμε τις επιδόσεις του GMl-ASLCS με αυτές του GMl-ASLCS$_0$ και τριών state-of-the-art πολυκατηγορικών αλγορίθμων (οι οποίοι δεν ανήκουν στην οικογένεια των ΜαΣΤ) στα παραπάνω σύνολα δεδομένων και βρήκαμε πως ο GMl-ASLCS κατατάσσεται πρώτος ανάμεσά τους για τη μετρική της ακρίβειας και δεύτερος για τη μετρικής της ακριβούς ορθότητας. Επιπρόσθετα, ο GMl-ASLCS εμφανίζει στατιστικά σημαντικές διαφορές στις επιδόσεις του με βάση την ακρίβεια σε σχέση με τον GMl-ASLCS$_0$. Αντιθέτως, για τη μετρική της ακριβούς ορθότητας, βρέθηκε ότι για τους πέντε υπό μελέτη αλγόριθμους δεν υπάρχει στατιστικά σημαντική διαφορά στις επιδόσεις τους.

Ακόμα, συγκρίναμε τις επιδόσεις του GMl-ASLCS, και του τροποποιημένου GMl-ASLCS που χρησιμοποιεί τον τελεστή Διασταύρωσης Ενός Σημείου και βρήκαμε πως αν και οι δυο αλγόριθμοι δεν εμφανίζουν στατιστικά σημαντική διαφορά όσον αφορά στις μετρικές της ακρίβειας και της ακριβούς ορθότητας, ο δεύτερος κατατάσσεται συστηματικά σε χαμηλότερες θέσεις από τον πρώτο, συνεπώς, στο πλαίσιο λειτουργίας του GMl-ASLCS, ο τελεστής Διασταύρωσης Δύο Τμημάτων κρίνεται καταλληλότερος και αποτελεσματικότερος από αυτόν που χρησιμοποιούσε ο GMl-ASLCS$_0$.

Όσον αφορά στο στόχο μας για αύξηση του αριθμού των δειγμάτων που καλύπτουν οι κανόνες του ΜαΣΤ, με ταυτόχρονη διατήρηση της προβλεπτικής του ικανότητας, η εισαγωγή του νέου τμήματος διαγραφής κανόνων με κριτήρια που τίθενται στα Match Sets φαίνεται ότι τον επιτυγχάνει, αυξάνοντας μάλιστα και τις επιδόσεις του GMl-ASLCS, όσον αφορά στις βαρυσήμαντες μετρικές της ακρίβειας και της ακριβούς ορθότητας, για τα έξι πραγματικά σύνολα πολυκατηγορικών δεδομένων που μελετήσαμε.

Επιπρόσθετα, αξιολογήσαμε την ικανότητα ταξινόμησης των τροποποιήσεων που προτείναμε στον ορισμό του GMl-ASLCS. Αποδείχθηκε ότι δύο από αυτές τις τροποποιήσεις βρίσκονται πολύ κοντά στις επιδόσεις του GMl-ASLCS και, συνεπώς, ότι αξίζει η περαιτέρω έρευνά και εντρύφηση στις εσωτερικές διεργασίες των τροποποιημένων GMl-ASLCS. Οι τροποποιήσεις αυτές αφορούν στην αρχικοποίηση του πληθυσμού του ΜαΣΤ μέσω της ομαδοποίησης του συνόλου εκπαίδευσης, με παραμέτρους $(\gamma, P_{\#A}, P_{\#L}) \equiv (0.2, 0, 0)$ και στο μη υπολογισμό των αδιαφοριών για ετικέτες στη συνάρτηση υπολογισμού της καταλληλότητας των κανόνων του ΜαΣΤ, με παραμέτρους $(\omega, \phi) \equiv (0, 0)$.

Τέλος, η εισαγωγή του τμήματος διαγραφής κανόνων από τα Match Sets και η πρόνοια για την απαγόρευση εισαγωγής κανόνων μηδενικής κάλυψης στον πληθυσμό των ΜαΣΤ, λόγω της ανεξαρτησίας τους από τον αριθμό των ετικετών του προβλήματος προς επίλυση, θεωρούμε ότι μπορεί να εφαρμοστεί με αντίστοιχη επιτυχία για την επίλυση προβλημάτων μονοκατηγορικής ταξινόμησης.

Περιορισμοι και Πλεονεκτηματα των ΜαΣΤ

Οι αλγόριθμοι πολυκατηγορικής ταξινόμησης με ΜαΣΤ φαίνεται ότι είναι ικανοί να παρέχουν ικανοποιητικότερα αποτελέσματα σε σχέση με τις ντετερμινιστικές προσεγγίσεις της βιβλιογραφίας. Υπάρχουν, όμως, και περιορισμοί στη λειτουργία τους και στα αποτελέσματα που εξάγουν. Μερικοί από αυτούς είναι:

