Επιλογή Χαρακτηριστικών για Προβλήματα Ταξινόμησης

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από τον

Οδυσσέα Πετρόχειλο

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Ιούνιος 2009

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Αριστείδη Λύκα για τη βοήθεια που μου προσέφερε κατά την εκπόνηση αυτής της εργασίας.

ΠΕΡΙΕΧΟΜΕΝΑ

	Σελ
ΕΥΧΑΡΙΣΤΙΕΣ	ii
ΠΕΡΙΕΧΟΜΕΝΑ	iii
ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ	v
ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ	vi
ΠΕΡΙΛΗΨΗ	vii
EXTENDED ABSTRACT IN ENGLISH	viii
ΚΕΦΑΛΑΙΟ 1. Εισαγωγή	1
1.1. Περιγραφή του προβλήματος	1
1.2. Αντικείμενο της εργασίας	3
1.3. Δομή της εργασίας	4
ΚΕΦΑΛΑΙΟ 2. Βασικές έννοιες και σχετική βιβλιογραφία	6
2.1. Το πρόβλημα της ταξινόμησης	6
2.2. Επιλογή χαρακτηριστικών για προβλήματα ταξινόμησης	8
2.3. Σημαντικές έννοιες για το πρόβλημα της επιλογής χαρακτηριστικών	9
2.4. Τεχνικές αναζήτησης	12
2.5. Επιλογή χαρακτηριστικών με wrapper	15
2.5.1. Η διαδικασία αξιολόγησης με χρήση wrapper	15
2.5.2. Χαρακτηριστικά των wrapper	16
2.6. Επιλογή χαρακτηριστικών με φίλτρα	17
2.6.1. Univariate μέθοδοι	17
2.6.2. Multivariate μέθοδοι	20
2.7. Embedded μέθοδοι	21
ΚΕΦΑΛΑΙΟ 3. Επιλογή χαρακτηριστικών με ελάχιστη περιττή πληροφορία	24
3.1. Το κριτήριο της μέγιστης εξάρτησης	24

3.2. Ο αλγόριθμος mRMR 26
3.3. Ο αλγόριθμος mRMR για συνεχή χαρακτηριστικά 28
3.4. Σχετικοί αλγόριθμοι 30
ΚΕΦΑΛΑΙΟ 4. Τροποποιήσεις του mRMR 33
4.1. Εκτίμηση της περιττής πληροφορίας 33
4.2. Βελτιστοποίηση του κριτηρίου αξιολόγησης 37
4.2.1. Καθορισμός του κριτηρίου αξιολόγησης 37
4.2.2. Εναλλακτικές τεχνικές αναζήτησης 39
4.2.3. Σύγκριση με τη λύση του mRMR 41
4.3. Κανονικοποιημένο mRMR 44
ΚΕΦΑΛΑΙΟ 5. Πειράματα 50
5.1. Τα σύνολα δεδομένων 50
5.2. Αναλυτική περιγραφή μεθοδολογίας σύγκρισης 53
5.2.1. Αποτίμηση ακρίβειας ταξινόμησης 53
5.2.2. Ο ταξινομητής και η επιλογή παραμέτρων 54
5.2.3. Σύγκριση αποτελεσμάτων 56
5.3. Σύγκριση mRMR και κανονικοποιημένου mRMR 57
5.4. Σύγκριση προσεγγίσεων για τον υπολογισμό της περιττής πληροφορίας 61
ΚΕΦΑΛΑΙΟ 6. Συμπεράσματα και μελλοντική εργασία 67
ΑΝΑΦΟΡΕΣ 69
ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ 73

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Πίνακας	Σελ
Πίνακας 2.1 Το πρόβλημα XOR	10
Πίνακας 4.1 Αξιολόγηση λύσεων ως προς το κριτήριο $J_2=V(S)-\frac{1}{2}W(S)$	42
Πίνακας 4.2 Αξιολόγηση λύσεων ως προς το κριτήριο J_1=V(S)-W(S)	44
Πίνακας 4.3 Αξιολόγηση λύσεων ως προς το κριτήριο J_3=V(S)-5W(S)	44
Πίνακας 5.1 Σύνοψη των συνόλων δεδομένων	53
Πίνακας 5.2 Παράμετροι που χρησιμοποιήθηκαν κατά την αποτίμηση υποσυνόλων και την εκπαίδευση του SVM	56
Πίνακας 5.3 Ακρίβεια ταξινόμησης που επιτυγχάνεται χρησιμοποιώντας όλα τα χαρακτηριστικά	58
Πίνακας 5.4 Αριθμός λαθών στο σύνολο δεδομένων Colon-cancer	59
Πίνακας 5.5 Αριθμός λαθών στο σύνολο δεδομένων Lymphoma	59
Πίνακας 5.6 Αριθμός λαθών στο σύνολο δεδομένων Lung-cancer	59
Πίνακας 5.7 Αριθμός λαθών στο σύνολο δεδομένων Leukemia	59
Πίνακας 5.8 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Multiple Features	60
Πίνακας 5.9 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Internet-Ads	60
Πίνακας 5.10 Αριθμός λαθών στο σύνολο δεδομένων Colon-cancer	62
Πίνακας 5.11 Αριθμός λαθών στο σύνολο δεδομένων Lymphoma	62
Πίνακας 5.12 Αριθμός λαθών στο σύνολο δεδομένων Lung-cancer	62
Πίνακας 5.13 Αριθμός λαθών στο σύνολο δεδομένων Leukemia	62
Πίνακας 5.14 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Multiple Features	63
Πίνακας 5.15 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Internet-Ads	63
Πίνακας 5.16 Αριθμός λαθών στο σύνολο δεδομένων Colon-cancer	65
Πίνακας 5.17 Αριθμός λαθών στο σύνολο δεδομένων Lymphoma	65
Πίνακας 5.18 Αριθμός λαθών στο σύνολο δεδομένων Lung cancer	66
Πίνακας 5.19 Αριθμός λαθών στο σύνολο δεδομένων Leukemia	66
Πίνακας 5.20 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Multiple Features	66
Πίνακας 5.21 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Internet-Ads	66

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

Σχήμα Σελ

Σχήμα 4.1 Ομοιότητες των λιγότερο περιττών χαρακτηριστικών ως προς τις προσεγγίσεις mean-redundancy και mean-top5 με τα ήδη επιλεγμένα χαρακτηριστικά. 35

Σχήμα 4.2 Η μεταβολή του κριτηρίου αξιολόγησης J_2 σε σχέση με τον αριθμό των χαρακτηριστικών που υπάρχουν στο υποσύνολο 39

Σχήμα 4.3 Συνάφεια και περιττή πληροφορία υποψήφιων χαρακτηριστικών σε δύο διαφορετικά σύνολα δεδομένων. 46

Σχήμα 4.4 Κατάταξη επιλεγμένων χαρακτηριστικών με βάση τη συνάφεια στο σύνολο δεδομένων Lymphoma 48

Σχήμα 4.5 Κατάταξη επιλεγμένων χαρακτηριστικών με βάση τη συνάφεια στο σύνολο δεδομένων Multiple features 48

Σχήμα 5.1 Συνάφεια και περιττή πληροφορία υποψηφίων χαρακτηριστικών κατά τη δέκατη επανάληψη του mRMR με βάση τις τρεις διαφορετικές προσεγγίσεις (σύνολο δεδομένων Lymphoma). 64

ΠΕΡΙΛΗΨΗ

Οδυσσέας-Ηρακλής Πετρόχειλος του Νικολάου και της Δωρίδας.
MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούνιος, 2009.
Επιλογή χαρακτηριστικών για προβλήματα ταξινόμησης.
Επιβλέπων: Αριστείδης Λύκας.

Το πρόβλημα της επιλογής χαρακτηριστικών για προβλήματα ταξινόμησης συνίσταται στον εντοπισμό μέσα από ένα σύνολο χαρακτηριστικών, εκείνων που είναι πρωτεύουσας σημασίας για το διαχωρισμό των κατηγοριών. Η χρήση αυτών των χαρακτηριστικών (και η απόρριψη των υπολοίπων, περιττών ή άσχετων με το πρόβλημα) κατά τη διαδικασία της ταξινόμησης μπορεί να βοηθήσει να φτιαχτούν πιο εύρωστα και πιο εύκολα ερμηνεύσιμα μοντέλα.

Η εργασία αυτή, αφού συνοψίσει τις προσεγγίσεις που έχουν προταθεί για την επίλυση του προβλήματος, εστιάζει στο δημοφιλή αλγόριθμο mRMR. Κοινό στοιχείο όλων των προσεγγίσεων είναι ότι χρησιμοποιούν μία συνάρτηση αξιολόγησης βάση της οποίας αποτιμάται η ποιότητα ενός υποσυνόλου χαρακτηριστικών. Ο αλγόριθμος mRMR χρησιμοποιεί μια συνάρτηση αξιολόγησης η οποία επιβραβεύει υποσύνολα που περιέχουν χαρακτηριστικά που είναι συσχετισμένα με την κατηγορία και ταυτόχρονα δεν είναι συσχετισμένα μεταξύ τους. Έτσι επιδιώκεται η μείωση της περιττής πληροφορίας που περιέχεται στο τελικό επιλεγμένο υποσύνολο.

Σε αυτή την εργασία μελετούμε εκτενώς διάφορες πτυχές του mRMR. Αρχικά εξετάζουμε τον τρόπο με τον οποίο εκτιμάται η συσχέτιση που υπάρχει μεταξύ των χαρακτηριστικών ενός υποσυνόλου. Επίσης μελετούμε την αποτελεσματικότητα της προς τα εμπρός (forward) αναζήτησης που χρησιμοποιεί ο mRMR. Τέλος εξετάζεται η συνάρτηση αξιολόγησης και ο τρόπος που αυτή συνδυάζει τους δύο στόχους, της αυξημένης συσχέτισης μεταξύ χαρακτηριστικών-κατηγορίας και της μειωμένης περιττής πληροφορίας. Για όλα τα παραπάνω προτείνονται τροποποιήσεις, των οποίων η αποτελεσματικότητα αποτιμάται πειραματικά σε προβλήματα υψηλής διάστασης κυρίως από το πεδίο της βιοπληροφορικής.

EXTENDED ABSTRACT IN ENGLISH

Petrocheilos, Odysseas-Iraklis, N.
MSc, Computer Science Department, University of Ioannina, Greece. June, 2009.
Thesis Tilte: Feature selection for classification.
Thesis Supervisor: Aristidis Likas.

Feature selection for classification is the task of discovering features that provide the information needed for the correct classification of examples into classes. These informative features are often called relevant since they are relevant to the target concept that the classifier tries to learn. The ultimate goal of feature selection is the dimensionality reduction of the problem being studied. Unlike other dimensionality reduction techniques such as feature extraction, feature selection does not involve any transformation of the original feature space. Instead, a small subset of useful features is selected and the rest of them are discarded. Using only a subset of relevant features and ignoring the irrelevant and redundant ones during the classification procedure, can yield significant benefits such as reduced risk of overfitting, more interpretable classification models, insight for the problem being studied and reduced computational cost.

In this thesis, after introducing some important concepts, we review the basic approaches to feature selection found in the literature. Feature selection algorithms can be divided into three categories: filters, wrappers and embedded methods. We focus on a filter method called mRMR (maximum relevance – minimum redundancy) which has proved to be quite efficient in some difficult high dimensional problems. A feature subset should fulfill two conditions in order to be selected by the mRMR algorithm: i) it should consist of features that are highly relevant with the class ii) it should consist of features that are not highly correlated with each other. The first condition aims at selecting highly informative features that can be used to classify correctly new examples. The second condition aims at reducing the redundant

information contained in the feature subset. This is important since it has been proven that little gain stems from using highly relevant features if they are redundant. Essentially, the selected feature subset gives a trade-off between high relevance and low redundancy.

The mRMR algorithm is basically a heuristic which lacks strict theoretical background. This means that there are some aspects of the algorithm that need to be better understood and possibly improvements might occur by introducing appropriate modifications. In this thesis we study some of these aspects. First of all, we study the approximation used for the estimation of redundant information contained in a feature subset. Furthermore, we examine the objective function used by mRMR to evaluate feature subsets. More specifically, we investigate if this objective function gives a good trade-off between increased relevance and reduced redundancy. For both of the above issues we suggest some modifications. Finally, assuming that the original objective function is used (mRMR's objective function with no modifications), we examine if a better solution can be found by using a different search strategy instead of the forward selection search used by mRMR.

The proposed modifications have been compared against the original mRMR algorithm. The evaluation criterion used for the comparison is the classification accuracy achieved when using the features selected by each method. The methods were tested mostly on high dimensional bioinformatics datasets using the state-of-the-art SVM classifier.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή του προβλήματος
1.2 Αντικείμενο της εργασίας
1.3 Δομή της εργασίας

1.1. Περιγραφή του προβλήματος

Στις μέρες μας η απόκτηση πληροφορίας είναι ευκολότερη από ότι στο παρελθόν. Τεχνολογικά προηγμένες μέθοδοι μας δίνουν τη δυνατότητα να καταγράφουμε δεδομένα με ολοένα αυξανόμενο ρυθμό. Ο τεράστιος όγκος δεδομένων που παράγεται επηρεάζει αρνητικά την ικανότητα μας να τα μελετήσουμε, να εξάγουμε συμπεράσματα και να λύσουμε προβλήματα. Μεγάλο μέρος των δεδομένων μπορεί να αφορά άχρηστες πληροφορίες, πληροφορίες που δεν παρουσιάζουν ενδιαφέρον σε σχέση με το πρόβλημα που μελετάται.

Σε ένα υποθετικό παράδειγμα ένας γιατρός θέλει να μελετήσει τις πληροφορίες που υπάρχουν σε μια βάση δεδομένων για έναν ασθενή προκειμένου να αποφανθεί αν υπάρχει σημαντικός κίνδυνος ο ασθενής να εμφανίσει κάποιο είδος καρδιοπάθειας. Τον ενδιαφέρει να μελετήσει χαρακτηριστικά του ασθενούς όπως το βάρος, το ύψος, η αρτηριακή πίεση, τα επίπεδα χοληστερίνης κ.α. Στη βάση δεδομένων μπορεί υπάρχουν πληροφορίες και για άλλα χαρακτηριστικά του ασθενούς όπως ο τόπος καταγωγής ή η οικογενειακή κατάσταση. Προφανώς αυτά τα χαρακτηριστικά είναι άχρηστα σε σχέση με το πρόβλημα που μελετάται. Ο γιατρός τελικά θα τα αγνοήσει αφού πρώτα ξοδέψει λίγο παραπάνω χρόνο μέχρι να βρει τα στοιχεία που τον ενδιαφέρουν.

Σε σύγχρονες εφαρμογές, τα δεδομένα περιγράφονται πολλές φορές από χιλιάδες χαρακτηριστικά, πολύ περισσότερα από αυτά που έχει την ικανότητα ένας άνθρωπος

να «φιλτράρει». Η επιλογή χαρακτηριστικών, που μελετάται σε αυτή την εργασία, είναι μια τεχνική προεπεξεργασίας των δεδομένων με σκοπό την ανακάλυψη χρήσιμων χαρακτηριστικών και την απόρριψη των υπόλοιπων, πριν τα δεδομένα τροφοδοτηθούν σε κάποιον αλγόριθμο μάθησης. Η αξιολόγηση ενός χαρακτηριστικού ως λίγο ή πολύ χρήσιμου δεν είναι εύκολη διαδικασία και είναι το κύριο αντικείμενο μελέτης στο πρόβλημα της επιλογής χαρακτηριστικών.

Η δυσκολία του προβλήματος της επιλογής ενός υποσυνόλου χαρακτηριστικών οφείλεται σε δύο κυρίως λόγους. Πρώτον, το πλήθος υποσυνόλων που θα μπορούσαν να επιλεγούν αυξάνεται εκθετικά σε σχέση με τον αριθμό των χαρακτηριστικών του αρχικού συνόλου. Αν υποθέσουμε ότι δίνεται ένα κριτήριο αξιολόγησης υποσυνόλων, η εύρεση του καλύτερου υποσυνόλου ως προς αυτό δεν είναι υπολογιστικά εφικτή. Δεύτερον, η ποιότητα ενός υποσυνόλου εξαρτάται από πολλούς παράγοντες και έτσι δεν μπορεί να οριστεί εύκολα ένα αντικειμενικό κριτήριο αξιολόγησης. Με άλλα λόγια δεν υπάρχει τρόπος να αποτιμηθεί με ακρίβεια η ποιότητα ενός υποσυνόλου, αντίθετα στην πράξη με χρήση ευρετικών τεχνικών επιλέγεται τελικά ένα υποσύνολο που αναμένεται ότι θα οδηγήσει σε καλή απόδοση τον αλγόριθμο μάθησης που θα το χρησιμοποιήσει.

Τα οφέλη που μπορούν να προκύψουν από τη μείωση της διάστασης μέσω της επιλογής χαρακτηριστικών είναι πολλά. Καταρχήν η απόδοση των αλγορίθμων μάθησης επηρεάζεται αρνητικά από τη μεγάλη διάσταση. Για παράδειγμα στο πρόβλημα της μάθησης με επίβλεψη, λόγω της μεγάλης διάστασης είναι αυξημένος ο κίνδυνος να υπάρξει υπερεκπαίδευση (overfitting). Η επιλογή χαρακτηριστικών βοηθάει να αμβλυνθεί αυτό το πρόβλημα. Επιπλέον, η επιλογή χαρακτηριστικών μπορεί να βοηθήσει στην κατανόηση των αποτελεσμάτων που παράγονται από έναν αλγόριθμο. Προφανώς ένας άνθρωπος δεν μπορεί να ερμηνεύσει μια απεικόνιση χιλιάδων μεταβλητών που παράχθηκε από έναν αλγόριθμο μάθησης με επίβλεψη, ίσως όμως μπορεί να ερμηνεύσει μία απεικόνιση π.χ. πέντε μεταβλητών. Φυσικά ένα επιπλέον κέρδος που έρχεται σε συνδυασμό με όλα τα παραπάνω είναι το μειωμένο υπολογιστικό κόστος.

1.2. Αντικείμενο της εργασίας

Η τεχνική της επιλογής χαρακτηριστικών μπορεί να χρησιμοποιηθεί είτε σε προβλήματα ομαδοποίησης είτε σε προβλήματα ταξινόμησης. Οι αλγόριθμοι επιλογής χαρακτηριστικών για το πρόβλημα της ταξινόμησης μπορούν να χωριστούν σε τρεις βασικές κατηγορίες με βάση τον τρόπο που αξιολογούν τα υποψήφια προς επιλογή υποσύνολα χαρακτηριστικών. Οι κατηγορίες αυτές είναι τα φίλτρα, οι wrapper και οι embedded μέθοδοι. Από όλες τις κατηγορίες, οι οποίες περιγράφονται στο δεύτερο κεφάλαιο, επικεντρωνόμαστε στα φίλτρα.