- Ο χρόνος εκπαίδευσης των ΜαΣΤ είναι πολύ μεγαλύτερος σε σχέση με τις αιτιοκρατικές προσεγγίσεις και, μάλιστα, αυξάνει όσο αυξάνει το μέγεθος του συνόλου εκπαίδευσης, η πολυπλοκότητά του (ο αριθμός των γνωρισμάτων, ετικετών και η πολυκατηγορική του πυκνότητα του) και ο αριθμός των κανόνων που απαιτούνται για τη λύση του προβλήματος. Ενδεικτικά, σε υπολογιστή με 2.66 GHz επεξεργαστικής ισχύος και 4 GB μνήμη, οι χρόνοι εκπαίδευσης του GMl-ASLCS στα πραγματικά σύνολα δεδομένων που χρησιμοποιήθηκαν σε αυτή την εργασία ήταν 280 λεπτά για το σύνολο *music*, 4120 για το *yeast*, 668 για το *genbase*, 350 για το *scene*, 640 για το *medical* και 1100 λεπτά για το *enron*, χρησιμοποιώντας αξιολόγηση 10-πλης διασταυρωμένης επικύρωσης για τα *music*, *yeast* και *enron* και προϋπάρχοντα χωρισμό σε σύνολα εκπαίδευσης και ελέγχου για τα *scene*, *medical* και *enron*. Επιπρόσθετα, ο χρόνος εύρεσης του κατωφλίου με εσωτερική αξιολόγηση στη συνιστώσα Επίδοσης αυξάνει με το μέγεθος των κανόνων και των ετικετών του προβλήματος, σε σημείο τέτοιο όπου σε σύνολα δεδομένων όπως το *enron*, ο χρόνος για την εύρεσή του να παίρνει 2 ώρες.

- Τα ΜαΣΤ διατηρούν ένα μεγάλο σύνολο παραμέτρων που πρέπει να ρυθμιστούν κατάλληλα ώστε να διεξάγουν αποτελεσματική ταξινόμηση. Σε συνδυασμό με την πολυπλοκότητα που εισάγει η ρύθμιση κάθε παραμέτρου ξεχωριστά και την αλληλεξάρτηση των παραμέτρων, όσον αφορά στη συμπεριφορά ενός ΜαΣΤ κατά την εκπαιδευτική διαδικασία, αλλά και την τελική του ικανότητα ταξινόμησης, ο χρόνος για την εξαγωγή ικανοποιητικών αποτελεσμάτων αυξάνει δραματικά.

- Κάποια σύνολα δεδομένων, όπως το *scene* και το *medical* που εξετάσαμε σε αυτή την εργασία, απαιτούν πολύ περισσότερες επαναλήψεις μάθησης από ότι άλλα σύνολα. Η αραιότητα των δειγμάτων τους συνηγορεί στην αργή σύγκλιση του Γενετικού Αλγορίθμου και κάνει απαραίτητη την επέκταση της εκπαιδευτικής διαδικασίας και, άρα, του χρόνου εκπαίδευσης ενός ΜαΣΤ.

- Το θεωρητικό υπόβαθρο που μπορεί να βοηθήσει στην πράξη την κατηγοριοποίηση με ΜαΣΤ είναι ακόμη μικρό, λόγω της στοχαστικής φύσης τους και το γεγονός πως συντελείται μικρότερο μέγεθος έρευνας προς την κατεύθυνση των ΜαΣΤ σε σχέση με άλλες μεθόδους κατηγοριοποίησης.

- Η έλλειψη των απαραίτητων εργαλείων για την επαρκέστερη γνώση του τι πραγματικά συμβαίνει στο εσωτερικό των διεργασιών ενός ΜαΣΤ, λόγω βεβαίως και της στοχαστικής τους φύσης, κάνει δύσκολη τη διείσδυσή της ανά-

λυσής μας σε βάθος, ώστε να αποκτήσουμε μία πληρέστερη εικόνα για τα διάφορα τμήματά του και να προχωρήσουμε μακρύτερα την έρευνα στον τομέα τους.

- Η αποτύπωση της πορείας των μετρικών αξιολόγησης ενός ΜαΣΤ και, άρα της προόδου του με το πέρασμα των επαναλήψεων μάθησης, μπορεί να γίνει μόνο σε σύνολα δεδομένων με μικρή πολυπλοκότητα, όπως στα τεχνητά σύνολα δεδομένων που εξετάσαμε. Η πορεία της μετρικής της ακρίβειας, για παράδειγμα, με ρύθμιση κατωφλίου με τη μέθοδο PCut, για κάθε επανάληψη μάθησης είναι χρονοβόρα σε τέτοιο βαθμό που τριπλασιάζει το χρόνο εκπαίδευσης. Ακόμα περισσότερο, η χρήση της εύρεσης κατωφλίου με τη μέθοδο εσωτερικής αξιολόγησης είναι απαγορευτική.

Αντιθέτως, όμως, με τις αιτιοκρατικές προσεγγίσεις της πολυκατηγορικής ταξινόμησης, τα ΜαΣΤ περιλαμβάνουν ένα μεγάλο φάσμα από διαφορετικές λειτουργίες, με χώρο για τη μελέτη, τη βελτίωσή και την παραλληλοποίησή τους, αλλά και πρόσφορο έδαφος για την εισαγωγή νέων τμημάτων των οποίων οι λειτουργίες βασίζονται στην ανάλυση των εσωτερικών διεργασιών και δομών ενός ΜαΣΤ. Η μη-αιτιοκρατική τους φύση μπορεί να φαίνεται ως εμπόδιο όσον αφορά στην αναλυτική προσιτότητα τους, ταυτόχρονα, όμως, είναι το στοιχείο εκείνο που τα κάνει ευέλικτα και εύρωστα. Η βασισμένη-στη-φύση προσέγγισή τους, με τη θεωρία της εξέλιξης των ειδών να αποτελεί τον πυλώνα της δημιουργίας λύσεων πάνω σε προβλήματα, καθιστά τα ΜαΣΤ πλήρως ουσιώδη και ικανά να προσεγγίσουν με ακρίβεια σχεδόν ό,τι πρόβλημα μπορεί να περιγραφεί με τη μορφή μίας συλλογής δεδομένων.

10
Μελλοντικές Επεκτάσεις

Το θέμα της παρούσας διπλωματικής εργασίας επιδέχεται πολλών μελλοντικών επεκτάσεων, αλλά και διαφορετικών προσεγγίσεων. Περαιτέρω τροποποιήσεις από αυτές που περιγράφονται σε αυτό το κεφάλαιο μπορεί να βρεί κανείς στο [Mil11]. Παρακάτω παραθέτουμε μερικές από τις τροποποιήσεις και επεκτάσεις του αλγορίθμου εκπαίδευσης που η ανάλυση του GMl-ASLCS μας έκανε να θεωρήσουμε ως σημαντικές.

Μη επιμονη ενημερωση παραμετρων των κανονων του ΜαΣΤ

Μετά το σχηματισμό του Match Set για ένα δεδομένο δείγμα εκπαίδευσης, ακολουθεί η ενημέρωση των παραμέτρων του κάθε κανόνα που συμμετέχει σε αυτό. Αυτή η διαδικασία είναι διαρκής: οι παράμετροι ενός κανόνα ανανεώνονται αέναα για κάθε Match Set στο οποίο συμμετέχει, μέχρι το τέλος της διαδικασίας εκπαίδευσης. Θα ήταν, ίσως, άσκοπο, και σίγουρα περισσότερο χρονοβόρο, να ενημερώνουμε τις παραμέτρους που δεν άπτονται, άμεσα ή έμμεσα, της αντικειμενικής αξιολόγησης του κανόνα, για το διάστημα στο οποίο του παρουσιάζεται το σύνολο δεδομένων D για πολλοστή φορά.

Πιο συγκεκριμένα, το σύστημα αποκτά μία μερική εικόνα της ποιότητας του κανόνα, μέσω των μεταβλητών $truepositive(tp)$, $matchsetappearances(msa)$, $accuracy = tp/msa$ και $fitness = f(tp/msa)$, αφού αυτός αξιολογηθεί για κάθε δείγμα που καλύπτει. Παρακάτω εξετάζουμε την κάθε μία.

- tp: Για κάθε *label* l για το οποίο συμφωνεί ο κανόνας με το δείγμα, η μεταβλητή tp αυξάνεται ισόποσα με το msa, στην περίπτωσή μας, κατά ένα. Αν ο κανόνας αδιαφορεί για την l, τιμωρείται και το μέγεθος tp αυξάνεται κατά μία θετική ποσότητα, πάντα μικρότερη από αυτήν με την οποία αυξάνεται εάν συμφωνεί. Στην περίπτωση που διαφωνεί, μένει ως έχει. Όταν ο κανόνας εξετάσει όλα τα δείγματα που μπορεί να καλύψει από το D για πρώτη φορά, η τιμή της

μεταβλητής θα ανήκει στο διάστημα $[0, coveredInstances \cdot labels]$. Από εκεί και έπειτα, η τιμή αυτή θα N-πλασιάζεται στο τέλος της N-ιοστής φοράς που ο κανόνας εξετάζει όλα τα δείγματα που καλύπτει. Όλη η χρήσιμη πληροφορία, όμως, έχει εξαχθεί ήδη από το τέλος της πρώτης φοράς που του παρουσιάζεται το D.

- msa: Το msa αυξάνεται *ακριβώς* κατά ένα σε κάθε περίπτωση. Αντίστοιχα με το tp, η τιμή του N-πλασιάζεται στο τέλος της N-ιοστής φοράς που ο κανόνας εξετάζει όλα τα δείγματα του D που καλύπτει.

- $accuracy$ και $fitness$: Την πρώτη φορά που ο κανόνας θα εξετάσει όλα τα δείγματα του D, θα αποκαλυφθεί η εικόνα της ποιότητας του κανόνα, μέσω του αντικειμενικού προσδιορισμού της ακρίβειας και της καταλληλότητάς του. Η τιμή των δύο μεταβλητών θα βρίσκεται στο διάστημα $[0, 1]$. Με εξαίρεση τους κανόνες που διαθέτουν $accuracy = fitness = 1$ ή 0, η τιμή αυτών των μεταβλητών κατά την ενημέρωση των παραμέτρων τους σε μία τυχαία επανάληψη θα ταλαντώνεται γύρω από την πραγματική τους τιμή: αυτή που απέκτησαν στο τέλος της πρώτης φοράς που ο κανόνας εξέτασε όλα τα δείγματα του συνόλου δεδομένων.

  Θεωρητικά, υπάρχουν περιπτώσεις στις οποίες η τοποθέτηση των δειγμάτων μέσα στο D, δηλαδή η σειρά με την οποία παρουσιάζονται τα δείγματά του στο ΜαΣΤ, μπορεί να αναπτύξει συνθήκες όπου ένας κανόνας υποσκάπτει έναν άλλο (για παράδειγμα στην επιλογή για γονέα στο γενετικό αλγόριθμο) σε ένα χρονικό σημείο όπου δύο κανόνες έχουν λανθασμένη εκτίμηση για την καταλληλότητά τους, με τον έναν να την υπερεκτιμά και τον άλλον να την υποεκτιμά. Ίσως, από την άλλη, αυτό προσδίδει μία ευελιξία στην επιλογή γονέων για το γενετικό αλγόριθμο αυξάνοντας την ποικιλομορφία των παραγόμενων κανόνων. Ίσως ακόμα έχει θετικές δράσεις σε κομμάτια του συστήματος που δεν μπορούμε να προβλέψουμε λόγω της στοχαστικής του φύσης. Σε κάθε περίπτωση όμως, θεωρούμε ότι η αμεταβλητότητα αυτών των μεγεθών πέραν της πρώτης παρουσίασής του συνόλου δεδομένων στο ΜαΣΤ, παρέχουν μία αντικειμενική γνώμη στο σύστημα για τους κανόνες που εξελίσσει.