Τα φίλτρα βασίζονται στην έννοια της συνάφειας μεταξύ χαρακτηριστικών και κατηγορίας, προκειμένου να ανιχνεύσουν χαρακτηριστικά χρήσιμα για τη διαδικασία της μάθησης. Τα φίλτρα στην απλούστερη μορφή τους αξιολογούν κάθε χαρακτηριστικό ξεχωριστά με βάση τη συσχέτιση του με την κατηγορία και τελικά επιλέγουν το υποσύνολο που αποτελείται από τα N πιο συσχετισμένα χαρακτηριστικά. Αυτή η επιλογή συνήθως δεν οδηγεί σε αρκετά καλά αποτελέσματα γιατί επιλέγονται πολλά *περιττά* χαρακτηριστικά. Αυτό σημαίνει ότι τα επιλεγμένα χαρακτηριστικά είναι πολύ όμοια μεταξύ τους και έτσι ο συνδυασμός τους δεν προσφέρει πολύ περισσότερη πληροφορία για την κατηγορία από αυτή που θα προσέφερε κάθε χαρακτηριστικό από μόνο του.

Για την αντιμετώπιση αυτού το προβλήματος, διάφοροι αλγόριθμοι όπως οι [4], [10], [18] θέτουν δύο στόχους που πρέπει εκπληρώνει ένα καλό υποσύνολο χαρακτηριστικών: i) να περιέχει χαρακτηριστικά που έχουν μεγάλη συσχέτιση με την κατηγορία και ii) τα χαρακτηριστικά να είναι όσο το δυνατόν ανόμοια μεταξύ τους. Οι μέθοδοι προσπαθούν να επιλέξουν υποσύνολα που επιτυγχάνουν ισορροπία ανάμεσα στους δύο στόχους τους οποίους συνδυάζουν σε μία συνάρτηση αξιολόγησης της μορφής:

$$J(S) = relevance(S) - \beta \cdot redundancy(S)$$

όπου S είναι το υποσύνολο που αποτιμάται, ενώ η ποσότητα relevance(S) εκφράζει την συσχέτιση των χαρακτηριστικών του S με την κατηγορία. Η ποσότητα redundancy(S) προσεγγίζει τη συσχέτιση που έχουν τα χαρακτηριστικά του S μεταξύ τους και εκφράζει την περιττή πληροφορία που περιέχει το υποσύνολο.

Η εργασία αυτή επικεντρώνεται στο δημοφιλή αλγόριθμο mRMR [10] και ασχολείται καταρχήν με την προσέγγιση που χρησιμοποιείται για την εκτίμηση της περιττής πληροφορίας. Δίνεται ένα παράδειγμα όπου η παρούσα προσέγγιση δεν συμφωνεί με τη διαίσθηση που υπάρχει για το πρόβλημα και προτείνεται ένας εναλλακτικός τρόπος προσέγγισης. Εξετάζεται επίσης κατά πόσο η προηγούμενη συνάρτηση αξιολόγησης καταφέρνει να επιτύχει ισορροπία στη βελτιστοποίηση των δύο στόχων και προτείνεται ένας τρόπος να διορθωθεί η ανισορροπία στις περιπτώσεις όπου υπάρχει. Τέλος εξετάζεται η ιδέα της βελτιστοποίησης του κριτηρίου J(S) με χρήση άλλων μεθόδων αναζήτησης πέραν της forward αναζήτησης που χρησιμοποιείται από τους ήδη γνωστούς αλγορίθμους.

1.3. Δομή της εργασίας

Αρχικά στο κεφάλαιο 2 της εργασίας γίνεται μια εισαγωγή στο πρόβλημα της επιλογής χαρακτηριστικών για το πρόβλημα της ταξινόμησης. Περιγράφονται το πρόβλημα της ταξινόμησης και σημαντικές έννοιες γύρω από το θέμα της επιλογής χαρακτηριστικών, ενώ στη συνέχεια παρουσιάζονται τεχνικές που επιστρατεύονται για την αναζήτηση καλών υποσυνόλων χαρακτηριστικών. Τέλος περιγράφονται οι τρεις βασικές κατηγορίες μεθόδων αξιολόγησης υποσυνόλων χαρακτηριστικών.

Το κεφάλαιο 3 αφορά αλγόριθμους που επιλέγουν χαρακτηριστικά φροντίζοντας να μειώσουν την περιττή πληροφορία που αυτά περιέχουν. Ιδιαίτερη έμφαση δίνεται στον αλγόριθμο mRMR. Το κεφάλαιο κλείνει παρουσιάζοντας τους υπόλοιπους αλγόριθμους της ίδιας κατηγορίας.

Στο κεφάλαιο 4 μελετώνται διάφορα προβλήματα που εμφανίζει ο mRMR και προτείνονται παραλλαγές ώστε να αντιμετωπιστούν. Αρχικά προτείνεται μια άλλη προσέγγιση σχετικά με την εκτίμηση της περιττής πληροφορίας, ακολουθεί η προσπάθεια βελτιστοποίησης του κριτηρίου αξιολόγησης με μεθόδους διαφορετικές από την forward αναζήτηση και τέλος παρουσιάζεται μια τροποποίηση της συνάρτησης αξιολόγησης.

Το κεφάλαιο 5 αφορά στην πειραματική αποτίμηση των προτεινόμενων τροποποιήσεων. Περιγράφεται αναλυτικά η πειραματική μεθοδολογία και τα σύνολα

δεδομένων που χρησιμοποιήθηκαν, παρουσιάζονται και αναλύονται τα αποτελέσματα και γίνεται αξιολόγηση των προτεινόμενων τροποποιημένων προσεγγίσεων.

ΚΕΦΑΛΑΙΟ 2. ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΚΑΙ ΣΧΕΤΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

2.1 Το πρόβλημα της ταξινόμησης
2.2 Επιλογή χαρακτηριστικών για προβλήματα ταξινόμησης
2.3 Σημαντικές έννοιες για το πρόβλημα της επιλογής χαρακτηριστικών
2.4 Τεχνικές αναζήτησης
2.5 Επιλογή χαρακτηριστικών με wrapper
2.6 Επιλογή χαρακτηριστικών με φίλτρα
2.7 Embedded μέθοδοι

Στο κεφάλαιο αυτό αρχικά παρουσιάζονται κάποιες βασικές έννοιες σχετικά με το πρόβλημα της ταξινόμησης και της επιλογής χαρακτηριστικών. Κατόπιν περιγράφονται οι μέθοδοι αναζήτησης που χρησιμοποιούνται συνήθως για να ερευνηθεί ο χώρος των υποψήφιων υποσυνόλων. Στη συνέχεια, με βάση τον τρόπο που γίνεται η αποτίμηση των υποσυνόλων χαρακτηριστικών, διακρίνονται τρεις βασικές κατηγορίες μεθόδων και παρουσιάζονται κάποιες αντιπροσωπευτικές μέθοδοι για κάθε κατηγορία.

2.1. Το πρόβλημα της ταξινόμησης

Η *ταξινόμηση (classification)* είναι μια διαδικασία κατάταξης αντικειμένων σε κατηγορίες ανάλογα με τα χαρακτηριστικά τους. Πιο συγκεκριμένα, σκοπός είναι η κατάταξη ενός διανύσματος $x \in X$, που αποτελεί την αναπαράσταση ενός αντικειμένου ως προς κάποια χαρακτηριστικά, σε μία κατηγορία $y \in Y$ όπου Y ένα σύνολο διακριτών τιμών. Μας ενδιαφέρει η κατάταξη να γίνει με βάση μία απεικόνιση $f : X \to Y$ και στόχος είναι να εκτιμηθεί η f από ένα σύνολο

παραδειγμάτων. Τα παραδείγματα είναι στην ουσία διανύσματα $x_1,\ldots,x_n \in X$ που έχουν ήδη καταταχθεί σε κατηγορίες $y_1,\ldots,y_n \in Y$. Το σύνολο των παραδειγμάτων $D=\{(x_1,y_1),\ldots,(x_n,y_n)\}$ ονομάζεται σύνολο εκπαίδευσης.

Χρησιμοποιώντας το σύνολο D πρέπει να κατασκευαστεί μία απεικόνιση $g: X \to Y$ που θα είναι καλή προσέγγιση της *f*. Η διαδικασία κατασκευής της *g* υλοποιείται από έναν *αλγόριθμο επαγωγής (induction algorithm)*. Η συμπεριφορά της *g* καθορίζεται από ένα σύνολο παραμέτρων, έστω *θ*, και ο αλγόριθμος επαγωγής επιλέγει τις παραμέτρους *θ* ώστε η *g* να κατατάσσει κατά το δυνατόν σωστότερα τα παραδείγματα του συνόλου D. Η απεικόνιση *g* ονομάζεται συνήθως *ταξινομητής (classifier)* και η παραπάνω διαδικασία επιλογής των κατάλληλων παραμέτρων είναι γνωστή ως εκπαίδευση του ταξινομητή μιας και ο ταξινομητής εκπαιδεύεται από τα παραδείγματα του συνόλου εκπαίδευσης.

Αφού επιλεγούν οι παράμετροι *θ*, πρέπει να ελεγχθεί αν όντως η απεικόνιση *g* είναι καλή προσέγγιση της *f*. Αυτό μπορεί να γίνει ελέγχοντας αν η *g* κατατάσσει σωστά κάποια παραδείγματα των οποίων η κατηγορία είναι γνωστή εκ των προτέρων. Το ποσοστό σωστών κατατάξεων ονομάζεται *ακρίβεια κατάταξης (classification accuracy)* και είναι το μέτρο που χρησιμοποιείται συνήθως για την αξιολόγηση ταξινομητών. Για τη μέτρηση της ακρίβειας κατάταξης δεν πρέπει να χρησιμοποιηθούν τα παραδείγματα του συνόλου εκπαίδευσης D γιατί αυτά χρησιμοποιήθηκαν ήδη για τον καθορισμό των παραμέτρων θ και είναι αναμενόμενο ότι η *g* θα τα κατατάσσει σωστά. Για το λόγο αυτό χρησιμοποιείται ένα σύνολο παραδειγμάτων που δεν λήφθηκαν υπόψιν κατά τη διαδικασία της εκπαίδευσης το οποίο ονομάζεται *σύνολο ελέγχου (test set)*.

Συχνά η απεικόνιση *g*, που παράγεται από τη διαδικασία εκπαίδευσης, ταιριάζει πολύ καλά στα δεδομένα εκπαίδευσης (δηλαδή τα κατατάσσει σωστά) αλλά παρόλα αυτά έχει κακή ικανότητα γενίκευσης, δίνει δηλαδή μεγάλο σφάλμα κατάταξης στα δεδομένα του συνόλου ελέγχου. Μάλιστα είναι δυνατόν κάποια άλλη απεικόνιση *g'* που δεν ταιριάζει τόσο καλά στα δεδομένα εκπαίδευσης, να έχει καλύτερη ικανότητα γενίκευσης. Το φαινόμενο αυτό ονομάζεται *υπερεκπαίδευση (overfitting)* και εμφανίζεται όταν μια απεικόνιση προσαρμόζεται πολύ καλά στα δεδομένα εκπαίδευσης με αποτέλεσμα να επηρεάζεται από τον τυχαίο θόρυβο που υπάρχει σε

αυτά. Για την αντιμετώπιση της υπερεκπαίδευσης, που είναι ιδιαίτερα έντονη όταν τα δεδομένα εκπαίδευσης είναι λίγα και έχουν μεγάλη διάσταση, μπορεί να χρησιμοποιηθεί, ανάμεσα σε άλλες μεθόδους, η μείωση της διάστασης μέσω επιλογής χαρακτηριστικών που είναι το αντικείμενο αυτής της εργασίας.

2.2. Επιλογή χαρακτηριστικών για προβλήματα ταξινόμησης

Σε ένα πρόβλημα ταξινόμησης υπάρχουν πολλά χαρακτηριστικά, των οποίων η γνώση δεν είναι απαραίτητη για το διαχωρισμό των αντικειμένων σε κατηγορίες. Αυτό συμβαίνει είτε γιατί τα χαρακτηριστικά αυτά δίνουν πληροφορία άσχετη με το πρόβλημα είτε γιατί είναι περιττά, περιέχουν δηλαδή πληροφορία που δίνεται και από άλλα χαρακτηριστικά. Στην επιλογή χαρακτηριστικών ζητούμενο είναι ο εντοπισμός εκείνων των χαρακτηριστικών που είναι απαραίτητα για τη σωστή κατάταξη των αντικειμένων σε κατηγορίες. Απώτερος στόχος είναι να χρησιμοποιηθεί μόνο αυτό το υποσύνολο χρήσιμων χαρακτηριστικών κατά τη διαδικασία της ταξινόμησης και να αγνοηθούν τα υπόλοιπα χαρακτηριστικά, άσχετα ή περιττά. Θεωρητικά, εφόσον αγνοηθούν μόνο άσχετα ή περιττά χαρακτηριστικά, η απόδοση του ταξινομητή που θα κατασκευαστεί δεν θα είναι χειρότερη από την απόδοση που θα επιτυγχανόταν αν χρησιμοποιούνταν όλα τα χαρακτηριστικά. Πολλές φορές στην πράξη, η απόδοση όχι μόνο δεν χειροτερεύει αλλά βελτιώνεται σημαντικά γιατί χάρη στη μείωση της διάστασης αμβλύνεται το φαινόμενο της υπερεκπαίδευσης.

Η επιλογή χαρακτηριστικών διαφέρει από άλλες μεθόδους μείωσης της διάστασης όπως η εξαγωγή χαρακτηριστικών ως προς το ότι δεν μετασχηματίζει το χώρο των δεδομένων, απλά απορρίπτει τις συνιστώσες που είναι λιγότερο σημαντικές. Αυτή η διαφορά μπορεί να είναι πολύ σημαντική για κάποιες εφαρμογές.

Εκτός από τη βελτίωση της απόδοσης, περισσότερα οφέλη μπορεί να προκύψουν μέσω της επιλογής χαρακτηριστικών. Ένα από αυτά είναι ότι μία απεικόνιση που ορίζεται βάσει λίγων χαρακτηριστικών είναι πιο εύκολα ερμηνεύσιμη από μία απεικόνιση που ορίζεται βάσει πολλών χαρακτηριστικών. Επίσης, διακρίνοντας ποια χαρακτηριστικά είναι πιο σημαντικά για το αποτέλεσμα μιας διαδικασίας, μπορεί να αποκτηθεί διαίσθηση για το πραγματικό πρόβλημα, επιτρέποντας στους ειδικούς του τομέα να το αντιμετωπίσουν αποτελεσματικότερα. Αυτή η πτυχή της επιλογής

χαρακτηριστικών είναι ιδιαιτέρως σημαντική για προβλήματα βιοπληροφορικής όπου για παράδειγμα δίνει τη δυνατότητα να αναγνωριστούν γονίδια που σχετίζονται με διάφορες νόσους. Τέλος υπάρχει κέρδος όσον αφορά στο υπολογιστικό κόστος. Χρησιμοποιώντας λιγότερα χαρακτηριστικά, η εκπαίδευση ταξινομητών είναι γρηγορότερη, το ίδιο και η διαδικασία της κατάταξης, ενώ μειώνονται και οι απαιτήσεις σε αποθηκευτικό χώρο.

2.3. Σημαντικές έννοιες για το πρόβλημα της επιλογής χαρακτηριστικών

Περιγράφοντας αφηρημένα την εκπαίδευση ενός ταξινομητή, θα λέγαμε ότι πρόκειται για μία διαδικασία κατά την οποία αναζητούνται στα δεδομένα εκπαίδευσης συσχετίσεις μεταξύ των τιμών των χαρακτηριστικών και της τιμής της κατηγορίας. Στη συνέχεια, στη φάση της κατάταξης οι συσχετίσεις αυτές χρησιμοποιούνται για να κατατάξουν σε κατηγορίες άγνωστα παραδείγματα. Για παράδειγμα, μπορεί να δίνεται ένα σύνολο εκπαίδευσης στο οποίο παρατηρείται ότι για τα παραδείγματα που το χαρακτηριστικό X_1 έχει τιμή 0, η κατηγορία έχει τιμή 1, ενώ όταν το X_1 έχει τιμή 1, η κατηγορία έχει τιμή 0. Με βάση αυτή τη συσχέτιση μπορούμε να συμπεράνουμε ότι η τιμή του χαρακτηριστικού X_1 επηρεάζει την τιμή της κατηγορίας και έτσι να κατατάσσουμε μελλοντικά παραδείγματα στην κατηγορία 0 ή 1 ανάλογα με την τιμή του X_1.

Γενικά σε ένα σύνολο δεδομένων, παρατηρούνται συσχετίσεις μεταξύ της τιμής κάποιων χαρακτηριστικών και της τιμής της κατηγορίας. Με βάση τη συζήτηση της ενότητας 2.1 θα λέγαμε ότι η έξοδος της άγνωστης απεικόνισης f φαίνεται να επηρεάζεται από αυτά τα χαρακτηριστικά, τα οποία ονομάζονται *συναφή (relevant)*. Προφανώς τα συναφή χαρακτηριστικά είναι χρήσιμα για τη διαδικασία της ταξινόμησης γιατί με βάση αυτά ορίζονται οι αντιστοιχήσεις που χρησιμοποιούνται για την κατάταξη των παραδειγμάτων. Τα συναφή χαρακτηριστικά δίνουν πληροφορία για την κατηγορία με την έννοια ότι αν δεν υπήρχαν, η κατάταξη των άγνωστων παραδειγμάτων θα γινόταν με τυχαίο τρόπο. Προφανώς, αν ζητείται η επιλογή ενός υποσυνόλου χαρακτηριστικών ώστε να χρησιμοποιηθεί για τη διαδικασία της ταξινόμησης, τα συναφή χαρακτηριστικά είναι αυτά που πρέπει να επιλεγούν.

Στη συνέχεια ορίζεται σε μαθηματική βάση η έννοια της *συνάφειας (relevance)* που περιγράφηκε έως τώρα αφηρημένα. Παράλληλα με τους ορισμούς που δίνονται, εξετάζονται κάποιες σημαντικές έννοιες σημαντικές για το πρόβλημα της επιλογής χαρακτηριστικών. Οι ορισμοί που δίνονται εκφράζονται βάσει πιθανοτήτων. Η τυχαία μεταβλητή Y αντιστοιχεί στην κατηγορία και η τυχαία μεταβλητή X_i αντιστοιχεί στο i-οστό χαρακτηριστικό. Σύμφωνα με τον πρώτο ορισμό, ένα χαρακτηριστικό είναι συναφές αν η γνώση του αλλάζει την εκτίμηση για την τιμή της κατηγορίας.