Συνολικά, θεωρούμε ότι η πειραματική αξιολόγηση της ανανέωσης των παραμέτρων tp, msa και $fitness = f(tp, msa)$ των κανόνων ενός ΜαΣΤ μόνο για τα πρώτα $|D|$ δείγματα που παρουσιάζονται σε αυτόν μετά τη δημιουργία του, θα άξιζε λόγω της μείωσης του χρόνου εκπαίδευσης και της "ορθολογικότητας" που εισάγει. Επιπλέον, αυτός ο τρόπος ανανέωσης είναι απαραίτητος για την επέκταση που προτείνουμε στην επόμενη παράγραφο.

Αντικατασταση της μεθοδου υπολογισμου της Ακριβειας των Κανονων

Η ακρίβεια των κανόνων, όπως έχουμε δει, βασίζεται στην αύξηση των μεγεθών tp και msa κατά τις ποσότητες ω και ϕ αντίστοιχα, για κάθε δείγμα που καλύπτουν και κάθε ετικέτα l.

$$accuracy(t+1) = \frac{tp(t) + \omega}{msa(t) + \phi} \tag{10.1}$$

Μία τολμηρή κίνηση, που ξεφεύγει από το ακριβειο-κεντρικό πλαίσιο λειτουργίας των ΜαΣΤ, είναι η αντικατάσταση της ακρίβειας, όπως αυτή υπολογίζεται με τον παραπάνω τρόπο, με ένα μέγεθος που αντικαθιστά την πρόσθεση με τον πολλαπλασιασμό. Για αντίστοιχες μεταβλητές ω' και ϕ', η καταλληλότητα ενός κανόνα μπορεί τότε να πάρει τη μορφή

$$fitness(t+1) = \left(\frac{tp(t) \cdot \omega'}{msa(t) \cdot \phi'} \right)^{\nu} \tag{10.2}$$

Όπως αναφέραμε στην προηγούμενη παράγραφο, αυτή η μέθοδος υπολογισμού είναι σημαντικό να χρησιμοποιείται με ενημέρωση των παραμέτρων tp και msa μόνο κατά την πρώτη φορά που παρουσιάζεται το σύνολο δεδομένων D σε έναν κανόνα, λόγω της βιαιότητας με την οποία μπορεί να επιδράσει η πράξη του πολλαπλασιασμού στις τιμές της καταλληλότητας των κανόνων.

Σε αυτόν τον τρόπο, η ρύθμιση των (ω', ϕ') θα πρέπει να είναι πολύ περισσότερο προσεκτική από αυτή των (ω, ϕ) και η τιμή της παραμέτρου ω' σίγουρα πολύ υψηλότερη από την τιμή 0.9, ίσως στα επίπεδα $\omega' = 0.99$ για $\phi' = 1$. Από αυτή τη μέθοδο δεν είναι σίγουρο ότι μπορούμε να περιμένουμε κάτι συγκλονιστικό, αν και σε μερικά σύνολα δεδομένων, όπως το τεχνητό $mlPosition_7$ έχει δείξει ότι το ΜαΣΤ συγκλίνει γρηγορότερα και καλύτερα από την ακριβειοκεντρική προσέγγιση του GMl-ASLCS.

Σε κάθε περίπτωση, η προσπάθεια για ριζοσπαστική σκέψη για τροποποίηση, ακόμα και των πιο βασικών τμημάτων ενός ΜαΣΤ, αξίζει γιατί μπορούν να εξαχθούν ενδιαφέροντα συμπεράσματα, όχι μόνο για την ίδια τη λειτουργία της μεθόδου τροποποίησης, αλλά και για την εν γένει συμπεριφορά ενός ΜαΣΤ.

Περι του θ_{GA}

Το κατώφλι εμπειρίας θ_{GA}, όπως έχουμε ξαναναφέρει, υποδεικνύει το ρυθμό με τον οποίο παράγονται οι κανόνες απόγονοι μέσω του Γενετικού Αλγορίθμου. Πέραν αυτού, για δεδομένο συνδυασμό

$$(|I|, \theta_{GA}, maxPopulationSize, P_{\#A})$$

όπου $|I|$ ο αριθμός επαναλήψεων μάθησης, $maxPopulationSize$ το μέγιστο πλήθος των μικρο-κανόνων του πληθυσμού και $P_{\#A}$ η πιθανότητα γενίκευσης γνωρισμάτων των κανόνων που παράγονται από το τμήμα Κάλυψης, το κατώφλι θ_{GA} ρυθμίζει έμμεσα το βαθμό γενίκευσης των κανόνων, τόσο μέχρι ο αριθμός τους να φτάσει το μέγιστο $maxPopulationSize$, όσο και από εκεί και έπειτα, μέχρι το τέλος των επαναλήψεων εκπαίδευσης.

Όλα τα ΜαΣΤ, από όσο γνωρίζουμε, χρησιμοποιούν μία και μοναδική σταθερή τιμή για αυτή τη μεταβλητή. Στην προσπάθειά μας να ρυθμίσουμε με ακριβέστερο τρόπο το βαθμό γενίκευσης των κανόνων ενός ΜαΣΤ, θα μπορούσαμε να χρησιμοποιήσουμε αντί για μία τιμή, δύο. Μία τιμή θ_{GA_0} μέχρι ο αριθμός των κανόνων να φτάσει το μέγιστο αριθμό τους, ώστε να οδηγήσουμε τον πληθυσμό στο επιθυμητό επίπεδο μέσης κάλυψης δειγμάτων, το οποίο θα είναι και το μέγιστό και μία δεύτερη θ_{GA_1}, εν γένει διαφορετική από την πρώτη, ώστε στο τέλος της εκπαιδευτικής διαδικασίας ο πληθυσμός να έχει φτάσει στο τελικό επίπεδο μέσης κάλυψης που έχουμε θέσει ως στόχο.

$$\theta_{GA} = \begin{cases} \theta_{GA_0}, & rouletteWheelDeletionsCommenced = false \\ \theta_{GA_1}, & rouletteWheelDeletionsCommenced = true \end{cases} \tag{10.3}$$

Με αυτό τον τρόπο, ο σχεδιαστής αποκτά μεγαλύτερη ευελιξία στο έργο του, καθώς μπορεί να ρυθμίσει το χρονικό σημείο στο οποίο η μέση κάλυψη των κανόνων θα αρχίσει να φθίνει (το σημείο στο οποίο ο πληθυσμός αποκτά τη μέγιστή του χωρητικότητα) και τον αριθμό των διαγραφών που θα συμβούν, λόγω της πρόσθεσης κανόνων στον πληθυσμό, ανεξάρτητα το ένα από το άλλο.

Περι τοϒ θ_{exp}

Όπως αναφέραμε στην Παρ. 6.4.2, ίσως είναι σκόπιμο να εγκαταλείψουμε την έκπτωση της καταλληλότητας των κανόνων βάσει της εμπειρίας τους στη συνιστώσα Εξερεύνησης, λόγω της δυσκολίας στον προσδιορισμό μίας βέλτιστης τιμής για κάθε σύνολο δεδομένων, αλλά και ώστε κάθε κανόνας, ανεξάρτητα από τον αριθμό των δειγμάτων που καλύπτει, να αντιμετωπίζεται ισότιμα σε σχέση με τους υπόλοιπους, όσον αφορά στο από ποιο χρονικό σημείο και έπειτα μπορεί να συμμετέχει στην εξελικτική διαδικασία. Το αντικειμενικό κριτήριο για κάθε κανόνα i θα μπορούσε να είναι το χρονικό σημείο στο οποίο ο κανόνας έχει αξιολογηθεί πάνω σε όλα τα δείγματα του συνόλου δεδομένων D που αυτός μπορεί να καλύψει, δηλαδή το σημείο στο οποίο για πρώτη φορά έχει παρουσιαστεί στον κανόνα το D στο σύνολό του.

$$fitness(i) = \begin{cases} 0, & instancesChecked < |D| \\ (accuracy(i))^{\nu}, & \text{αλλού} \end{cases} \tag{10.4}$$

Βέβαια, με αυτό τον τρόπο θα περιμέναμε μία χαμηλότερη πίεση προς γενίκευση, γιατί πλέον καθίσταται άσχετος ο αριθμός δειγμάτων που καλύπτει ένας κανόνας και άρα το πότε αυτός καταφέρνει να υπερπηδήσει το κατώφλι θ_{exp}.

Περι των διαγραφων

Στον GMl-ASLCS υπάρχουν δύο συνιστώσες διαγραφών: η πλέον καθιερωμένη διαγραφή με επιλογή κανόνων από τον πληθυσμό (εδώ με επιλογή ρουλέτας) και η διαγραφή κανόνων με κριτήρια διαγραφής πάνω στα σύνολα Match Set. Η λογική και των διαγραφών μπορεί να επεκταθεί περαιτέρω. Ίσως είναι σκόπιμο να προσπαθήσουμε να διαγράφουμε, υπό συνθήκες, κανόνες με τον πρώτο τρόπο, όχι μόνο από τον πληθυσμό, αλλά και από τα σύνολα Match Set $[M]$ ή από τα Incorrect Set $[!C]$, λόγω του μικρότερου αριθμού τους, αλλά και των αντικειμενικών συνθηκών κάτω από τις οποίες σχηματίζονται τα παραπάνω σύνολα. Αντίστοιχα, η δεύτερη μέθοδος διαγραφής, όπως προαναφέραμε στην Παρ. 6.3.2, ίσως χρειάζεται να χρησιμοποιείται σε διαφορετικά σύνολα εκτός του Match Set.

Επιπρόσθετα, ανεξάρτητα από την προσπάθειά μας για αύξηση του μέσου αριθμού δειγμάτων που καλύπτουν οι κανόνες ενός ΜαΣΤ, η λογική της διαγραφής κανόνων από τα Match Sets μπορεί να επεκταθεί και σε περισσότερα επίπεδα κάλυψης εκτός από το χαμηλότερο. Η διαγραφή κανόνων σε *κάθε* επίπεδο κάλυψης ίσως αποτελέσει κάτι το βοηθητικό προς την εκπαιδευτική διαδικασία, διαγράφοντας κανόνες με χαμηλή καταλληλότητα εν γένει. Στην παραπάνω περίπτωση, ίσως να ήταν χρήσιμο να θεωρήσουμε ένα επιπλέον ανώφλι καταλληλότητας (ή και κάποιο άλλο κριτήριο) λόγω της μεγάλης αλλαγής που η παραπάνω λειτουργία θα επιφέρει στην εξελικτική διαδικασία.

Συσχετισεις Ετικετων

Από την ανάλυσή μας μέχρι στιγμής απουσιάζει μία μεθοδολογία που να λαμβάνει υπόψη τη συσχέτιση των ετικετών ενός πολυκατηγορικού προβλήματος για τον ακριβέστερο συμπερασμό των ετικετών αταξινόμητων δειγμάτων. Στην Παρ. 2.5.3 είδαμε τους δύο τρόπους με τους οποίους μπορούμε οπτικά να αναγνωρίσουμε το βαθμό αλληλοσυσχετίσεων των ετικετών ενός προβλήματος, αλλά η συνολική μεθοδολογία μας δεν χρησιμοποιεί κάποιον από αυτούς. Από τους δύο τρόπους, οι Χάρτες Θερμότητας θα ήταν περισσότερο πρόσφοροι για την εφαρμογή τους στο συνολικό πλαίσιο των πολυκατηγορικών ΜαΣΤ, λόγω της άμεσης μετάφρασής τους σε πίνακες πιθανοτήτων.