Ορισμός 1*: Ένα χαρακτηριστικό X_i είναι συναφές αν και μόνο αν υπάρχουν τιμές x_i και y με $P(X_i=x_i)>0$ τέτοια ώστε $P(Y=y|X_i=x_i)\neq P(Y=y)$.*

Ο ορισμός αυτός δεν είναι πλήρης γιατί είναι πιθανό, ένα χαρακτηριστικό που δεν δίνει πληροφορία για την κατηγορία από μόνο του, να είναι χρήσιμο σε συνδυασμό με άλλα χαρακτηριστικά. Για την περιγραφή της κατάστασης αυτής, όπου μία ομάδα χαρακτηριστικών δρουν συμπληρωματικά, χρησιμοποιείται συνήθως ο όρος *αλληλεπίδραση (interaction)*. Ένα απλό παράδειγμα χαρακτηριστικών που αλληλεπιδρούν είναι το πρόβλημα XOR. Έστω ότι ο χώρος X περιλαμβάνει 2 δυαδικά χαρακτηριστικά X_1, X_2 και έστω μία απεικόνιση $f:X\rightarrow Y$, με $f(X_1,X_2) = X_1 \oplus X_2$, όπου $\oplus$ το σύμβολο XOR (exclusive or). Η έξοδος της f έχει τιμή 1 όταν μόνο μία από τις μεταβλητές X_1, X_2 ισούται με 1, και τιμή 0 σε κάθε άλλη περίπτωση. Έστω ότι δίδονται τα δεδομένα εκπαίδευσης του πίνακα 2.1.

Πίνακας 2.1 Το πρόβλημα XOR

Y	X_1	X_2
0	0	0
1	0	1
1	1	0
0	1	1

Από τα δεδομένα εκπαίδευσης παρατηρούμε ότι η κατηγορία μπορεί να παίρνει την τιμή μηδέν ή ένα με την ίδια πιθανότητα. Οι πιθανότητες για την τιμή της κατηγορίας παραμένουν ίδιες ακόμα και όταν το χαρακτηριστικό X_1 είναι γνωστό. Συγκεκριμένα όταν $X_1=0$ τότε $P(Y=0|X_1=0)=P(Y=1|X_1=0)=1/2$ και ομοίως όταν $X_1=1$ τότε

$P(Y=0|X_1=1)=P(Y=1|X_1=1)=1/2$. Το ίδιο ισχύει και για το χαρακτηριστικό X_2. Αν και τα δύο χαρακτηριστικά όταν κρίνονται μεμονωμένα φαίνονται να μην είναι χρήσιμα, ο συνδυασμός τους καθορίζει με μοναδικό τρόπο την τιμή της κατηγορίας.

Ένας δεύτερος ορισμός που λαμβάνει υπόψιν τα παραπάνω προβλέπει ότι ένα χαρακτηριστικό X_i είναι συναφές αν η απουσία του οδηγεί στο να μην μπορεί να υπάρξει μονοσήμαντη αντιστοίχηση των τιμών των χαρακτηριστικών και της τιμής της κατηγορίας. Το X_i είναι δηλαδή συναφές, αν υπάρχουν δύο παραδείγματα Α και Β που ανήκουν σε διαφορετικές κατηγορίες και διαφέρουν μόνο στην τιμή τους για το χαρακτηριστικό X_i [6]. Έστω S το σύνολο όλων των χαρακτηριστικών $X_1,...,X_m$ και S_i το σύνολο όλων των χαρακτηριστικών εξαιρουμένου του X_i.

Ορισμός 2*: Ένα χαρακτηριστικό X_i είναι συναφές αν και μόνο αν υπάρχουν x_i, s_i και y με $P(S_i=s_i, X_i=x_i)>0$ τέτοια ώστε $P(Y=y|S_i=s_i,X_i=x_i)\neq P(Y=y|S_i=s_i)$.*

Ακόμα και αυτός ο ορισμός δεν είναι πλήρως επαρκής αφού, αν υπάρχουν πολλά χαρακτηριστικά στο σύνολο δεδομένων, είναι πολύ πιθανό να μην υπάρχει κανένα ζεύγος παραδειγμάτων που να διαφέρει μόνο σε ένα χαρακτηριστικό και έτσι όλα τα χαρακτηριστικά μπορούν να θεωρηθούν μη-συναφή. Στην εργασία [21] οι συγγραφείς εκφράζουν την άποψη ότι πρέπει να διακριθούν δύο είδη συνάφειας, ισχυρής και αδύναμης. Στην πρώτη κατηγορία ανήκουν χαρακτηριστικά για τα οποία ισχύει ο ορισμός 2. Αδύναμα-συναφή είναι τα χαρακτηριστικά που γίνονται ισχυρά-συναφή αν αφαιρεθεί ένα υποσύνολο χαρακτηριστικών.

Ορισμός 3*: Ένα χαρακτηριστικό X_i είναι αδύναμα-συναφές αν και μόνο αν δεν είναι ισχυρά-συναφές και υπάρχει ένα υποσύνολο S_i' του S_i για το οποίο υπάρχουν x_i, s_i' και y με $P(S_i'=s_i',X_i=x_i)>0$, τέτοια ώστε $P(Y=y|S_i'=s_i',X_i=x_i)\neq P(Y=y|S_i'=s_i')$.*

Ένα ισχυρά-συναφές χαρακτηριστικό δεν πρέπει να αγνοηθεί γιατί τότε δεν μπορεί να υπάρξει μονοσήμαντη αντιστοιχία μεταξύ της τιμής των χαρακτηριστικών και της τιμής της κατηγορίας στα δεδομένα εκπαίδευσης. Από την άλλη, ένα αδύναμα συναφές χαρακτηριστικό, έστω X_i, περιέχει μεν πληροφορία για την κατηγορία, αλλά μπορεί να αγνοηθεί γιατί υπάρχει κάποιο υποσύνολο χαρακτηριστικών που δίνει την ίδια πληροφορία. Λέγεται ότι το χαρακτηριστικό X_i καθίσταται *περιττό (redundant)* λόγω της παρουσίας των υπόλοιπων. Το X_i είναι δυνατόν να καταστεί απαραίτητο αν αφαιρεθούν κάποια χαρακτηριστικά. Ένα χαρακτηριστικό που δεν είναι ισχυρά ή

αδύναμα συναφές, ονομάζεται *μη-συναφές (irrelevant)*. Ένα τέτοιο χαρακτηριστικό δεν προσφέρει σε καμία περίπτωση πληροφορία για την κατηγορία.

Σύμφωνα με τα παραπάνω, στόχος στην επιλογή χαρακτηριστικών για το πρόβλημα της ταξινόμησης είναι η εύρεση ενός υποσυνόλου συναφών και μη περιττών χαρακτηριστικών. Η επιλογή περιττών και μη-συναφών χαρακτηριστικών αυξάνει τη διάσταση χωρίς να δίνει επιπλέον πληροφορία και έτσι αυξάνει τον κίνδυνο να υπάρξει υπερεκπαίδευση.

2.4. Τεχνικές αναζήτησης

Το πρόβλημα της επιλογής χαρακτηριστικών είναι δύσκολο για πολλούς λόγους. Ένας από αυτούς είναι ότι το πλήθος των πιθανών λύσεων στο πρόβλημα αυξάνεται εκθετικά ως προς το πλήθος των χαρακτηριστικών. Αν υποθέσουμε ότι υπάρχει ένα κριτήριο αξιολόγησης *J* βάσει του οποίου μπορεί να αποτιμηθεί η ποιότητα ενός υποσυνόλου χαρακτηριστικών, και ότι το πλήθος των χαρακτηριστικών είναι N, τότε 2^N διαφορετικά υποσύνολα πρέπει να αποτιμηθούν προκειμένου να βρεθεί το καλύτερο. Προφανώς η εξαντλητική εξερεύνηση του χώρου των υποσυνόλων δεν είναι εφικτή και έτσι διαφορές τεχνικές αναζήτησης επιστρατεύονται ώστε να επιτευχθεί η εύρεση ενός αρκετά καλού υποσυνόλου μέσα σε λογικά χρονικά πλαίσια. Το κριτήριο *J* χρησιμοποιείται όχι μόνο για την επιλογή του καλύτερου υποσυνόλου, αλλά και για να καθοδηγήσει την αναζήτηση. Στη συνέχεια παρουσιάζονται συνοπτικά μερικές συχνά χρησιμοποιούμενες τεχνικές αναζήτησης όπως η προς-τα-εμπρός επιλογή (forward selection) και η προς-τα-πίσω απαλοιφή (backward elimination) που είναι δύο διαφορετικές εκδοχές άπληστης αναζήτησης, η αναζήτηση πρώτα στο καλύτερο (best first search), οι γενετικοί αλγόριθμοι κ.α.

Forward selection

Ο αλγόριθμος χτίζει αυξητικά το υποσύνολο των επιλεγμένων χαρακτηριστικών. Αρχικά το υποσύνολο επιλεγμένων χαρακτηριστικών είναι κενό. Σε ένα τυπικό βήμα του αλγόριθμου, εξετάζονται όλα τα υποσύνολα που προκύπτουν από την προσθήκη ενός χαρακτηριστικού στο τρέχον υποσύνολο. Το χαρακτηριστικό που οδηγεί στη

μεγαλύτερη αύξηση απόδοσης σύμφωνα με το κριτήριο αποτίμησης ενσωματώνεται στο τρέχον υποσύνολο. Η επέκταση με την προσθήκη ενός χαρακτηριστικού κάθε φορά συνεχίζεται ωσότου ικανοποιηθεί κάποια συνθήκη τερματισμού.

Συνήθως η επέκταση σταματάει όταν κανέναν από τα υποσύνολα-παιδιά δεν οδηγεί σε βελτίωση της απόδοσης. Μια τόσο αυστηρή συνθήκη τερματισμού μπορεί να οδηγήσει τον αλγόριθμο σε πρόωρο σταμάτημα. Μια πιο χαλαρή συνθήκη τερματισμού προβλέπει τη συνέχιση των επεκτάσεων εφόσον υπάρχει κάποιο υποσύνολο-παιδί που οδηγεί σε απόδοση το ίδιο καλή με την έως τώρα καλύτερη και τερματισμό αν η απόδοση δεν βελτιωθεί μετά από N διαδοχικές επεκτάσεις.

Backward elimination

Το αρχικό υποσύνολο περιέχει όλα τα χαρακτηριστικά. Ακολουθώντας πορεία αντίστροφη σε σχέση με την προηγούμενη μέθοδο, σε κάθε επανάληψη εξετάζονται όλα τα υποσύνολα που προκύπτουν από τη διαγραφή ενός χαρακτηριστικού από το τρέχον υποσύνολο. Τελικά διαγράφεται αυτό του οποίου η απουσία οδηγεί στη μεγαλύτερη απόδοση ως προς το κριτήριο αξιολόγησης.

Best first search

Στις δύο προηγούμενες μεθόδους, υπάρχει μία κατάσταση (ένα υποσύνολο) η οποία επεκτείνεται είτε προσθέτοντας είτε διαγράφοντας χαρακτηριστικά. Οι καταστάσεις παιδιά που προκύπτουν αξιολογούνται, η καλύτερη ανάμεσα τους επιλέγεται και επεκτείνεται ενώ οι υπόλοιπες αγνοούνται. Στην παρούσα μέθοδο δεν αγνοείται καμία κατάσταση, αντίθετα όλες οι καταστάσεις-παιδιά αφού αξιολογηθούν τοποθετούνται σε μία λίστα. Στη συνέχεια, σε κάθε βήμα εντοπίζεται η καλύτερη κατάσταση από τη λίστα και αυτή είναι που επεκτείνεται. Οι καταστάσεις-παιδιά που προκύπτουν με τη σειρά τους αποτιμώνται και τοποθετούνται στη λίστα. Η διαδικασία τερματίζει αν μετά από N διαδοχικές επεκτάσεις δεν προκύψει κάποια κατάσταση που να βελτιώνει την απόδοση. Η μέθοδος έχει μεγάλη πολυπλοκότητα.

Plus l– take away r

Σε αυτή τη μέθοδο l βήματα forward selection ακολουθούνται από r βήματα backward elimination. Σε μια τυπική επανάληψη του αλγορίθμου, ξεκινώντας από ένα υποσύνολο μεγέθους N, ακολουθεί η προσθήκη l χαρακτηριστικών. Στη συνέχεια

εξετάζονται τα σύνολα που προκύπτουν από τη διαγραφή ενός χαρακτηριστικού με σκοπό να βρεθεί κάποιο υποσύνολο μεγέθους N+ℓ-1 με απόδοση καλύτερη από το προηγούμενο ίδιου μεγέθους. Το σκεπτικό είναι ότι ένα χαρακτηριστικό που συμπεριλήφθηκε αρχικά στο υποσύνολο γιατί έδωσε μεγάλη απόδοση όταν κρίθηκε μεμονωμένα, μπορεί να μην συνδυάζεται τόσο καλά με χαρακτηριστικά που προστέθηκαν στη συνέχεια. Αντίθετα κάποιο χαρακτηριστικό που από μόνο του δεν φαίνεται αξιόλογο, μπορεί να δίνει έναν ισχυρό συνδυασμό με χαρακτηριστικά που προστέθηκαν πρόσφατα. Τελικά γίνονται r διαγραφές, πριν η μέθοδος συνεχίσει και πάλι με την προσθήκη ℓ χαρακτηριστικών.

Floating search

Η μέθοδος αυτή [27] είναι παρόμοια με την προηγούμενη μέθοδο. Η διαφορά τους έγκειται στο ότι ο αριθμός των backward βημάτων που ακολουθούν τα forward βήματα δεν είναι προκαθορισμένος. Τα backward βήματα συνεχίζονται όσο οι διαγραφές βελτιώνουν το κριτήριο αξιολόγησης σε σχέση με την προηγούμενη καλύτερη τιμή.

Γενετικοί αλγόριθμοι

Ένας γενετικός αλγόριθμος είναι μία στοχαστική μέθοδος βελτιστοποίησης η οποία είναι εμπνευσμένη από το τη βιολογική διαδικασία της φυσικής επιλογής. Αρχικά δίνεται ένα σύνολο από πιθανές λύσεις του προβλήματος, που ονομάζεται πληθυσμός. Τα στοιχεία του πληθυσμού αποτιμώνται με βάση κάποια συνάρτηση αξιολόγησης. Κάποια από τα στοιχεία του πληθυσμού επιλέγονται για τη φάση της αναπαραγωγής. Τα επιλεγμένα στοιχεία συνδυάζονται μεταξύ τους και παράγουν την επόμενη γενιά του πληθυσμού. Η πιθανότητα επιλογής ενός στοιχείου για αναπαραγωγή είναι ανάλογη του πόσο καλό είναι σύμφωνα με τη συνάρτηση αποτίμησης. Έτσι υπάρχει η τάση με την πάροδο των γενεών να «επιβιώσουν» στοιχεία του πληθυσμού με καλά χαρακτηριστικά και από τους συνδυασμούς τους να προκύψουν καλές λύσεις.

Στο πρόβλημα της επιλογής χαρακτηριστικών, τα στοιχεία του πληθυσμού είναι υποσύνολα χαρακτηριστικών που αναπαριστώνται από δυαδικά διανύσματα. Κατά τη φάση της αναπαραγωγής, δύο υποσύνολα S_1, S_2 συνδυάζονται και δίνουν ένα νέο που περιέχει ένα μέρος των χαρακτηριστικών του S_1 και ένα μέρος των χαρακτηριστικών

του S_2. Η συνάρτηση που χρησιμοποιείται για την αποτίμηση είναι συνήθως η απόδοση του αντίστοιχου ταξινομητή.

2.5. Επιλογή χαρακτηριστικών με wrapper

Στην κατηγορία των wrapper μεθόδων εντάσσονται όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιούν την ακρίβεια ταξινόμησης ως κριτήριο αξιολόγησης των υποσυνόλων. Η αξιολόγηση με βάση την απόδοση του ταξινομητή απαιτεί την κατασκευή του ταξινομητή για κάθε υποσύνολο χαρακτηριστικών που εξετάζεται και έχει ως αρνητικό επακόλουθο το αυξημένο υπολογιστικό κόστος σε σχέση με τις πιο εξελιγμένες embedded μεθόδους ή τα φίλτρα. Παρόλα αυτά οι wrapper έχουν και κάποια πλεονεκτήματα που κάνουν τη χρήση τους ελκυστική ιδίως σε συνδυασμό με φίλτρα για την ανάπτυξη υβριδικών αλγορίθμων.

2.5.1. Η διαδικασία αξιολόγησης με χρήση wrapper

Η αξιολόγηση ενός υποψήφιου υποσυνόλου, έστω S, γίνεται ως εξής. Τα δεδομένα εκπαίδευσης χωρίζονται σε δύο σύνολα, εκπαίδευσης (training) και *επικύρωσης (validation)*. Από τα παραδείγματα και των δύο συνόλων διαγράφονται τα χαρακτηριστικά που δεν ανήκουν στο υποψήφιο υποσύνολο S. Ο ταξινομητής εκπαιδεύεται στο τροποποιημένο σύνολο εκπαίδευσης που προκύπτει και στη συνέχεια κατατάσσει τα παραδείγματα του τροποποιημένου συνόλου επικύρωσης. Η ακρίβεια ταξινόμησης που επιτυγχάνεται στο σύνολο επικύρωσης είναι το κριτήριο αξιολόγησης.

Όταν τα διαθέσιμα παραδείγματα είναι λίγα δεν υπάρχει περιθώριο να σχηματιστεί μεγάλο σύνολο επικύρωσης γιατί τότε το σύνολο εκπαίδευσης γίνεται υπερβολικά μικρό. Μικρό όμως σύνολο επικύρωσης συνεπάγεται μη αξιόπιστη εκτίμηση της ακρίβειας. Σε αυτή την περίπτωση χρησιμοποιείται συνήθως η τεχνική cross validation [5]. Τα παραδείγματα εκπαίδευσης χωρίζονται σε k ξένα υποσύνολα (υποσύνολα παραδειγμάτων). Ο ταξινομητής εκπαιδεύεται στα παραδείγματα των k-1 υποσυνόλων ενώ ένα υποσύνολο παίζει το ρόλο του συνόλου επικύρωσης. Η διαδικασία επαναλαμβάνεται k φορές έτσι ώστε κάθε υποσύνολο να παίξει το ρόλο

του συνόλου επικύρωσης μία φορά. Ο μέσος όρος της ακρίβειας ταξινόμησης στα διαφορετικά σύνολα επικύρωσης είναι το κριτήριο αξιολόγησης.