Πιθανές μεθοδολογίες θα μπορούσαν να χρησιμοποιούν τις παραπάνω πιθανότητες για την εφαρμογή κάποιου είδους επιπρόσθετης έκπτωσης στην καταλληλότητα των κανόνων κατά την εξελικτική διαδικασία, ή ακόμα την τροποποίηση του τμήματος Κάλυψης ώστε να υπάρχει κάποια "πρόταση" ετικετών σε κανόνες. Επιπρόσθετα, εκτός από τη ρύθμιση του κατωφλίου στη συνιστώσα Επίδοσης, το ΜαΣΤ θα μπορούσε να επιλέξει ένα μικρότερο υποσύνολο των κανόνων που θα αποφασίσουν για την κατηγοριοποίηση ενός αταξινόμητου δείγματος με βάση το πόσο πιθανό είναι να κατηγοριοποιηθεί σε μία συγκεκριμένη ετικέτα.

Περι της εκτιμησης του μεγεθους cs

Όπως είδαμε στην Παρ. 6.2.5, ο GMl-ASLCS χρησιμοποιεί την εκτίμηση του μέσου ελαχίστου μεγέθους των Correct Sets στα οποία συμμετέχει ένας κανόνας, αντί για την ίδια την τιμή του, για όλα τα Correct Sets στα οποία συμμετέχει. Αυτή η μεθοδολογία είδαμε πώς επηρεάζει την αντιμετώπιση των κανόνων του ΜαΣΤ, μέσω της υπερ-εκτίμησης και υπο-εκτίμησης της πραγματικής τιμής του cs. Για τον υπολογισμό του πραγματικού cs ενός κανόνα $rule$, αντί της Εξ. 6.13 που χρησιμοποιεί το ρυθμό μάθησης β, μπορεί να χρησιμοποιηθεί ο απευθείας υπολογισμός του cs μέσω της

$$rule.cs(t) = \frac{(t-1) \cdot rule.cs(t-1) + cs(t)}{t} \tag{10.5}$$

όπου t και $t-1$ είναι ο αριθμός των φορών που έχει συμμετέχει ο $rule$ σε Correct Set, $cs(t)$ το μέγεθος του Correct Set στο οποίο συμμετέχει ο $rule$ την t φορά και $rule.cs(t)$ η μέση τιμή του cs του $rule$ την τρέχουσα (t) στιγμή.

Διαμοιρασμος Καταλληλοτητας

Ο διαμοιρασμός καταλληλότητας (fitness sharing) που χρησιμοποιεί ο UCS στοχεύει στο να υπάρχει ποικιλομορφία στον πληθυσμό των κανόνων που εξελίσσει, τιμωρώντας τους υπεργενικούς κανόνες που συμμετέχουν στα ίδια Correct Sets με κανόνες που είναι καταλληλότεροι από αυτούς. Αυτό παράγει μία επιπρόσθετη πίεση προς διαγραφή υπεργενικών κανόνων και μία επιλεκτική πίεση προς ανακάλυψη ειδικών κανόνων [OPBM08].

Ο GMl-ASLCS θα μπορούσε να χρησιμοποιήσει τις παραπάνω ευεργετικές ιδιότητες του διαμοιρασμού καταλληλότητας σε συνδυασμό με τις ήδη υπάρχουσες λειτουργίες του, ώστε να βελτιώσει την ποιότητα των γενικών του κανόνων, αλλά και να ανακαλύψει περισσότερους ειδικούς κανόνες που ίσως δεν καταφέρνει να βρει. Αν και ο διαμοιρασμός καταλληλότητας δεν έχει υλοποιηθεί, υποπτευόμαστε ότι στον πολυκατηγορικό χώρο θα αυξήσει την υπολογιστική πολυπλοκότητα του GMl-ASLCS ανάλογα με τον αριθμό των ετικετών του προβλήματος και το μέγεθος του πληθυσμού, αν διαμοιράσουμε και εδώ την καταλληλότητα των κανόνων των Correct Sets, δηλαδή ανά ετικέτα. Ένας εναλλακτικός τρόπος θα ήταν, αντί για το διαμοιρασμό της καταλληλότητας για κάθε ετικέτα, να συγκεντρώσουμε όλους τους κανόνες των επιμέρους Correct Sets σε ένα σύνολο, δηλαδή όλους τους κανόνες που συμφωνούν με τις επιμέρους ετικέτες ενός δείγματος, και να διαμοιράσουμε την καταλληλότητά τους σε αυτό το σύνολο.

ΜΕΡΟΣ IV

Παραρτήματα

Παραρτημα Α΄

Αντιστοιχιση επιστημονικων ορων στα Αγγλικα

Παρακάτω παρατίθενται κάποιοι από τους πιο συχνά χρησιμοποιούμενους όρους στην παρούσα διπλωματική, μαζί με τη μετάφρασή τους στην Αγγλική γλώσσα, όπως εμφανίζονται στη διεθνή βιβλιογραφία.

Ακρίβεια	Accuracy
Αμοιβαία Αποκλειόμενοι	Mutually Exclusive
Αξιοποίηση	Exploitation
Γενετικοί Αλγόριθμοι	Genetic Algorithms
Γνώρισμα	Features, Attributes
Διασταύρωση	Crossover
Ενισχυτική Μάθηση	Reinforcement Learning
Εξαντλητικό	Exhaustive
Εξόρυξη Δεδομένων	Data Mining
Επιβλεπόμενη Μάθηση	Supervised Learning
Ετικέτα	Label
Επιλογή Ρουλέτας	Roulette wheel selection
Διαμοιρασμός Καταλληλότητας	Fitness Sharing
Διασταυρωμένη Επικύρωση	Cross-Validation
Διασταύρωση	Crossover
Μηχανές Διανυσμάτων Υποστήριξης	Support Vector Machines
Μηχανική Μάθηση	Machine Learning
Μηδενική Κάλυψη	Zero coverage
Κατηγορία - Κλάση	Class
Κατηγοριοποίηση	Classification
Κατηγορική Πληθικότητα	Label Cardinality
Κατηγορική Πυκνότητα	Label Density
Κατώφλι	Threshold
Μετάλλαξη	Mutation
Πλήρης Χάρτης Αποφάσεων	Complete Action Map
Πληθυσμιακή Εξισορρόπηση	Niching
Ρύθμιση	Calibration

Ταξινόμηση	Classification
Ταξινομητής	Classifier
Ταξινομητής Συνόλων	Ensemble Classifier
Τμήμα	Component
Σταθερή Κατάσταση	Steady State
Συγχώνευση - Αφομοίωση - Ὑπαγωγή	Subsumption
Συμβιβασμός	Trade-off
Σύνολο Δεδομένων	Data Set
Σύνολο Δεδομένων Εκπαίδευσης	Train Set
Σύνολο Δεδομένων Ελέγχου	Test Set
Σύνολο Κανόνων	Rule Set
Σύνολο Ορθής Απόφασης	Correct Set
Μικρο-κανόνας	Microclassifier
Μαθάνοντα Συστήματα Ταξινομητών	Learning Classifier Systems
Μακρο-κανόνας	Macroclassifier
Φυσική Επιλογή	Natural Selection
Χάρτης Βέλτιστων Αποφάσεων	Best Action Map
Χώρος Ετικετών	Label Space
Χώρος Χαρακτηριστικών	Feature Space

Βιβλιογραφια

[AKA91] D.W. Aha, D. Kibler, and M.K. Albert. Instance-based learning algorithms. *Machine learning*, 6(1):37–66, 1991.

[Aly05] M. Aly. Survey on Multiclass Classification Methods. 2005.

[ATM13] Miltiadis Allamanis, Fani A. Tzima, and Pericles A. Mitkas. Effective rule-based multi-label classification with learning classifier systems. In *ICANNGA*, pages 466–476, 2013.

[BBMH08] Larry Bull, Ester Bernadó-Mansilla, and John H. Holmes, editors. *Learning Classifier Systems in Data Mining*, volume 125 of *Studies in Computational Intelligence*. Springer, 2008.

[BGV92] B.E. Boser, I.M. Guyon, and V.N. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152. ACM, 1992.

[BLSB04] M.R. Boutell, J. Luo, X. Shen, and C.M. Brown. Learning multi-label scene classification* 1. *Pattern Recognition*, 37(9):1757–1771, 2004.

[BMGG03] E. Bernadó-Mansilla and J.M. Garrell-Guiu. Accuracy-based learning classifier systems: models, analysis and applications to classification tasks. *Evolutionary Computation*, 11(3):209–238, 2003.

[Buc02] M. Buckland. *AI techniques for game programming*. Course Technology, 2002.

[BW01] M. Butz and S. Wilson. An algorithmic description of xcs. *Advances in Learning Classifier Systems*, pages 267–274, 2001.

[CK01] A. Clare and R. King. Knowledge discovery in multi-label phenotype data. *Principles of Data Mining and Knowledge Discovery*, pages 42–53, 2001.

[DTMV05] S. Diplaris, G. Tsoumakas, P. Mitkas, and I. Vlahavas. Protein classification with multiple algorithms. *Advances in Informatics*, pages 448–456, 2005.

[EW02] A. Elisseeff and J. Weston. Kernel methods for multi-labelled classification and categorical regression problems. *Advances in neural information processing systems*, 14:681–687, 2002.

[Fre05] Alex A. Freitas. *Evolutionary Algorithms for Data Mining*, volume The Data Mining and Knowledge Discovery Handbook, pages 435–467. Springer, January 2005.

[Fri40] Milton Friedman. A Comparison of Alternative Tests of Significance for the Problem of m Rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[Gol02] David E. Goldberg. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[HK06] J. Han and M. Kamber. *Data mining: concepts and techniques*. Morgan Kaufmann, 2006.

[Hol75] John H. Holland. *Adaptation in Natural and Artificial Systems: An Introductory Analysis with Applications to Biology, Control and Artificial Intelligence*. University of Michigan Press, Ann Arbor, MI, USA, 1975.

[HR78] J. H. Holland and J. S. Reitman. Cognitive systems based on adaptive algorithms. In D. A. Waterman and F. Hayes-Roth, editors, *Pattern directed inference systems*, pages 313–329. Academic Press, New York, 1978.

[HSCB91] R. Hanson, J. Stutz, P. Cheeseman, and Ames Research Center. Artificial Intelligence Research Branch. *Bayesian classification theory*. Citeseer, 1991.

[KK01] Tim Kovacs and Manfred Kerber. What Makes a Problem Hard for XCS? In *IWLCS '00: Revised Papers from the Third International Workshop on Advances in Learning Classifier Systems*, pages 80–102, London, UK, 2001. Springer-Verlag.