Οι wrapper μπορούν να συνδυαστούν με οποιαδήποτε μέθοδο αναζήτησης όπως για παράδειγμα με τις μεθόδους που παρουσιάστηκαν στην ενότητα 2.4. Πιο συνηθισμένες στη βιβλιογραφία είναι οι μέθοδοι forward selection και backward elimination.

2.5.2. Χαρακτηριστικά των wrapper

Οι wrapper μεθοδολογίες μπορούν να χρησιμοποιηθούν με οποιονδήποτε ταξινομητή καθώς δεν εξαρτώνται από τον τρόπο λειτουργίας τους, παρά μόνο χρησιμοποιούν την απόδοση τους για να αξιολογήσουν υποψήφια υποσύνολα χαρακτηριστικών. Το ισχυρότερο επιχείρημα υπέρ της χρήσης των wrapper είναι ότι λαμβάνουν υπόψη την *επαγωγική πόλωση (inductive bias)* του ταξινομητή [21],[6]. Κάθε ταξινομητής έχει τα δικά του ιδιαίτερα χαρακτηριστικά και τον δικό του τρόπο που απεικονίζει την είσοδο που δέχεται σε έξοδο. Αυτό σημαίνει ότι το καλύτερο υποσύνολο χαρακτηριστικών για έναν ταξινομητή δεν είναι απαραίτητα το καλύτερο υποσύνολο για έναν ταξινομητή άλλου τύπου. Η ακρίβεια ταξινόμησης είναι το πιο αξιόπιστο κριτήριο για να ελεγχθεί αν ένα υποσύνολο χαρακτηριστικών δουλεύει καλά σε συνδυασμό με έναν ταξινομητή.

Θεωρητικά η χρήση wrapper δίνει τη δυνατότητα ανακάλυψης αλληλεπιδράσεων μεταξύ χαρακτηριστικών κι αυτό γιατί τα χαρακτηριστικά δεν αξιολογούνται μεμονωμένα αλλά ως μέρη ενός υποσυνόλου. Φυσικά αν υπάρχουν χαρακτηριστικά που αλληλεπιδρούν, η ανακάλυψη τους εξαρτάται από το αν θα τύχει να βρεθούν στο ίδιο υποψήφιο υποσύνολο ώστε να αξιολογηθούν ως ομάδα. Τελικά αυτό είναι κάτι που εξαρτάται από το μηχανισμό αναζήτησης και κυρίως από το χρόνο που δίνεται στον αλγόριθμο για να εκτελεστεί.

Το βασικό μειονέκτημα των wrapper είναι το μεγάλο υπολογιστικό κόστος. Η αποτίμηση κάθε υποψήφιου υποσυνόλου συνεπάγεται την εκπαίδευση του ταξινομητή και τη μέτρηση της απόδοσης στο σύνολο επικύρωσης που είναι συνήθως μια αργή διαδικασία.

Τέλος, ένα μεγάλο πρόβλημα των wrapper είναι ότι η επιλογή υποσυνόλου με βάση την ακρίβεια ταξινόμησης μπορεί να αποδειχθεί τελείως αναξιόπιστη ως μέθοδος παρά το γεγονός ότι η ακρίβεια αυτή μετράται σε ένα ξεχωριστό σύνολο επικύρωσης που δεν χρησιμοποιείται κατά τη φάση της εκπαίδευσης. Επειδή ο αριθμός των υποσυνόλων που εξετάζονται από έναν wrapper είναι μεγάλος, είναι πολύ πιθανό να βρεθεί τελικά από τύχη ένα υποσύνολο που δίνει πολύ καλή απόδοση στο σύνολο επικύρωσης, χωρίς όμως να έχει καλή ικανότητα γενίκευσης [20],[28]. Την ίδια στιγμή μπορεί άλλα υποσύνολα με σημαντικά μικρότερη ακρίβεια ταξινόμησης στο σύνολο επικύρωσης να επιτυγχάνουν καλύτερη ικανότητα γενίκευσης. Το πρόβλημα αυτό είναι γενικά γνωστό ως το *πρόβλημα πολλαπλών συγκρίσεων* και είναι πολύ έντονο όταν το σύνολο με τα διαθέσιμα δεδομένα εκπαίδευσης είναι πολύ μικρό [29].

Στη βιβλιογραφία υπάρχουν πολλές εργασίες που έχουν χρησιμοποιήσει wrapper μεθοδολογίες παρόλα αυτά δεν θα γίνει κάποια λεπτομερέστερη ανάλυση καθώς οι μέθοδοι ακολουθούν την ίδια φιλοσοφία χωρίς να παρουσιάζουν ιδιαίτερες διαφορές. Μια επισκόπηση των wrapper μεθόδων παρουσιάζεται στα [21] και [6].

2.6. Επιλογή χαρακτηριστικών με φίλτρα

Στην κατηγορία των φίλτρων ανήκουν όσοι αλγόριθμοι δεν βασίζονται σε κάποιο ταξινομητή προκειμένου να εκτιμήσουν την ποιότητα ενός υποσυνόλου χαρακτηριστικών, αντίθετα χρησιμοποιώντας στατιστικά μέτρα προσπαθούν να εντοπίσουν συναφή χαρακτηριστικά. Δύο βασικές κατηγορίες φίλτρων μπορούν να διακριθούν που παρουσιάζονται στις επόμενες ενότητες.

2.6.1. Univariate μέθοδοι

Οι univariate μέθοδοι αξιολογούν κάθε χαρακτηριστικό με βάση τη συσχέτισή του με τις κατηγορίες. Όσο μεγαλύτερη συσχέτιση υπάρχει, τόσο πιο χρήσιμο θεωρείται το χαρακτηριστικό. Κατόπιν επιλέγονται τα N πιο συσχετισμένα χαρακτηριστικά όπου N ένας αριθμός που καθορίζεται εμπειρικά.

Οι univariate μέθοδοι εμφανίζουν κάποιες αδυναμίες με κυριότερη ότι δεν προλαμβάνουν την επιλογή περιττών χαρακτηριστικών, αφού κάθε χαρακτηριστικό

αξιολογείται ξεχωριστά χωρίς να λαμβάνεται υπόψιν ποια άλλα χαρακτηριστικά έχουν επιλεγεί. Προφανώς για τον ίδιο λόγο δεν είναι σε θέση να εντοπίσουν αλληλεπιδράσεις μεταξύ των χαρακτηριστικών. Παρόλα αυτά οι univariate μέθοδοι χρησιμοποιούνται αρκετά συχνά κυρίως σε προβλήματα βιοπληροφορικής.

Στη συνέχεια παρουσιάζονται μερικά από τα κριτήρια που έχουν χρησιμοποιηθεί στη βιβλιογραφία για τη μέτρηση της συσχέτισης. Το κριτήριο του Fischer μπορεί να χρησιμοποιηθεί σε προβλήματα δύο κατηγοριών. Σύμφωνα με αυτό η συσχέτιση του i-οστού χαρακτηριστικού υπολογίζεται ως [25]:

$$w_i = \frac{\left(\mu_{i1} - \mu_{i2}\right)^2}{\sigma_{i1}^2 + \sigma_{i2}^2} \qquad \text{Εξ. 2.1}$$

Τα μ_{i1} και μ_{i2} αντιστοιχούν στη μέση τιμή του i-οστού χαρακτηριστικού για τα παραδείγματα της πρώτης και δεύτερης κατηγορίας αντίστοιχα. Ομοίως, σ_{i1} και σ_{i2} είναι οι τυπικές αποκλίσεις του i-οστου χαρακτηριστικού για τα παραδείγματα της πρώτης και δεύτερης κατηγορίας αντίστοιχα. Μεγάλη τιμή του βάρους w_i σημαίνει ότι τα παραδείγματα της πρώτης κατηγορίας διαφέρουν σημαντικά από τα παραδείγματα της δεύτερης ως προς το χαρακτηριστικό i και επομένως το χαρακτηριστικό έχει ισχυρή συσχέτιση.

Το F-test [10],[12] μπορεί να χρησιμοποιηθεί για προβλήματα με K κατηγορίες και ορίζεται ως εξής:

$$w_j = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_j)^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2} \qquad \text{Εξ. 2.2}$$

όπου $\bar{x}_{kj}$ είναι η μέση τιμή του χαρακτηριστικού j για τα παραδείγματα της κατηγορίας c_k, ενώ $\bar{x}_j$ είναι η μέση τιμή του χαρακτηριστικού j υπολογισμένη με βάση όλα τα παραδείγματα. Η έκφραση I(A) ισούται με ένα αν η πρόταση A είναι αληθής, διαφορετικά ισούται με μηδέν. Ο αριθμητής αυξάνεται όταν οι μέσες τιμές $\bar{x}_{kj}$ διαφέρουν μεταξύ τους για διαφορετικά k και άρα διαφέρουν και από το $\bar{x}_j$. Ο παρονομαστής μειώνεται όταν υπάρχει μικρή διακύμανση μεταξύ παραδειγμάτων της ίδιας κατηγορίας ως προς το χαρακτηριστικό j.

Ένα άλλο μέτρο της εξάρτησης μεταξύ δύο τυχαίων μεταβλητών είναι η *αμοιβαία πληροφορία (mutual information)*. Η αμοιβαία πληροφορία, σε αντίθεση με τους προηγούμενους συντελεστές συσχέτισης, μπορεί να ανιχνεύσει μη γραμμικές εξαρτήσεις μεταξύ των μεταβλητών [4],[16]. Αν X και Y δύο διακριτές τυχαίες μεταβλητές με περιθώριες κατανομές p(x) και p(y) και από κοινού κατανομή p(x,y), η αμοιβαία πληροφορία τους I(X,Y) ισούται με:

$$I(X,Y) = \sum_{x,y} p(x,y) \log \frac{p(x,y)}{p(x)\,p(y)} \qquad \text{Εξ. 2.3}$$

Η τιμή της αμοιβαίας πληροφορίας είναι μηδέν όταν οι μεταβλητές X και Y είναι ανεξάρτητες και μεγαλύτερη του μηδενός σε κάθε άλλη περίπτωση. Η τιμή της αυξάνεται όσο μεγαλύτερη είναι η εξάρτηση μεταξύ των δύο μεταβλητών.

Η αμοιβαία πληροφορία σχετίζεται με την εντροπία που είναι μέτρο της αβεβαιότητας για την τιμή μιας τυχαίας μεταβλητής. Έστω ότι ζητείται η πρόβλεψη της κατηγορίας Y για ένα παράδειγμα για το οποίο δεν είναι γνωστό κανένα χαρακτηριστικό. Η πρόβλεψη της κατηγορίας θα πρέπει να γίνει αναγκαστικά στην τύχη. Η εντροπία της κατηγορίας Y εκφράζει την αβεβαιότητα που υπάρχει στην πρόβλεψη αυτή. Αν η κατηγορία Y παίρνει N διαφορετικές τιμές $y_1,\ldots,y_n$ με πιθανότητες $p(y_1),\ldots,p(y_N)$ αντίστοιχα, η εντροπία της υπολογίζεται ως:

$$H(Y) = -\sum_{i=1}^{N} p(y_i) \log(p(y_i)) \qquad \text{Εξ. 2.4}$$

Έστω τώρα ότι γνωρίζουμε ότι για το νέο παράδειγμα που πρέπει να ταξινομηθεί το χαρακτηριστικό X έχει τιμή x. Έχοντας παρατηρήσει στο σύνολο δεδομένων ότι υπάρχει συσχέτιση μεταξύ της τιμής του X και της τιμής της κατηγορίας, η πρόβλεψη μπορεί να γίνει με μικρότερη αβεβαιότητα που ισούται με:

$$H(Y \mid X = x) = -\sum_{i=1}^{N} p(y_i \mid x) \log(p(y_i \mid x)) \qquad \text{Εξ. 2.5}$$

Γενικά η γνώση της τιμής του X, βοηθάει στο να υπάρχει μικρότερη η αβεβαιότητα στην πρόβλεψη της Y που κατά μέσο όρο ισούται με:

$$H(Y \mid X) = -\sum_{j=1}^{m} p(x_j) H(Y \mid X = x_j) \qquad \text{Εξ. 2.6}$$

όπου $x_1, ..., x_m$ οι τιμές που παίρνει το χαρακτηριστικό X. Η μείωση της αβεβαιότητας που υπάρχει χάρη στη γνώση του χαρακτηριστικού X ισούται με την αμοιβαία πληροφορία του χαρακτηριστικού X με την κατηγορία Y [9], δηλαδή ισχύει:

$$I(X;Y) = H(Y) - H(Y \mid X) \qquad \text{Εξ. 2.7}$$

Άρα όσο μεγαλύτερη είναι η αμοιβαία πληροφορία ενός χαρακτηριστικού με την κατηγορία τόσο χρησιμότερο είναι το χαρακτηριστικό.

2.6.2. Multivariate μέθοδοι

Οι multivariate μέθοδοι αποτιμούν χαρακτηριστικά λαμβάνοντας υπόψιν την παρουσία άλλων, προσπαθώντας έτσι να αποφύγουν την επιλογή περιττών χαρακτηριστικών. Στο επόμενο κεφάλαιο ασχολούμαστε περισσότερο με τις multivariate μεθόδους, οπότε στην ενότητα αυτή παρουσιάζεται μόνο μία μέθοδος αυτής της κατηγορίας. Η μέθοδος προτείνεται στο [23] και βασίζεται στην έννοια του markov blanket. Έστω Y η κατηγορία, F το σύνολο όλων των χαρακτηριστικών, M_i ένα υποσύνολο του F και F_i ένα χαρακτηριστικό που ανήκει στο F αλλά δεν ανήκει στο M_i. Έστω επίσης F'=F-M_i-F_i. Το M_i είναι markov-blanket για το χαρακτηριστικό Fi αν ισχύει

$$P(Y = y, F' = f' \mid M_i = m_i, F_i = f_i) = P(Y = y, F' = f' \mid M_i = m_i) \qquad \text{Εξ. 2.8}$$

για οποιαδήποτε ανάθεση τιμών y,f',m_i,f_i. Η ισχύς της (2.8) συνεπάγεται ότι:

$$P(Y = y \mid M_i = m_i, F_i = f_i) = P(Y = y \mid M_i = m_i) \qquad \text{Εξ. 2.9}$$

Η ύπαρξη ενός markov-blanket M_i για ένα χαρακτηριστικό F_i επομένως σημαίνει ότι το F_i δεν δίνει επιπλέον πληροφορία για την κατηγορία από αυτή που δίνουν τα χαρακτηριστικά του M_i και επομένως μπορεί να αφαιρεθεί. Επιπλέον, αν υποθέσουμε ότι χρησιμοποιείται η backward elimination μέθοδος αναζήτησης, η αφαίρεση του F_i λόγω ύπαρξης του M_i, εγγυάται ότι κάποιο άλλο χαρακτηριστικό F_j που αφαιρέθηκε νωρίτερα, δεν θα καταστεί και πάλι χρήσιμο. Η ισχύς της (2.9) συνεπάγεται ότι

$$\sum_{m_i, f_i} P(M_i = m_i, F_i = f_i) \cdot D(P(Y \mid M_i = m_i, F_i = f_i), P(Y \mid M_i = m_i)) = 0 \qquad \text{Εξ. 2.10}$$

όπου $D(P,Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$ είναι η KL-απόσταση των κατανομών P και Q [9].

Όσο μικρότερο είναι το άθροισμα της εξίσωσης (2.10) τόσο μικρότερη είναι η πληροφορία που συνεισφέρει το F_i για την κατηγορία δοθέντος του M_i. Όταν το άθροισμα είναι μηδέν το χαρακτηριστικό F_i θεωρείται περιττό.

Επειδή η αναζήτηση όλων των συνδυασμών από χαρακτηριστικά που θα μπορούσαν να αποτελέσουν markov-blanket για ένα χαρακτηριστικό F_i δεν είναι εφικτή, ως προσέγγιση του markov-blanket για το F_i θεωρείται το σύνολο M_i που αποτελείται από τα k πιο όμοια με το F_i χαρακτηριστικά. Η προσέγγιση αυτή βασίζεται στο σκεπτικό ότι τα πιο όμοια χαρακτηριστικά είναι αυτά που έχουν τη μεγαλύτερη πιθανότητα να καταστήσουν το F_i περιττό. Η παράμετρος k επιλέγεται εκ των προτέρων εμπειρικά.

Ο αλγόριθμος χρησιμοποιεί μία backward-ellimination τεχνική αναζήτησης αφαιρώντας σε κάθε επανάληψη το χαρακτηριστικό που συνεισφέρει τη μικρότερη πληροφορία. Η συνεισφορά κάθε χαρακτηριστικού υπολογίζεται με βάση την εξίσωση (2.10).

2.7. Embedded μέθοδοι

Στην κατηγορία αυτή συναντούμε μεθόδους που έχουν σχεδιαστεί με στόχο να δουλεύουν σε συνεργασία με ένα ταξινομητή συγκεκριμένου τύπου. Σε αντίθεση με τους wrapper που απλώς χρησιμοποιούν την έξοδο ενός ταξινομητή, οι μέθοδοι αυτές επιλέγουν χαρακτηριστικά με βάση το πως επηρεάζεται κάποια συνάρτηση κόστους που εμπλέκεται στη διαδικασία εκπαίδευσης του ταξινομητή. Από την ενσωμάτωση της επιλογής χαρακτηριστικών στη διαδικασία εκπαίδευσης προκύπτουν διάφορα πλεονεκτήματα σε σχέση με τους wrapper όπως μεγάλο κέρδος σε υπολογιστικό κόστος. Επίσης, οι embedded μέθοδοι καταφέρνουν να κάνουν καλύτερη χρήση των διαθέσιμων δεδομένων αφού δεν υπάρχει η ανάγκη αυτά να χωριστούν σε σύνολα εκπαίδευσης και επικύρωσης. Σε σχέση με τα φίλτρα έχουν το πλεονέκτημα ότι

λαμβάνουν υπόψιν την επαγωγική πόλωση (inductive bias) του ταξινομητή. Μία καλή επισκόπηση των embedded μεθόδων γίνεται στο [16].

Ιδιαίτερα σημαντικός αριθμός embedded μεθόδων έχουν σχεδιαστεί για τον ταξινομητή SVM, εκ των οποίων δύο περιγράφονται ενδεικτικά στη συνέχεια. Ο ταξινομητής SVM (support vector machine) [7],[30] αρχικά σχεδιάστηκε για προβλήματα δύο κατηγοριών. Η κατάταξη ενός άγνωστου διανύσματος $x \in R^n$ γίνεται με βάση τη συνάρτηση:

$$f(x) = w^T x + b \qquad \text{Εξ. 2.11}$$

όπου w το διάνυσμα βαρών και b η πόλωση (στην εξίσωση 2.11 υποθέτουμε ότι χρησιμοποιείται γραμμική συνάρτηση πυρήνα (linear kernel)). Η εξίσωση $w^T x + b = 0$ ορίζει ένα υπερεπίπεδο διάστασης n-1 που χρησιμοποιείται για να διαχωρίσει τα αντικείμενα των δύο κατηγοριών. Ένα διάνυσμα x κατατάσσεται στην κατηγορία C_1 αν f(x)<0 και στην κατηγορία C_2 αν f(x)>0. Η εκπαίδευση του SVM γίνεται μέσω της λύσης ενός προβλήματος βελτιστοποίησης που οδηγεί στον υπολογισμό παράμερων w και b τέτοιων ώστε το αντίστοιχο υπερεπίπεδο να διαχωρίζει τα παραδείγματα των κατηγοριών C_1 και C_2 αφήνοντας όσο το δυνατόν μεγαλύτερο περιθώριο (περιθώριο είναι η απόσταση του υπερεπιπέδου από τα κοντινότερα σημεία των δύο κατηγοριών).

Ο αλγόριθμος SVM-RFE [17] εκτελεί μια backward elimination διαδικασία, σε κάθε βήμα της οποίας διαγράφεται το λιγότερο σημαντικό χαρακτηριστικό. Λιγότερο σημαντικό θεωρείται το χαρακτηριστικό του οποίου η διαγραφή θα προκαλούσε τη μικρότερη μείωση του περιθωρίου. Ο αλγόριθμος ξεκινά εκπαιδεύοντας τον ταξινομητή SVM χρησιμοποιώντας όλα τα χαρακτηριστικά. Από την εκπαίδευση υπολογίζονται το διάνυσμα βαρών w και η πόλωση b. Αν υποθέσουμε ότι χρησιμοποιείται γραμμικός πυρήνας, το χαρακτηριστικό που αντιστοιχεί στην ελάχιστη κατά απόλυτη τιμή συνιστώσα του διανύσματος w, είναι αυτό του οποίου η διαγραφή θα προκαλούσε τη μικρότερη μείωση του περιθωρίου. Το χαρακτηριστικό απορρίπτεται και ο ταξινομητής εκπαιδεύεται λαμβάνοντας υπόψιν τα εναπομείναντα χαρακτηριστικά. Με τον ίδιο τρόπο το λιγότερο σημαντικό χαρακτηριστικό διαγράφεται και η διαδικασία συνεχίζει ωσότου απομείνει ένας προκαθορισμένος αριθμός χαρακτηριστικών.

Ο αλγόριθμος SVM-RFE απαιτεί μία εκπαίδευση του ταξινομητή SVM σε κάθε backward βήμα προκειμένου να αποφασιστεί ποιο χαρακτηριστικό θα διαγραφεί. Χρειάζονται n εκπαιδεύσεις για ολόκληρη τη διαδικασία, όπου n ο αριθμός των χαρακτηριστικών που πρέπει να αφαιρεθούν. Μία wrapper μέθοδος που εκτελεί backward elimination αναζήτηση, απαιτεί τόσες εκπαιδεύσεις όσες και τα εναπομείναντα χαρακτηριστικά πριν αποφασίσει πιο θα αφαιρεθεί. Έτσι η διαγραφή n χαρακτηριστικών απαιτεί συνολικά $O(n^2)$ εκπαιδεύσεις.

Σε περίπτωση που δεν χρησιμοποιείται γραμμικός πυρήνας, η εύρεση του χαρακτηριστικού του οποίου η αφαίρεση οδηγεί στην μικρότερη μείωση του περιθωρίου απαιτεί πιο πολύπλοκους υπολογισμούς αλλά και πάλι χρειάζεται μία εκπαίδευση του SVM σε κάθε γύρο.

Στην εργασία [31] η ελαχιστοποίηση μίας συνάρτησης κόστους που αποτελεί ένα άνω φράγμα για την πιθανότητα λανθασμένης πρόβλεψης από το SVM κατά τη διαδικασία leave-one-out cross-validation [5],[22] χρησιμοποιείται για την επιλογή χαρακτηριστικών. Η ποσότητα που πρέπει να ελαχιστοποιηθεί είναι $R^2\|w\|^2$ όπου R το μέγεθος της μικρότερης σφαίρας που περικλείει τα δεδομένα εκπαίδευσης και w το διάνυσμα της εξίσωσης (2.11). Τα R και w εκφράζονται συναρτήσει ενός διανύσματος $\sigma \in R^n$ όπου σ_i ο συντελεστής κλιμάκωσης του i-οστου χαρακτηριστικού. Η εύρεση του καλύτερου υποσυνόλου m χαρακτηριστικών ανάγεται στην εύρεση ενός σ* τέτοιου ώστε $\sigma^* = \arg\min_{\sigma}\{R^2(\sigma)W^2(\sigma)\}$, με τους περιορισμούς ότι $\sigma \in \{0,1\}^n$ και $\sum_i \sigma_i = m$.

Για τη λύση αυτού του προβλήματος βελτιστοποίησης επιτρέπεται στο σ να ανήκει στο R^n κάτι που δίνει τη δυνατότητα παραγώγισης του $R^2(\sigma)W^2(\sigma)$ ως προς σ και εκτέλεσης αλγορίθμου gradient-descent. Περιορισμοί και όροι ποινής στη συνάρτηση κόστους χρησιμοποιούνται ώστε η μέθοδος να συγκλίνει σε μία λύση σ_{sol} τέτοια ώστε n-m συνιστώσες του σ_{sol} να έχουν τιμή κοντά στο μηδέν. Το m είναι μια παράμετρος που τίθεται από τον χρήστη και καθορίζει το μέγεθος που πρέπει να έχει το τελικό υποσύνολο. Τελικά επιλέγονται τα χαρακτηριστικά που αντιστοιχούν στις m μεγαλύτερες συνιστώσες του σ_{sol}.

ΚΕΦΑΛΑΙΟ 3. ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕ ΕΛΑΧΙΣΤΗ ΠΕΡΙΤΤΗ ΠΛΗΡΟΦΟΡΙΑ

3.1 Το κριτήριο της μέγιστης εξάρτησης
3.2 Ο αλγόριθμος mRMR
3.3 Ο αλγόριθμος mRMR για συνεχή χαρακτηριστικά
3.4 Σχετικοί αλγόριθμοι

3.1. Το κριτήριο της μέγιστης εξάρτησης

Χρησιμοποιώντας όρους από τη θεωρία πληροφορίας, το ιδανικό υποσύνολο χαρακτηριστικών είναι αυτό που έχει τη μέγιστη αμοιβαία πληροφορία με την κατηγορία, δηλαδή αυτό από το οποίο η κατηγορία έχει τη μεγαλύτερη εξάρτηση. Το υποσύνολο χαρακτηριστικών που θα έπρεπε να επιλεγεί είναι αυτό που μεγιστοποιεί την ποσότητα:

$$I(S;Y) = \sum_{s \in S} \sum_{y \in Y} p(S = s, Y = y) \log \frac{p(S = s, Y = y)}{p(S = s)\, p(Y = y)} \qquad \text{Εξ. 3.1}$$

όπου $S=\{X_1,\ldots,X_N\}$ ένα υποσύνολο χαρακτηριστικών, Y η κατηγορία, p(S) η από κοινού κατανομή των $X_1,\ldots,X_N$ και p(S,Y) η από κοινού κατανομή των $X_1,\ldots,X_N$,Y.

Δυστυχώς η αμοιβαία πληροφορία του συνόλου μεταβλητών S με τη μεταβλητή Y δεν μπορεί να εκτιμηθεί με αξιόπιστο τρόπο. Αυτό συμβαίνει γιατί ο αριθμός παραδειγμάτων που απαιτούνται για την εκτίμηση των από κοινού κατανομών p(s) και p(s,y) αυξάνεται εκθετικά ως προς των αριθμό των χαρακτηριστικών στο σύνολο S. Για παράδειγμα ας υποθέσουμε ότι ζητείται η εκτίμηση της κατανομής p(s) και ότι τα $X_1,\ldots,X_N$ είναι διακριτά χαρακτηριστικά με δυαδικό πεδίο ορισμού. Υπάρχουν 2^N διαφορετικές τιμές που μπορούν να ανατεθούν στο s, χρειάζεται επομένως ο

υπολογισμός 2^N πιθανοτήτων. Ο αριθμός αυτός θα είναι πολύ μεγαλύτερος από το πλήθος των διαθέσιμων παραδειγμάτων εκτός και αν το N είναι πολύ μικρό.

Εφόσον η αμοιβαία πληροφορία I(S;Y) δεν μπορεί να υπολογιστεί, κάποια προσέγγιση αυτής πρέπει να χρησιμοποιηθεί ώστε να βρεθεί και να επιλεγεί το υποσύνολο από το οποίο η κατηγορία έχει τη μεγαλύτερη εξάρτηση. Η univariate προσέγγιση στο πρόβλημα αυτό είναι να υπολογιστεί η εξάρτηση της κατηγορίας από κάθε χαρακτηριστικό ξεχωριστά και εν συνεχεία να επιλεγούν τα K χαρακτηριστικά από τα οποία υπάρχει μεγαλύτερη εξάρτηση. Δηλαδή, το πρόβλημα εύρεσης του υποσυνόλου S^* με μέγεθος K για το οποίο ισχύει

$$S^* = \arg\max_{S \subset F, |S|=K} \{I(S;Y)\} \qquad \text{Εξ. 3.2}$$

μετατρέπεται στο ευκολότερο πρόβλημα εύρεσης ενός S^* για το οποίο ισχύει

$$S^* = \arg\max_{S \subset F, |S|=K} \left\{ \sum_{X_i \in S} I(X_i;Y) \right\} \qquad \text{Εξ. 3.3}$$

Το βασικό πρόβλημα της univariate προσέγγισης είναι ότι επιλέγεται μεγάλος αριθμός περιττών χαρακτηριστικών. Αν ένα χαρακτηριστικό επιλέγεται γιατί έχει υψηλή αμοιβαία πληροφορία με την κατηγορία, τότε χαρακτηριστικά πολύ όμοια με αυτό θα έχουν επίσης υψηλή αμοιβαία πληροφορία με την κατηγορία και θα επιλεγούν. Όμως μια ομάδα πολύ όμοιων χαρακτηριστικών προσφέρει λίγη παραπάνω πληροφορία για την κατηγορία από αυτήν που προσφέρουν μερικά μόνο χαρακτηριστικά της ομάδας. Ακραίο παράδειγμα είναι η επιλογή δύο χαρακτηριστικών που έχουν ίδιες μεταξύ τους τιμές σε κάθε παράδειγμα του συνόλου εκπαίδευσης. Ίσως να είναι χρήσιμη η επιλογή ενός εκ των δύο αλλά η γνώση του δεύτερου δεν προσφέρει κανένα κέρδος σε πληροφορία. Ένα καλύτερο υποσύνολο χαρακτηριστικών μπορεί να προκύψει αν επιλέγονται χαρακτηριστικά που έχουν μεν μεγάλη συνάφεια με την κατηγορία, αλλά την ίδια στιγμή είναι μεταξύ τους όσο το δυνατόν ανόμοια. Σε αυτή την ιδέα βασίζεται η μέθοδος mRMR που παρουσιάζεται στην επόμενη ενότητα.

3.2. Ο αλγόριθμος mRMR

Η μέθοδος mRMR [10] θέτει δύο συνθήκες οι οποίες πρέπει να ικανοποιούνται από ένα υποσύνολο χαρακτηριστικών: i) τα χαρακτηριστικά του υποσυνόλου πρέπει να έχουν όσο το δυνατόν μεγαλύτερη συνάφεια με την κατηγορία, ii) τα χαρακτηριστικά του υποσυνόλου πρέπει να είναι όσο το δυνατόν ανόμοια μεταξύ τους.

Η συνάφεια ενός υποσυνόλου με την κατηγορία εκτιμάται με τον ίδιο τρόπο που χρησιμοποιείται από την univariate προσέγγιση. Αν S το υποσύνολο χαρακτηριστικών και Y η κατηγορία, τότε η συνάφεια του S με την Y ισούται με:

$$V(S) = \frac{1}{|S|} \sum_{Xi \in S} I(X_i; Y) \qquad \text{Εξ. 3.4}$$

Όσον αφορά στη δεύτερη συνθήκη, δεν υπάρχει κάποιος προφανής τρόπος για το πώς μπορεί να μετρηθεί ο βαθμός στον οποίο τα χαρακτηριστικά του υποσυνόλου S είναι μεταξύ τους όμοια. Ως ευρετικό χρησιμοποιείται η μέση τιμή της ομοιότητας μεταξύ όλων των πιθανών ζευγών από χαρακτηριστικά του S. Η ποσότητα αυτή θα αναφέρεται στο εξής ως *περιττή πληροφορία (redundancy)* του S και συμβολίζεται με W(S). Αν για τη μέτρηση της ομοιότητας μεταξύ δύο χαρακτηριστικών X_1 και X_2 χρησιμοποιηθεί η αμοιβαία πληροφορία $I(X_1,X_2)$, τότε:

$$W(S) = \frac{1}{|S|} \frac{1}{|S|-1} \sum_{i=1}^{|S|} \sum_{j=1, j \neq i}^{|S|} I(X_i; X_j) \qquad \text{Εξ. 3.5}$$

Στόχος είναι η εύρεση ενός υποσυνόλου S^* που ελαχιστοποιεί την περιττή πληροφορία W(S) και μεγιστοποιεί τη συνάφεια V(S). Κατά κανόνα η αύξηση της συνάφειας συνοδεύεται από αύξηση της περιττής πληροφορίας και έτσι δεν υπάρχει μοναδικό υποσύνολο που να υπερέχει έναντι των άλλων με βάση και τα δύο κριτήρια. Για να επιλέξει τελικά μία λύση, ο αλγόριθμος mRMR συνδυάζει τους δύο στόχους σε μία συνάρτηση αξιολόγησης. Το πρόβλημα βελτιστοποίησης επαναδιατυπώνεται και το σύνολο S^* που επιλέγεται είναι αυτό για το οποίο ισχύει:

$$S^* = \arg\max_S \{V(S) - W(S)\} \qquad \text{Εξ. 3.6}$$

Αντί της (3.6) θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε συνάρτηση αξιολόγησης επιβραβεύει την αύξηση της συνάφειας και παράλληλα τιμωρεί την αύξηση της

περιττής πληροφορίας. Οι συγγραφείς προτείνουν μία δεύτερη συνάρτηση αξιολόγησης που δίνει μεγαλύτερη έμφαση στη μείωση της περιττής πληροφορίας. Σύμφωνα με αυτήν, το βέλτιστο υποσύνολο S^* υπολογίζεται ως:

$$S^* = \arg\max_S \{Vs / Ws\} \qquad \text{Εξ. 3.7}$$

Η εύρεση της βέλτιστης λύσης για τις εξισώσεις (3.6) και (3.7) απαιτεί την αποτίμηση $O(N^{|S|})$ υποσυνόλων όπου N το πλήθος όλων των χαρακτηριστικών και |S| το μέγεθος του επιλεγμένου υποσυνόλου. Προκειμένου η διαδικασία να γίνει υπολογιστικά εφικτή, το υποσύνολο σχηματίζεται χρησιμοποιώντας μία forward μέθοδο άπληστης αναζήτησης. Αρχικά επιλέγεται το πιο συναφές χαρακτηριστικό (αυτό με τη μεγαλύτερη αμοιβαία πληροφορία με την κατηγορία). Στη συνέχεια, σε κάθε επανάληψη επιλέγεται ένα χαρακτηριστικό που έχει μεγάλη συνάφεια με την κατηγορία και ταυτόχρονα μικρή ομοιότητα με τα ήδη επιλεγμένα χαρακτηριστικά. Τροποποιώντας κατάλληλα την εξίσωση (3.6) παίρνουμε μία συνάρτηση που αντί να αξιολογεί υποσύνολα χαρακτηριστικών, αξιολογεί μεμονωμένα χαρακτηριστικά. Αν S είναι το σύνολο των έως τώρα επιλεγμένων χαρακτηριστικών και F το σύνολο όλων των χαρακτηριστικών, τότε επιλέγεται το χαρακτηριστικό X^*:

$$X^* = \arg\max_{X \in F-S} \left\{ I(X;Y) - \frac{1}{|S|} \sum_{Xi \in S} I(X_i;X) \right\} \qquad \text{Εξ. 3.8}$$

Τροποποιώντας αναλόγως την εξίσωση (3.7), το χαρακτηριστικό X^* που επιλέγεται σε κάθε επανάληψη είναι το:

$$X^* = \arg\max_{X \in F-S} \frac{I(X;Y)}{\frac{1}{|S|} \sum_{Xi \in S} I(X_i;X)} \qquad \text{Εξ. 3.9}$$

Η διαδικασία επιλογής χαρακτηριστικών συνοψίζεται από τον αλγόριθμο 3.1.

Αλγόριθμος 3.1 Ο Αλγόριθμος mRMR

Είσοδος: σύνολο χαρακτηριστικών F, διάνυσμα με τις κατηγορίες των παραδειγμάτων εκπαίδευσης Y, μέγεθος τελικού υποσυνόλου K

Έξοδος: υποσύνολο επιλεγμένων χαρακτηριστικών S

1) S = {}
2) $X^* = \arg\max_{X \in F} \{I(X;Y)\}$
3) $S = S \cup X^*$
4) Για k=2 έως K
5) $X^* = \arg\max_{X \in F-S} \left\{ I(X;Y) - \frac{1}{|S|} \sum_{Xi \in S} I(X_i;X) \right\}$
6) $S = S \cup X^*$
7) τέλος

3.3. Ο αλγόριθμος mRMR για συνεχή χαρακτηριστικά

Η αμοιβαία πληροφορία, που χρησιμοποιείται από τον αλγόριθμο mRMR για την εκτίμηση της συνάφειας και της περιττής πληροφορίας που περιέχεται σε ένα υποσύνολο, μπορεί εύκολα να υπολογιστεί χρησιμοποιώντας τη σχέση (3.1), όταν οι μεταβλητές είναι διακριτές. Όταν δύο μεταβλητές X, Y είναι συνεχείς τότε η αμοιβαία πληροφορία δίνεται από τη σχέση:

$$I(X;Y) = \int_{x \in X} \int_{y \in Y} p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, dydx \qquad \text{Εξ. 3.10}$$

Ο υπολογισμός της αμοιβαίας πληροφορίας μεταξύ δύο συνεχών μεταβλητών είναι δύσκολος γιατί απαιτεί την εκτίμηση πυκνότητας πιθανότητας για τις επί μέρους και από κοινού κατανομές και κατόπιν την ολοκλήρωση σύμφωνα με τον τύπο (3.10).