[Kov00] Tim Kovacs. Strength or Accuracy? Fitness Calculation in Learning Classifier Systems. In *Learning Classifier Systems, From Foundations to Applications*, pages 143–160, London, UK, 2000. Springer-Verlag.

[LZZ06] T. Li, C. Zhang, and S. Zhu. Empirical studies on multi-label classification. In *Proceedings of the 18th IEEE international conference on tools with artificial intelligence*, pages 86–92. Citeseer, 2006.

[MBK07] James Marshall, Gavin Brown, and Tim Kovacs. Bayesian estimation of rule accuracy in ucs. In *Proceedings of the 2007 GECCO Conference Companion on Genetic and Evolutionary Computation*, pages 2831–2834. ACM Press, July 2007.

[Mil11] Allamanis Miltiadis. Multilabel classification with learning classifier systems. diploma thesis, Aristotle University of Thessaloniki, 2011.

[Mit97] Thomas M. Mitchell. *Machine Learning*. McGraw-Hill Higher Education, 1997.

[Mur98] Sreerama K. Murthy. Automatic Construction of Decision Trees from Data: A Multi-Disciplinary Survey. *Data Mining and Knowledge Discovery*, 2(4):345–389, 1998.

[Nem63] P.B Nemenyi. *Distribution-free multiple comparisons.* PhD thesis, Princeton University, 1963.

[OPBM08] A. Orriols-Puig and E. Bernadó-Mansilla. Revisiting UCS: Description, fitness sharing, and comparison with xcs. *Learning Classifier Systems*, pages 96–116, 2008.

[PBM+07] J.P. Pestian, C. Brew, P. Matykiewicz, DJ Hovermale, N. Johnson, K.B. Cohen, and W. Duch. A shared task involving multi-label classification of clinical free text. In *Proceedings of the Workshop on BioNLP 2007: Biological, Translational, and Clinical Language Processing*, pages 97–104. Association for Computational Linguistics, 2007.

[Rea10] J. Read. Scalable Multi-label Classification. 2010.

[RLS04] Arturo M. Ráez, Luís A. López, and Ralf Steinberger. Adaptive Selection of Base Classifiers in One-Against-All Learning for Large Multi-labeled Collections. In José L. Vicedo, Patricio Martínez-Barco, Rafael Muñoz, and Maximiliano Saiz Noeda, editors, *Advances in Natural Language Processing*, volume 3230 of *Lecture Notes in Computer Science*, chapter 1, pages 1–12. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2004.

[RPH08] J. Read, B. Pfahringer, and G. Holmes. Multi-label classification using ensembles of pruned sets. In *2008 Eighth IEEE International Conference on Data Mining*, pages 995–1000. IEEE, 2008.

[RPHF09] J. Read, B. Pfahringer, G. Holmes, and E. Frank. Classifier chains for multi-label classification. *Machine Learning and Knowledge Discovery in Databases*, pages 254–269, 2009.

[SS00] R.E. Schapire and Y. Singer. Boostexter: A boosting-based system for text categorization. *Machine learning*, 39(2):135–168, 2000.

[TK07] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[Tzi11] Fani A. Tzima and Pericles A. Mitkas. Strength-based learning classifier systems revisited: effective rule evolution in supervised classification tasks. *Engineering Applications of Artificial Intelligence*, 20(2):818–832, 2013.

[TSK05] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar. *Introduction to Data Mining, (First Edition).* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 2005.

[TTKV08] K. Trohidis, G. Tsoumakas, G. Kalliris, and I. Vlahavas. Multilabel classification of music into emotions. In *Proc. 9th International Conference on Music Information Retrieval (ISMIR 2008), Philadelphia, PA, USA*, volume 2008, 2008.

[TTM12] Fani A. Tzima, John B. Theocharis, and Pericles A. Mitkas. Clustering-based initialization of learning classifier systems. effects on model performance, readability and induction time. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 16:1267–1286, 2012. 10.1007/s00500-012-0811-y.

[TV07] G. Tsoumakas and I. Vlahavas. Random k-labelsets: An ensemble method for multilabel classification. *Machine Learning: ECML 2007*, pages 406–417, 2007.

[Tzi12] A.F. Tzima. *Learning Classifier Systems for Supervised Classification Problems.* PhD thesis, Aristotle University of Thessaloniki, 2012.

[Wat89] C.J.C.H. Watkins. Learning from delayed rewards. 1989.

[WF05] Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques (2nd Edition).* Morgan Kaufmann, San Francisco, 2005.

[Wil94] Stewart W. Wilson. ZCS: A Zeroth-level Classifier System. *Evolutionary Computation*, 2(1):1–18, 1994.

[Wil95] Stewart W. Wilson. Classifier fitness based on accuracy. *Evolutionary Computation*, 3(2):149–175, 1995.

[YTS07] Rong Yan, Jelena Tesic, and John R. Smith. Model-shared subspace boosting for multi-label classification. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '07, pages 834–843, New York, NY, USA, 2007. ACM.

[Zha00] G.P. Zhang. Neural networks for classification: a survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 30(4):451–462, 2000.

[ZZ06] Min-Ling Zhang and Zhi-Hua Zhou. Multilabel neural networks with applications to functional genomics and text categorization. *IEEE Transactions on Knowledge and Data Engineering*, 18:1338–1351, October 2006.

[ZZ07] M.L. Zhang and Z.H. Zhou. ML-KNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 40(7):2038–2048, 2007.

[ZZ10] M.L. Zhang and K. Zhang. Multi-label learning by exploiting label dependency. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 999–1008. ACM, 2010.