Η παραπάνω διαδικασία μπορεί να αποφευχθεί χρησιμοποιώντας κάποιους από τους συντελεστές συσχέτισης που περιγράφηκαν στην ενότητα 2.4.1 οι οποίοι μπορούν να χειριστούν συνεχή χαρακτηριστικά. Στην εργασία [10] προτείνεται η χρήση του F-test για τη μέτρηση της συσχέτισης μεταξύ χαρακτηριστικών και κατηγορίας. Το F-test κρίνεται πιο κατάλληλο σε σχέση με το κριτήριο του Fisher γιατί μπορεί να

χρησιμοποιηθεί για προβλήματα όπου υπάρχουν πολλές κατηγορίες. Όσον αφορά στην ομοιότητα μεταξύ δύο χαρακτηριστικών όμως, κάποιο άλλο κριτήριο πρέπει να χρησιμοποιηθεί γιατί το F-test προϋποθέτει ότι μία εκ δύο μεταβλητών είναι διακριτή. Στο [10] προτείνεται η χρήση του συντελεστή συσχέτισης του Pearson. Ο συντελεστής συσχέτισης του Pearson ορίζεται ως:

$$R(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\,\mathrm{var}(Y)}} \qquad \text{Εξ. 3.11}$$

όπου var(X) είναι η διακύμανση της μεταβλητής X και cov(X,Y) η συμμεταβλητότητα (covariance) των μεταβλητών X,Y. Η συμμεταβλητότητα των μεταβλητών X,Y από ένα δείγμα πλήθους m υπολογίζεται ως:

$$\mathrm{cov}(X,Y) = \sum_{k=1}^{m}(x_k - \bar{x})(y_k - \bar{y}) \qquad \text{Εξ. 3.12}$$

όπου $\bar{x}$ και $\bar{y}$ οι μέσες τιμές των X,Y. Η συμμεταβλητότητα είναι θετική αν υπάρχει τάση αύξησης της μεταβλητής Y όταν αυξάνεται η τιμή της X, αρνητική όταν υπάρχει τάση μείωσης της Y όταν αυξάνεται η X και τείνει στο μηδέν όταν δεν υπάρχει συσχέτιση μεταξύ των μεταβλητών.

Μια άλλη προσέγγιση στο χειρισμό συνεχών χαρακτηριστικών είναι η διακριτοποίηση, η μετατροπή δηλαδή των συνεχών τιμών σε διακριτές. Κατά την διαδικασία της διακριτοποίησης ενός συνεχούς χαρακτηριστικού πρέπει να καθοριστεί ο αριθμός των περιοχών στις οποίες θα χωριστεί το πεδίο ορισμού του, καθώς και τα όρια βάσει των οποίων οι συνεχείς τιμές μετατρέπονται σε διακριτές. Ο κατάλληλος τρόπος διακριτοποίησης εξαρτάται από το πρόβλημα και η διαδικασία θα πρέπει να λαμβάνει υπόψιν οποιαδήποτε γνώση υπάρχει σχετικά με αυτό ώστε οι περιοχές που θα προκύψουν να έχουν φυσική σημασία.

Στο [10], η επιλογή χαρακτηριστικών χρησιμοποιείται για προβλήματα βιοπληροφορικής με σκοπό τον εντοπισμό γονιδίων που σχετίζονται με την εκδήλωση διαφόρων νόσων. Τα παραδείγματα εκπαίδευσης είναι ιστοί από ασθενείς οι οποίοι περιγράφονται από το επίπεδο έκφρασης διαφόρων γονιδίων, συνήθως αρκετών χιλιάδων. Στο πρόβλημα αυτό, το επίπεδο έκφρασης κάθε γονιδίου είναι μία συνεχής μεταβλητή. Η διακριτοποίηση της γίνεται χωρίζοντας το εύρος τιμών της σε

τρεις περιοχές που αντιστοιχούν σε επίπεδο έκφρασης χαμηλό, μεσαίο και υψηλό. Τα όρια που χρησιμοποιούνται για τη διακριτοποίηση υπολογίζονται βάσει της μέσης τιμής της μεταβλητής, έστω μ, και της τυπικής απόκλισης, έστω σ (τα μ και σ υπολογίζονται από τις τιμές της μεταβλητής στα παραδείγματα εκπαίδευσης). Η τιμή -1 ανατίθεται σε ένα παράδειγμα αν το επίπεδο έκφρασης είναι μικρότερο από $\mu - \sigma/2$, η τιμή 1 ανατίθεται αν το επίπεδο έκφρασης είναι μεγαλύτερο από $\mu + \sigma/2$, διαφορετικά ανατίθεται η τιμή 0.

Η διακριτοποίηση με τον παραπάνω τρόπο γενικά ταιριάζει στα προβλήματα βιοπληροφορικής, μάλιστα οι συγγραφείς αναφέρουν βελτίωση της απόδοσης όταν στον ταξινομητή παρέχονται διακριτοποιημένα δεδομένα και όχι συνεχή. Σε ένα άλλο πρόβλημα η διακριτοποίηση σε μόνο τρεις περιοχές μπορεί να οδηγήσει στην απώλεια πληροφορίας (αν π.χ. τα δεδομένα δικαιολογούν την ύπαρξη 10 διαφορετικών διακριτών τιμών). Για αυτό το λόγο μπορούν να χρησιμοποιηθούν μέθοδοι διακριτοποίησης που καθορίζουν δυναμικά τον αριθμό των περιοχών. Μερικές τέτοιες μέθοδοι παρουσιάζονται συνοπτικά στα [11] και [19].

3.4. Σχετικοί αλγόριθμοι

Στη βιβλιογραφία εκτός από τον αλγόριθμο mRMR έχουν προταθεί μερικοί ακόμα αλγόριθμοι που διέπονται από την ίδια φιλοσοφία, επιχειρούν δηλαδή να βρουν ένα υποσύνολο συναφών με την κατηγορία χαρακτηριστικών που παράλληλα δεν έχουν μεγάλη συσχέτιση μεταξύ τους. Ο Battiti [4] χρησιμοποιεί επίσης την αμοιβαία πληροφορία για τη μέτρηση των συσχετίσεων σε μία forward selection μέθοδο. Η μέθοδος ξεκινά επιλέγοντας το πιο συναφές με την κατηγορία χαρακτηριστικό και στη συνέχεια σε κάθε επανάληψη επιλέγεται το χαρακτηριστικό X που μεγιστοποιεί την ποσότητα $I(X;Y) - \beta \sum_{Xi \in S} I(X_i;X)$, όπου S το σύνολο των ήδη επιλεγμένων χαρακτηριστικών, Y η κατηγορία και β μία σταθερά που ρυθμίζει τη σημασία που αποδίδεται στη μείωση της περιττής πληροφορίας. Η μόνη διαφορά σε σχέση με τον αλγόριθμο mRMR είναι ότι στη συνάρτηση αξιολόγησης η σταθερά β τίθεται εκ των προτέρων εμπειρικά, ενώ στον mRMR ισούται με 1/|S|, αλλάζει δηλαδή σε κάθε επανάληψη καθώς προστίθενται χαρακτηριστικά στο S. Η χρήση ενός σταθερού β,

οδηγεί στην αύξηση του όρου $\beta \sum_{Xi \in S} I(X_i; X)$ καθώς το μέγεθος του S αυξάνει. Έτσι δίνεται συνεχώς μεγαλύτερη βαρύτητα στην επιλογή χαρακτηριστικών με μικρή περιττή πληροφορία ακόμα και αν αυτά δεν έχουν μεγάλη συσχέτιση με την κατηγορία.

Μία άλλη μέθοδος που μπορεί να ενταχθεί στην ίδια κατηγορία, περιγράφεται στο [18]. Η συνάρτηση αξιολόγησης ενός υποσυνόλου S είναι:

$$J(S) = \frac{k \cdot \overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}} \qquad \text{Εξ. 3.13}$$

όπου $\overline{r_{cf}}$ η μέση τιμή της συσχέτισης χαρακτηριστικών του S και κατηγορίας, $\overline{r_{ff}}$ η μέση τιμή των συσχετίσεων των χαρακτηριστικών του S ανα δύο, και τέλος k το μέγεθος του συνόλου S. Ο αριθμητής εκφράζει τη συνάφεια χαρακτηριστικών-κατηγορίας ενώ ο παρονομαστής εκφράζει την περιττή πληροφορία στο σύνολο S. Χρησιμοποιείται η best first search τεχνική για τη βελτιστοποίηση του J, και το τελικό υποσύνολο δημιουργείται αυξητικά (forward μέθοδος).

Η μέθοδος CMIM που προτείνεται στο [13] χρησιμοποιεί επίσης όρους θεωρίας πληροφορίας για την επίτευξη των ίδιων στόχων, καταλήγει όμως σε μία ελαφρώς διαφορετική συνάρτηση αξιολόγησης. Τα χαρακτηριστικά επιλέγονται με μία forward άπληστη μέθοδο. Έστω ότι k forward βήματα έχουν ολοκληρωθεί και το σύνολο S έχει επιλεγεί ήδη. Σκοπός στο (k+1)-οστό βήμα είναι η επιλογή του χαρακτηριστικού X που θα ελαχιστοποιήσει την υπό συνθήκη εντροπία H(Y|S,X) που εκφράζει την αβεβαιότητα στην πρόβλεψη της κατηγορίας όταν το υποσύνολο χαρακτηριστικών $S \cup \{X\}$ είναι γνωστό. Ο στόχος αυτός μπορεί ισοδύναμα να εκφραστεί ως:

$$X^* = \arg\min_{X \in F-S} H(Y \mid S, X) \Leftrightarrow \qquad \text{Εξ. 3.14}$$

$$X^* = \arg\max_{X \in F-S} \{H(Y \mid S) - H(Y \mid S, X)\} \Leftrightarrow \qquad \text{Εξ. 3.15}$$

$$X^* = \arg\max_{X \in F-S} I(Y; X \mid S) \qquad \text{Εξ. 3.16}$$

Η ποσότητα I(Y;X|S) εκφράζει την πληροφορία που δίνει το X για την κατηγορία και δεν δίνεται ήδη από το S. Ισχύει ότι:

$$I(Y;X \mid S) \le \min_{X_i \in S} I(Y \mid X;X_i) \qquad \text{Εξ. 3.17}$$

Επειδή η ποσότητα I(Y;X|S) δεν μπορεί να εκτιμηθεί αξιόπιστα, η ποσότητα $\min_{X_i \in S} I(Y \mid X;X_i)$ χρησιμοποιείται για την αξιολόγηση των υποψήφιων χαρακτηριστικών. Αρχικά επιλέγεται το χαρακτηριστικό X_1 που έχει τη μεγαλύτερη αμοιβαία πληροφορία για την κατηγορία, και στη συνέχεια στον (k+1)-οστό γύρο το χαρακτηριστικό X_{k+1}

$$X_{k+1} = \max_{X \in F-S} \left\{ \min_{X_i \in S} I(Y;X \mid X_i) \right\} \qquad \text{Εξ. 3.18}$$

Η ποσότητα $\min_{X_i \in S} I(Y;X \mid X_i)$ είναι μικρή είτε όταν το X δεν περιέχει σημαντική πληροφορία για την Y, είτε όταν η πληροφορία που περιέχει δίνεται από κάποιο άλλο ήδη επιλεγμένο χαρακτηριστικό X_i.

ΚΕΦΑΛΑΙΟ 4. ΤΡΟΠΟΠΟΙΗΣΕΙΣ ΤΟΥ MRMR

4.1 Εκτίμηση της περιττής πληροφορίας
4.2 Βελτιστοποίηση του κριτηρίου αξιολόγησης
4.3 Κανονικοποιημένο mRMR

Στο κεφάλαιο αυτό μελετώνται οι αδυναμίες του αλγορίθμου mRMR και εξετάζονται τρόποι με τους οποίους αυτές μπορούν να διορθωθούν. Η ενότητα 4.1 επικεντρώνεται στην προσέγγιση που χρησιμοποιείται για την εκτίμηση της περιττής πληροφορίας που περιέχεται στα υποψήφια χαρακτηριστικά και προτείνονται εναλλακτικές προσεγγίσεις. Στην ενότητα 4.2 μελετάται κατά πόσο υπάρχει περιθώριο εύρεσης ενός καλύτερου υποσυνόλου χαρακτηριστικών σε σχέση με αυτό που βρίσκει ο mRMR, χρησιμοποιώντας άλλες μεθόδους αναζήτησης αντί για την άπληστη forward αναζήτηση. Τέλος, στην ενότητα 4.3 εξετάζεται η συνάρτηση αξιολόγησης που χρησιμοποιεί ο mRMR και προτείνεται μία τροποποίηση της.

4.1. Εκτίμηση της περιττής πληροφορίας

Οι αλγόριθμοι [4],[10] που παρουσιάστηκαν στο προηγούμενο κεφάλαιο, ακολουθούν ένα forward σχήμα αναζήτησης όπου ένα χαρακτηριστικό επιλέγεται σε κάθε βήμα και προστίθεται στο υποσύνολο των επιλεγμένων. Η χρησιμότητα ενός υποψήφιου χαρακτηριστικού αναλύεται βάσει δύο παραγόντων: της συνάφειάς του με την κατηγορία και της περιττής πληροφορίας του σε σχέση με τα ήδη επιλεγμένα χαρακτηριστικά. Για την εκτίμηση της περιττής πληροφορίας που περιέχει ένα χαρακτηριστικό X όταν έχει επιλεγεί το υποσύνολο S, η αμοιβαία πληροφορία I(S;X) θα ήταν το ιδανικό μέτρο. Καθώς όμως το I(S;X) δεν μπορεί να εκτιμηθεί αξιόπιστα, προσεγγίζεται από τον αλγόριθμο mRMR με τον τύπο

$$redundancy(X) = \frac{1}{|S|} \sum_{X_i \in S} I(X_i; X) \qquad \text{Εξ. 4.1}$$

μία προσέγγιση στην οποία θα αναφερόμαστε με το όνομα mean-redundancy.

Πρέπει να σημειωθεί ότι λίγα μόνο χαρακτηριστικά του συνόλου S είναι αρκετά για να καταστήσουν περιττή την πληροφορία που περιέχει το X. Ως ακραίο παράδειγμα μπορούμε να θεωρήσουμε την περίπτωση όπου ένα χαρακτηριστικό που ανήκει στο S, έχει ίδια τιμή με το υποψήφιο χαρακτηριστικό X για όλα τα παραδείγματα του συνόλου εκπαίδευσης. Το χαρακτηριστικό X δεν προσφέρει καμία επιπλέον πληροφορία και επομένως δεν πρέπει να επιλεγεί. Παρόλα αυτά, είναι πιθανό σύμφωνα με την προσέγγιση mean-redundancy να θεωρηθεί ότι το X περιέχει μικρή περιττή πληροφορία, αρκεί να υπάρχουν χαρακτηριστικά με τα οποία μοιάζει λίγο έτσι ώστε ο μέσος όρος της 4.1 να είναι χαμηλός. Σκοπός σε αυτή την ενότητα είναι η μελέτη εναλλακτικών τρόπων εκτίμησης της περιττής πληροφορίας που θα μπορούσαν να διορθώσουν το παραπάνω πρόβλημα.

Η διαίσθηση είναι ότι δεν έχει σημασία αν το υποψήφιο χαρακτηριστικό X μοιάζει λίγο ή καθόλου με κάποια χαρακτηριστικά του S, τη στιγμή που υπάρχει ένα χαρακτηριστικό με το οποίο μοιάζει πολύ. Έτσι, μία εναλλακτική προσέγγιση που ονομάζουμε max-redundancy, είναι η περιττή πληροφορία που περιέχει το X να υπολογίζεται μόνο με βάση το πιο όμοιο χαρακτηριστικό, δηλαδή

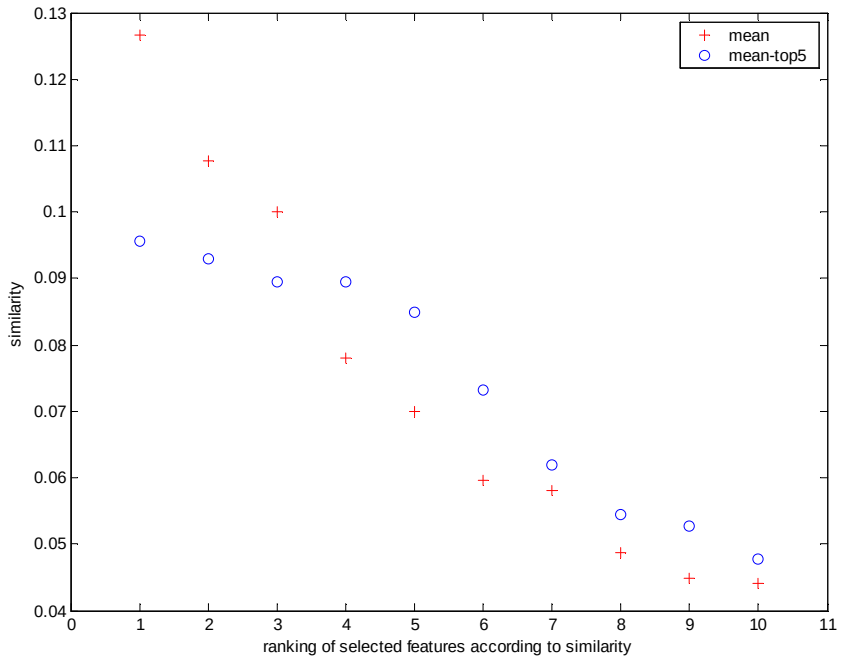
$$redundancy(X) = \max_{X_i \in S} I(X_i; X) \qquad \text{Εξ. 4.2}$$

Αν και το προηγούμενο σενάριο όπου δύο χαρακτηριστικά είναι τελείως όμοια, βοηθάει να γίνει αντιληπτή μία αδυναμία στη χρήση του μέσου όρου για τον υπολογισμό της περιττής πληροφορίας, δεν αναμένεται να συμβαίνει συχνά. Όταν το χαρακτηριστικό X δεν εξαρτάται απόλυτα από κανένα χαρακτηριστικό, τότε περισσότερα χαρακτηριστικά πρέπει να ληφθούν υπόψιν για να εκτιμηθεί η περιττή πληροφορία του X. Αυτό μπορεί να γίνει καλύτερα κατανοητό αν αναλύσουμε την ποσότητα I(S;X), που προσπαθούμε να προσεγγίσουμε, σε επιμέρους όρους. Έστω ότι $X_1,...,X_n$ τα χαρακτηριστικά που ανήκουν στο σύνολο S. Ισχύει ότι [9]:

$$I(S;X) = I(X;X_1) + I(X;X_2 \mid X_1) + ... + I(X;X_n \mid X_1, X_2, ..., X_{n-1}) \qquad \text{Εξ. 4.3}$$

Έστω X_1 το πιο όμοιο χαρακτηριστικό με το X. Αν τα X_1 και X ταυτίζονται, τότε $I(X_1;X) = H(X)$ οπότε $I(X;X_2 \mid X_1) + ... + I(X;X_n \mid X_1, X_2, ..., X_{n-1}) = 0$ γιατί $I(S;X) \le H(X)$, οπότε πράγματι τα χαρακτηριστικά $X_2,...,X_n$ δεν χρειάζεται να ληφθούν υπόψιν στον υπολογισμό της περιττής πληροφορίας. Όταν όμως $I(X_1;X) < H(X)$, τότε $I(X;X_2 \mid X_1) + ... + I(X;X_n \mid X_1, X_2, ..., X_{n-1}) > 0$. Μάλιστα όσο μικρότερος είναι ο όρος $I(X_1;X)$, τόσο σημαντικότεροι γίνονται οι υπόλοιποι όροι του αθροίσματος.

Μια ενδιάμεση προσέγγιση ανάμεσα στις mean-redundancy και max-redundancy είναι η περιττή πληροφορία του χαρακτηριστικού X να υπολογίζεται ως μέσος όρος των K πιο όμοιων χαρακτηριστικών, όπου K ένας αριθμός προκαθορισμένος εμπειρικά. Έστω mean-topK το όνομα της προσέγγισης αυτής.



Σχήμα 4.1 Ομοιότητες των λιγότερο περιττών χαρακτηριστικών ως προς τις προσεγγίσεις mean-redundancy και mean-top5 με τα ήδη επιλεγμένα χαρακτηριστικά.

Το σχήμα 4.1 δείχνει πως μπορεί να επηρεαστεί η επιλογή χαρακτηριστικών από την χρήση της προσέγγισης mean-topK. Έχοντας επιλέξει ήδη δέκα χαρακτηριστικά με

τον αλγόριθμο mRMR στο σύνολο δεδομένων της λευχαιμίας (παρουσιάζεται στο επόμενο κεφάλαιο), αναζητούμε το λιγότερο περιττό ανάμεσα στα εναπομείναντα χαρακτηριστικά με τις δύο διαφορετικές προσεγγίσεις: τις mean-redundancy και mean-topK με K=5. Έστω X_1 και X_2 τα λιγότερο περιττά χαρακτηριστικά σύμφωνα με την πρώτη και δεύτερη παραλλαγή αντίστοιχα. Στο σχήμα φαίνονται οι ομοιότητες των X_1, X_2 με τα ήδη επιλεγμένα χαρακτηριστικά. Για το X_1 υπάρχουν τρία χαρακτηριστικά με τα οποία η ομοιότητα είναι μεγαλύτερη από την ομοιότητα που έχει το X_2 με οποιοδήποτε χαρακτηριστικό. Παρόλα αυτά το X_1 θεωρείται καλύτερο από την mean-redundancy προσέγγιση γιατί υπάρχουν κάποια χαρακτηριστικά με τα οποία η ομοιότητα είναι πολύ μικρή. Η παραλλαγή mean-topK αντίθετα θεωρεί καλύτερο το χαρακτηριστικό X_2, όπως ήταν ο αρχικός στόχος.

Ένα πρόβλημα της παραλλαγής mean-topK είναι ότι η παράμετρος K θέτει ένα αυθαίρετο όριο που καθορίζει ποια χαρακτηριστικά λαμβάνονται υπόψιν και ποια αγνοούνται. Προκειμένου να γίνει λιγότερο απόλυτος ο διαχωρισμός αυτός, διατυπώνεται μια τελευταία παραλλαγή για τον υπολογισμό της περιττής πληροφορίας. Η ιδέα είναι να υπολογίζεται η περιττή πληροφορία ενός υποψηφίου χαρακτηριστικού ως σταθμισμένος μέσος όρος όλων των ομοιοτήτων. Η μικρή ομοιότητα που πιθανόν υπάρχει με κάποια χαρακτηριστικά, δεν θα πρέπει να μπορεί να ελαττώσει σημαντικά τον μέσο όρο. Για αυτό το λόγο τα βάρη πρέπει να έχουν μικρές τιμές για χαρακτηριστικά τα οποία μοιάζουν λίγο με το υποψήφιο. Ένα απλό σχήμα καθορισμού των βαρών που ικανοποιεί αυτή την απαίτηση είναι το εξής. Έστω ότι X το υποψήφιο χαρακτηριστικό και $X_1,...,X_n$ τα επιλεγμένα χαρακτηριστικά ταξινομημένα με βάση την ομοιότητα με το X σε φθίνουσα σειρά (το X_1 είναι το περισσότερο όμοιο με το X και το X_n το λιγότερο όμοιο). Η περιττή πληροφορία του X υπολογίζεται ως:

$$redundancy(\mathrm{X}) = \sum_{i=1}^{n} w_i I(X_i; X) \text{ με } w_i = \frac{(i+1)^{-1}}{\sum_{j=1}^{n} (j+1)^{-1}} \qquad \text{Εξ. 4.4}$$

Ονομάζουμε την προσέγγιση αυτή weighted-redundancy. Στο κεφάλαιο 5 ελέγχεται πειραματικά η απόδοση του αλγορίθμου mRMR όταν χρησιμοποιεί τις προσεγγίσεις max-redundancy και weighted-redundancy για τον υπολογισμό της ομοιότητας.

4.2. Βελτιστοποίηση του κριτηρίου αξιολόγησης

Η μέθοδος mRMR δημιουργεί αυξητικά το υποσύνολο επιλεγμένων χαρακτηριστικών εκτελώντας άπληστη αναζήτηση. Η άπληστη αναζήτηση έχει το πλεονέκτημα ότι είναι γρήγορη, συχνά όμως δεν δίνει ικανοποιητική λύση γιατί παγιδεύεται σε τοπικά ελάχιστα της συνάρτησης αξιολόγησης. Στην ενότητα αυτή εξετάζεται κατά πόσον είναι εφικτή η εύρεση κάποιου καλύτερου (ως προς τη συνάρτηση αξιολόγησης) υποσυνόλου από αυτό που επιλέγεται από τη μέθοδο mRMR, χρησιμοποιώντας άλλες μεθόδους αναζήτησης. Στις υποενότητες που ακολουθούν περιγράφεται η συνάρτηση αξιολόγησης που βελτιστοποιείται και οι μέθοδοι που χρησιμοποιούνται για τη βελτιστοποίηση. Τέλος ελέγχεται πειραματικά σε τι βαθμό μπορεί να βελτιωθεί η λύση που δίνει ο αλγόριθμος mRMR.

4.2.1. Καθορισμός του κριτηρίου αξιολόγησης

Για να χρησιμοποιηθεί κάποια από τις τεχνικές αναζήτησης που παρουσιάστηκαν στο κεφάλαιο 2, πρέπει καταρχήν να οριστεί μία συνάρτηση για την αξιολόγηση υποσυνόλων. Όπως αναφέρθηκε στο κεφάλαιο 3, μια συνάρτηση που μπορεί να χρησιμοποιηθεί είναι η εξής:

$$J(S) = V(S) - W(S) \qquad \text{Εξ. 4.5}$$

όπου S το υποσύνολο επιλεγμένων χαρακτηριστικών, V(S) η μέση τιμή της συνάφειας των χαρακτηριστικών του S με την κατηγορία με βάση την εξίσωση (3.4), και W(S) η μέση τιμής της ομοιότητας των χαρακτηριστικών του S μεταξύ τους όπως ορίστηκε στην εξίσωση (3.5). Ισοδύναμη με τη μεγιστοποίηση του J(S) είναι η μεγιστοποίηση του J_1(S):

$$J_1(S) = \sum_{i=1}^{|S|} I(X_i;Y) - \frac{1}{|S|-1}\sum_{i=1}^{|S|}\left\{\sum_{j=1, j\neq i}^{|S|} I(X_i;X_j)\right\} \qquad \text{Εξ. 4.6}$$

Ας υποθέσουμε ότι το κριτήριο J_1 χρησιμοποιείται για την επιλογή χαρακτηριστικών σε συνδυασμό με μία μέθοδο που εκτελεί forward άπληστη αναζήτηση και ότι ένα χαρακτηριστικό επιλέγεται σε κάθε βήμα. Ας υποθέσουμε επίσης ότι το σύνολο $S=\{X_1,...,X_N\}$ έχει ήδη επιλεγεί και ότι ζητείται η επέκταση του S με την προσθήκη

ενός χαρακτηριστικού που ανήκει στο σύνολο των εναπομεινάντων χαρακτηριστικών F-S. Το καλύτερο χαρακτηριστικό σύμφωνα με το κριτήριο J_1 είναι το X^*:

$$X^* = \arg\max_{X \in F-S} J_1(S \cup X) \Leftrightarrow \quad \text{Εξ. 4.7}$$

$$X^* = \arg\max_{X \in F-S} \left\{ \sum_{i=1}^{|S|} I(X_i;Y) + I(X;Y) - \frac{1}{|S|} \left(\sum_{i=1}^{|S|} \sum_{j=1, j \neq i}^{|S|} I(X_i;X_j) + 2 \cdot \sum_{j=1}^{|S|} I(X;X_j) \right) \right\} \quad \text{Εξ. 4.8}$$

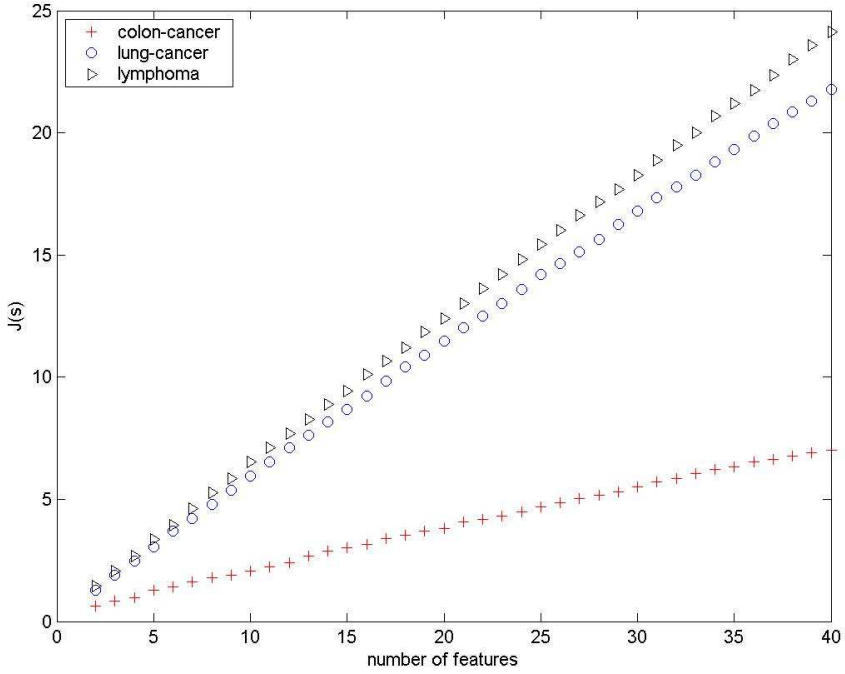
Στην εξίσωση 4.8 το άθροισμα των ομοιοτήτων του υποψήφιου χαρακτηριστικού X με τα ήδη επιλεγμένα πολλαπλασιάζεται επί δύο γιατί σύμφωνα με την εξίσωση (4.6) η ομοιότητα μεταξύ δύο οποιωνδήποτε χαρακτηριστικών λαμβάνεται υπόψιν δύο φορές. Από την εξίσωση (4.8) παίρνουμε ότι:

$$X^* = \arg\max_{X \in F-S} \left\{ I(X;Y) - \frac{2}{|S|} \sum_{j=1}^{|S|} I(X;X_j) \right\} \quad \text{Εξ. 4.9}$$

Από την τελευταία εξίσωση βλέπουμε η μεγιστοποίηση του J_1 δεν είναι ισοδύναμη με το κριτήριο που χρησιμοποιεί ο αλγόριθμος mRMR, καθώς η μεγιστοποίηση του J_1 δίνει διπλάσιο βάρος στον δεύτερο όρο που αντιστοιχεί στη μείωση της περιττής πληροφορίας. Το κριτήριο J_2 του οποίου η μεγιστοποίηση είναι ισοδύναμη με το κριτήριο που χρησιμοποιεί ο mRMR είναι:

$$J_2(S) = \sum_{i=1}^{|S|} I(X_i;Y) - \frac{1}{2} \sum_{i=1}^{|S|} \left\{ \frac{1}{|S|-1} \sum_{j=1, j \neq i}^{|S|} I(X_i;X_j) \right\} \quad \text{Εξ. 4.10}$$

Σκοπός σε αυτή την ενότητα είναι να μελετήσουμε πόσο μπορεί να βελτιωθεί η λύση που δίνει η forward άπληστη αναζήτηση που χρησιμοποιεί ο mRMR, με χρήση άλλων μεθόδων. Για να είναι αντικειμενική η σύγκριση, πρέπει να χρησιμοποιηθεί για την αξιολόγηση των υποσυνόλων το κριτήριο J_2 γιατί αυτό είναι που επιχειρεί να βελτιστοποιήσει ο mRMR.



Σχήμα 4.2 Η μεταβολή του κριτηρίου αξιολόγησης J_2 σε σχέση με τον αριθμό των χαρακτηριστικών που υπάρχουν στο υποσύνολο

Η βελτιστοποίηση του J_2 έχει νόημα υπό την προϋπόθεση ότι το μέγεθος των υποψηφίων υποσυνόλων είναι σταθερό. Αυτό το συμπέρασμα βγαίνει εξετάζοντας το σχήμα 4.2 όπου φαίνεται η τιμή του J_2 για τα *εμφωλευμένα υποσύνολα (nested subsets)* που παράγει ο αλγόριθμος mRMR για τρία διαφορετικά σύνολα δεδομένων. Έστω $S_{MRMR}=\{X_1,\ldots,X_k\}$ το υποσύνολο μεγέθους k που επιλέγεται από τον mRMR και S_i το εμφωλευμένο υποσύνολο μεγέθους i, δηλαδή $S_i=\{X_1,\ldots,X_i\}$. Παρατηρείται ότι $J_2(S_i)>J_2(S_j)$ αν i>j, δηλαδή η τιμή του J_2 έχει την τάση να αυξάνεται καθώς αυξάνει ο αριθμός των επιλεγμένων χαρακτηριστικών. Επομένως υποσύνολα διαφορετικού μεγέθους δεν είναι συγκρίσιμα.

4.2.2. Εναλλακτικές τεχνικές αναζήτησης

Έχοντας καθορίσει το κριτήριο αξιολόγησης που θα χρησιμοποιηθεί, εξετάζουμε για ποιο λόγο θα μπορούσε η forward-άπληστη αναζήτηση να μη δίνει καλή λύση. Το βασικό πρόβλημα είναι ότι μπορεί ένα χαρακτηριστικό που επιλέχτηκε σε κάποιο

πρώιμο στάδιο της αναζήτησης να μην συνδυάζεται τόσο καλά με χαρακτηριστικά που επιλέχτηκαν μετέπειτα. Με βάση το κριτήριο J_2, αυτό θα συνέβαινε αν το χαρακτηριστικό που επελέγη στο βήμα k, έστω X, είχε μεγάλη ομοιότητα με τα χαρακτηριστικά που επιλέχτηκαν στα επόμενα βήματα, έστω $X_{k+1,...,} X_n$. Σε αυτή την περίπτωση ένα καλύτερο υποσύνολο θα μπορούσε να προκύψει αντικαθιστώντας το X με κάποιο άλλο χαρακτηριστικό X' που έχει μικρότερη ομοιότητα με τα $X_{k+1,...,} X_n$.

Η υπόθεση που κάνουμε είναι ότι η αντιμετώπιση του παραπάνω προβλήματος μπορεί να οδηγήσει στην εύρεση καλύτερων υποσυνόλων. Το πρόβλημα μπορεί να αντιμετωπιστεί με την εκ των υστέρων εξέταση κάθε επιλεγμένου χαρακτηριστικού και την αντικατάσταση του με άλλο καλύτερο αν υπάρχει τέτοιο. Αν S_{MRMR} το υποσύνολο χαρακτηριστικών που επιλέχτηκε από τον mRMR και F το σύνολο όλων των χαρακτηριστικών, τότε κάθε χαρακτηριστικό που ανήκει στο S_{MRMR} μπορεί να διαγραφεί και στη θέση του να τοποθετηθεί ένα χαρακτηριστικό που ανήκει στο F-S_{MRMR}. Υπάρχουν $|S_{MRMR}|x|F-S_{MRMR}|$ τέτοιες πιθανές αντικαταστάσεις και κάθε μία δίνει ένα διαφορετικό υποψήφιο υποσύνολο. Κάθε υποψήφιο υποσύνολο αξιολογείται με βάση το κριτήριο J_2. Αν το καλύτερο υποσύνολο που προκύπτει από αυτή τη διαδικασία, έστω S', είναι καλύτερο από το αρχικό, τότε η διαδικασία επαναλαμβάνεται δοκιμάζοντας αντικαταστάσεις στο υποσύνολο S'. Αν δεν προκύψει κανένα βελτιωμένο υποσύνολο η διαδικασία τερματίζει. Η διαδικασία περιγράφεται από τον ψευδοκώδικα 4.1.

Η παραπάνω διαδικασία είναι επίσης μία άπληστη αναζήτηση γιατί αφού εξετάσει όλες τις πιθανές αντικαταστάσεις επιλέγει αυτήν που έδωσε το καλύτερο αποτέλεσμα. Αντί της άπληστης αναζήτησης, θα μπορούσε να χρησιμοποιηθεί η διαδικασία της προσομοιούμενης ανόπτησης (simulated annealing). Η κύρια διαφορά ανάμεσα σε προσομοιούμενη ανόπτηση και άπληστη αναζήτηση, είναι ότι η μετάβαση προς μία κατάσταση μπορεί να γίνει ακόμα και αν αυτή είναι χειρότερη ως προς το κριτήριο αξιολόγησης με κάποια πιθανότητα. Έτσι δίνεται η δυνατότητα στην αναζήτηση να ξεφύγει από τοπικά μέγιστα και να βρει μία καλύτερη λύση.

Αλγόριθμος 4.1 Βελτιστοποίηση του κριτηρίου J_2 με άπληστη αναζήτηση

Είσοδος: το αρχικό υποσύνολο επιλεγμένων χαρακτηριστικών S, το σύνολο όλων των χαρακτηριστικών F.

Έξοδος: το καλύτερο υποσύνολο που βρέθηκε S_best

1) S2 = S
2) Επανέλαβε
3) Για κάθε X που ανήκει στο S
4) Για κάθε X' που ανήκει στο F-S
5) S' = S-X+X'
6) Αν $J_2(S') > J_2(S2)$
7) S2 = S'
8) Αν S2≠S
9) S=S2
10) αλλιώς
11) Επίστρεψε το S

4.2.3. Σύγκριση με τη λύση του mRMR

Στην ενότητα αυτή εξετάζεται η βελτίωση που μπορεί να επιφέρει η εφαρμογή του αλγορίθμου διαδοχικών αντικαταστάσεων (αλγόριθμος 4.1) στα υποσύνολα που επιλέγονται από τον mRMR. Η σύγκριση έγινε στα εξής 4 σύνολα δεδομένων που περιγράφονται αναλυτικότερα στο 4^{ο} κεφάλαιο: Colon-cancer, Lymphoma, Lung-cancer και Leukemia. Όπως ήδη αναφέρθηκε, υποσύνολα διαφορετικού μεγέθους δεν είναι συγκρίσιμα για αυτό επιλέχθηκε να γίνει σύγκριση υποσυνόλων με 40 χαρακτηριστικά.

Για τα σύνολα δεδομένων colon-cancer και lung-cancer καμία αντικατάσταση δεν βρέθηκε που να δίνει καλύτερη απόδοση ως προς το κριτήριο J_2. Για τα σύνολα δεδομένων lymphoma και leukemia υπήρξε βελτίωση που επιτεύχθηκε με τρεις και δύο αντικαταστάσεις αντίστοιχα. Στην 2^{η} στήλη του πίνακα 4.1 παρουσιάζεται η αξιολόγηση ως προς το J_2 του υποσυνόλου που επιλέχθηκε από τον mRMR. Στην 3^{η} στήλη παρουσιάζεται η αξιολόγηση του υποσυνόλου που βρέθηκε μετά από διαδοχικές αντικαταστάσεις.

Πίνακας 4.1 Αξιολόγηση λύσεων ως προς το κριτήριο $J_2=V(S)-\frac{1}{2}W(S)$

	Λύση mRMR	Αρχικοποίηση με mRMR + διαδοχικές αντικαταστάσεις	10 τυχαίες αρχικές θέσεις + διαδοχικές αντικαταστάσεις	Αξιολόγηση αρχικών καταστάσεων ως προς J_2 (μ.ο.)
Colon-cancer	7.460618	7.460618	7.460618	1.9665
Lymphoma	24.638070	24.642288	24.642288	18.5337
Lung-cancer	22.175331	22.175331	22.175331	15.8781
Leukemia	11.920799	11.925338	11.925338	4.684811

Επειδή η μέθοδος αναζήτησης είναι άπληστη, η λύση που θα βρεθεί εξαρτάται από την αρχική κατάσταση. Είναι πιθανό να μπορούν να βρεθούν καλύτερες λύσεις αν η αναζήτηση ξεκινήσει από την κατάλληλη αρχική κατάσταση. Για να διαπιστωθεί αν κάτι τέτοιο όντως ισχύει, η διαδικασία των διαδοχικών αντικαταστάσεων επαναλήφθηκε 10 φορές ξεκινώντας όμως από τυχαίες αρχικές θέσεις (τυχαία υποσύνολα από 40 χαρακτηριστικά). Και τις 10 φορές, οι διαδοχικές αντικαταστάσεις οδήγησαν τελικά στο ίδιο υποσύνολο παρά το γεγονός ότι ξεκινούσαν από διαφορετικές αρχικές καταστάσεις. Η αξιολόγηση αυτού του υποσυνόλου ως προς το κριτήριο J_2 παρουσιάζεται στην $4^{η}$ στήλη του πίνακα 4.1. Το φαινόμενο παρατηρήθηκε και στα 4 σύνολα δεδομένων. Στην $5^{η}$ στήλη του 4.1 παρουσιάζεται ποια είναι κατά μέσο όρο η αξιολόγηση των 10 τυχαίων αρχικών καταστάσεων με βάση το J_2.

Το γεγονός ότι η άπληστη αναζήτηση καταλήγει πάντοτε στην ίδια λύση φαίνεται παράδοξο παρόλα αυτά μπορεί να εξηγηθεί. Η συνάρτηση που μεγιστοποιείται έχει τη γενική μορφή:

$$J(S)=V(S)-\beta\cdot W(S) \qquad \text{Εξ. 4.11}$$

όπου V η συνάφεια και W η περιττή πληροφορία που περιέχεται στο S. Για αρκετά μικρό β η επίδραση του όρου W(S) είναι μικρή και υπάρχει η τάση να επιλέγονται τα πιο συναφή χαρακτηριστικά. Όταν το β τείνει στο μηδέν, η συνάρτηση έχει ολικό μέγιστο που αποτελείται από τα πιο συναφή χαρακτηριστικά. Μάλιστα στη λύση αυτή μπορεί να φτάσει η αναζήτηση μέσω διαδοχικών αντικαταστάσεων ξεκινώντας από οποιαδήποτε αρχική κατάσταση, αντικαθιστώντας πάντα το λιγότερο συναφές

από τα επιλεγμένα χαρακτηριστικά με το περισσότερο συναφές από τα μη επιλεγμένα. Στην περίπτωση του κριτηρίου J_2, έχουμε β=1/2 και στην πράξη αποδεικνύεται ότι η τιμή αυτή του β είναι αρκετά μικρή ώστε να ισχύει το παραπάνω σενάριο.

Η κατάσταση δεν είναι ίδια όταν χρησιμοποιείται μια συνάρτηση αξιολόγησης που δίνει μεγαλύτερο βάρος στην περιττή πληροφορία. Χρησιμοποιώντας ως συνάρτηση αξιολόγησης το J_1 (εξίσωση (4.6)), που δίνει διπλάσιο βάρος στην περιττή πληροφορία σε σχέση με το J_2, παίρνουμε τα αποτελέσματα του πίνακα 4.2. Ξεκινώντας από 20 τυχαίες αρχικές θέσεις και εφαρμόζοντας διαδοχικές αντικαταστάσεις, βρέθηκαν αρκετές διαφορετικές λύσεις σε κάθε σύνολο δεδομένων. Η χειρότερη λύση που βρέθηκε παρουσιάζεται στην $4^{η}$ στήλη του πίνακα 4.2. Στην $5^{η}$ στήλη παρουσιάζεται η καλύτερη λύση που βρέθηκε ενώ στην $6^{η}$ στήλη φαίνεται ο αριθμός των διαφορετικών λύσεων που βρέθηκαν από τις 20 διαφορετικές αναζητήσεις. Η αξιολόγηση του υποσυνόλου που επιλέχθηκε με forward αναζήτηση (ξεκινώντας από το πιο συναφές και κατόπιν προσθέτοντας το χαρακτηριστικό που μεγιστοποιεί το κριτήριο της εξίσωσης (4.9)) φαίνεται στη $2^{η}$ στήλη. Εκτελώντας διαδοχικές αντικαταστάσεις στο υποσύνολο αυτό, βρέθηκε καλύτερο υποσύνολο σε 3 από τα 4 σύνολα δεδομένων ($3^{η}$ στήλη του πίνακα 4.2).

Χρησιμοποιώντας ως συνάρτηση αξιολόγησης την 4.11 και θέτοντας β=5, το φαινόμενο είναι ακόμα πιο έντονο. Οι 20 αναζητήσεις από τυχαίες αρχικές θέσεις κατέληξαν σε 20 διαφορετικές λύσεις σε 3 από τα 4 σύνολα δεδομένων εκτός από το σύνολο lung-cancer όπου βρέθηκαν 18 διαφορετικές λύσεις. Τα αποτελέσματα παρουσιάζονται στον πίνακα 4.3.

Από τα παραπάνω ευρήματα βγαίνει το συμπέρασμα ότι η χρήση μεθόδων καθολικής βελτιστοποίησης για τη βελτιστοποίηση της συνάρτησης αξιολόγησης J_2 δεν έχει ιδιαίτερο νόημα, καθώς αυτή δεν έχει πολλά τοπικά μέγιστα και έτσι η forward-άπληστη αναζήτηση καταφέρνει από μόνη της να βρει αρκετά καλές λύσεις. Αντίθετα, για συναρτήσεις αξιολόγησης που δίνουν μεγαλύτερη έμφαση στην ελαχιστοποίηση της περιττής πληροφορίας, πράγματι υπάρχει περιθώριο βελτίωσης από τη χρήση μεθόδων καθολικής βελτιστοποίησης.

Πίνακας 4.2 Αξιολόγηση λύσεων ως προς το κριτήριο J_1=V(S)-W(S)

	Λύση forward αναζήτησης	Αρχικοποίηση με forward + διαδοχικές αντικατα-στάσεις	20 τυχαίες αρχικές θέσεις+διαδοχικές αντικαταστάσεις (χειρότερη λύση)	20 τυχαίες αρχικές θέσεις+διαδοχικές αντικαταστάσεις (καλύτερη λύση)	Πλήθος διαφορε-τικών λύσεων	Αξιολόγηση αρχικών καταστάσεων ως προς J_1 (μ.ο.)
Colon	4.403746	4.403746	4.398164	4.409577	8	0.077696
Lymphoma	21.362736	21.383785	21.378602	21.389693	8	16.188258
Lung	19.201697	19.213106	19.213106	19.217707	2	13.650817
Leukemia	7.419353	7.492386	7.482838	7.501392	16	2.762569

Πίνακας 4.3 Αξιολόγηση λύσεων ως προς το κριτήριο J_3=V(S)-5W(S)

	Λύση forward αναζήτησης	Αρχικοποίηση με forward + διαδοχικές αντικατα-στάσεις	20 τυχαίες αρχικές θέσεις+ διαδοχικές αντικαταστάσεις (χειρότερη λύση)	20 τυχαίες αρχικές θέσεις+ διαδοχικές αντικαταστάσεις (καλύτερη λύση)	Πλήθος διαφορε-τικών λύσεων	Αξιολόγηση αρχικών καταστάσεων ως προς J_3 (μ.ο.)
Colon	-6.618445	-6.315082	-6.557311	-6.248328	20	-14.911825
Lymphoma	8.290338	8.547330	8.408317	8.712725	20	-4.781448
Lung	4.296803	5.096365	4.936465	5.126387	18	-6.644378
Leukemia	-4.341086	-3.979523	-4.022178	-3.798025	20	-13.880850

4.3. Κανονικοποιημένο mRMR

Ο αλγόριθμος mRMR έχει ως στόχο τη βελτιστοποίηση δύο συνθηκών. Στο πρόβλημα αυτό δεν υπάρχει μοναδική λύση αλλά ένα σύνολο λύσεων. Προκειμένου να επιλεγεί τελικά μία από αυτές, ο αλγόριθμος mRMR συνδυάζει τις δύο συνθήκες σε μία και τελικός στόχος είναι η μεγιστοποίηση του $J(S)=V(S)-W(S)$, όπου η μεταβλητή V προσεγγίζει τη συνάφεια του συνόλου S και η μεταβλητή W την περιττή πληροφορία που περιέχεται σε αυτό. Με αυτή την προσέγγιση ανταμείβεται το ίδιο μια αύξηση της συνάφειας κατά μία ποσότητα με μια μείωση της περιττής πληροφορίας κατά την ίδια ποσότητα. Το πρόβλημα είναι ότι η μεταβολή κατά αυτήν την ποσότητα μπορεί να έχει πολύ μεγάλη σημασία όταν αφορά στην περιττή

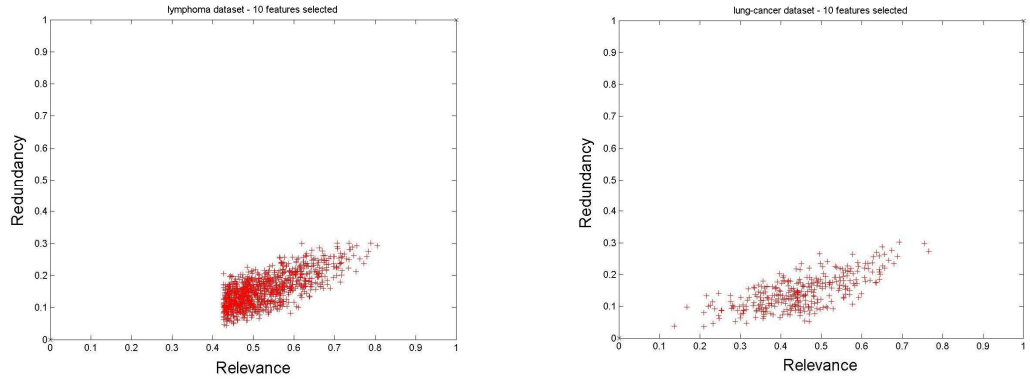
πληροφορία και πολύ μικρή σημασία όταν αφορά στη συνάφεια (ή και το αντίστροφο).

Το πόσο μεγάλη είναι η σημασία μεταβολής μιας μεταβλητής κατά μία ποσότητα μπορεί να εκτιμηθεί αν ληφθεί υπόψιν το εύρος τιμών που παίρνει η μεταβλητή. Για παράδειγμα ας υποθέσουμε ότι X_1 και X_2 είναι δύο υποψήφια προς επιλογή χαρακτηριστικά με $V(X_1)=100$ και $W(X_1)=1$, και $V(X_2)=98$ και $W(X_2)=0.1$. Με βάση το κριτήριο αξιολόγησης του αλγορίθμου mRMR έχουμε ότι: $J(X_1)=V(X_1)-W(X_1)=100-1=99$ και $J(X_2)=V(X_2)-W(X_2)=98-0.1=97.9$ οπότε το X_1 θεωρείται καλύτερο.

Αν εκτός από τα παραπάνω γνωρίζουμε ότι η συνάφεια των υποψηφίων χαρακτηριστικών παίρνει τιμές από 0 έως 100, ενώ η περιττή πληροφορία παίρνει τιμές από 0 έως 1, τότε η κατάσταση είναι τελείως διαφορετική. Το X_1 δεν μπορεί να θεωρηθεί καλό χαρακτηριστικό γιατί αν και έχει τη μέγιστη δυνατή συνάφεια, έχει επίσης τη μέγιστη περιττή πληροφορία. Το χαρακτηριστικό X_2 έχει αρκετά μεγάλη συνάφεια, ενώ η περιττή πληροφορία που περιέχει είναι κατά 0.9 μικρότερη από την περιττή πληροφορία του X_1, διαφορά που είναι τεράστια λαμβάνοντας υπόψιν τις τιμές που παίρνει το συγκεκριμένο μέγεθος.

Από το παραπάνω παράδειγμα είναι εμφανές ότι το μέγεθος που παρουσιάζει μεγαλύτερη διακύμανση, αποκτά μεγαλύτερη σημασία. Μια ασήμαντη μεταβολή του μπορεί να είναι κατά απόλυτη τιμή μεγαλύτερη από μία σημαντική μεταβολή του άλλου μεγέθους. Στο παράδειγμα η ασήμαντη διαφορά της συνάφειας θεωρήθηκε σπουδαιότερη από τη διαφορά της περιττής πληροφορίας.

Από τα παραπάνω γίνεται κατανοητό ότι η απευθείας σύγκριση συνάφειας και περιττής πληροφορίας δεν είναι δίκαιη όταν οι τιμές των δύο μεγεθών παρουσιάζουν διαφορετική διακύμανση. Είναι πιθανό τα δύο μεγέθη να παίρνουν περίπου τις ίδιες τιμές, πολλές φορές όμως αυτό δεν συμβαίνει όπως φαίνεται στο σχήμα 4.3 όπου παρουσιάζεται για δύο διαφορετικά σύνολα δεδομένων η συνάφεια και η περιττή πληροφορία των υποψηφίων χαρακτηριστικών στο δέκατο βήμα της προς τα εμπρός αναζήτησης.



a) Σύνολο δεδομένων Lymphoma b) Σύνολο δεδομένων Lung-cancer

Σχήμα 4.3 Συνάφεια και περιττή πληροφορία υποψήφιων χαρακτηριστικών σε δύο διαφορετικά σύνολα δεδομένων.

Για την αντιμετώπιση του παραπάνω προβλήματος προτείνεται η κανονικοποίηση των δύο μεγεθών, συνάφειας και περιττής πληροφορίας, έτσι ώστε να έχουν την ίδια διακύμανση. Ο αλγόριθμος mRMR τροποποιείται ως εξής: σε κάθε επανάληψη υπολογίζεται για κάθε μη επιλεγμένο χαρακτηριστικό, έστω X_i, η συνάφεια του με την κατηγορία, έστω $V(X_i)$, και η περιττή πληροφορία ως προς το σύνολο των ήδη επιλεγμένων χαρακτηριστικών, έστω $W(X_i)$. Με βάση τις τιμές αυτές, υπολογίζεται η μέση τιμή $\bar{V}$ και η τυπική απόκλιση σ_V της συνάφειας:

$$\bar{V} = \frac{1}{|F-S|}\sum_{X_i \in F-S} V(X_i) \qquad \text{Εξ. 4.12}$$

$$\sigma_V = \sqrt{\frac{1}{|F-S|}\sum_{X_i \in F-S} (V(X_i)-\bar{V})^2} \qquad \text{Εξ. 4.13}$$

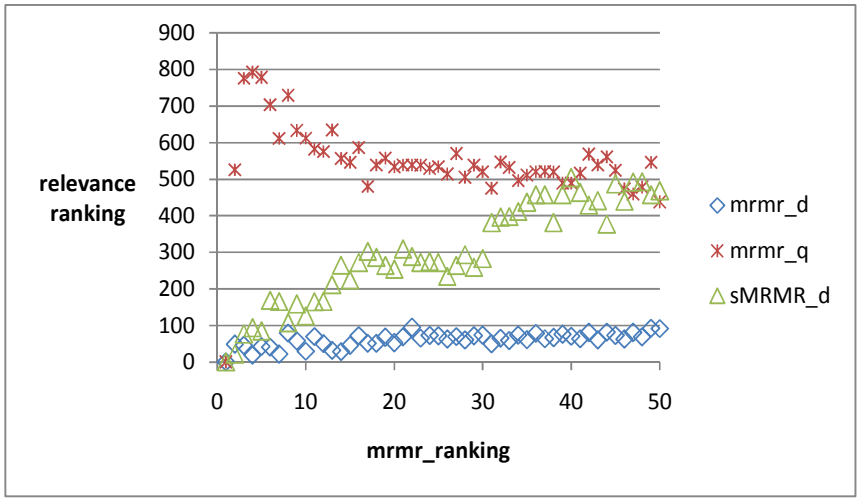
όπου F το σύνολο όλων των χαρακτηριστικών και S το σύνολο των επιλεγμένων χαρακτηριστικών. Με τον ίδιο τρόπο υπολογίζεται η μέση τιμή και η τυπική απόκλιση της περιττής πληροφορίας που περιέχεται από τα υποψήφια χαρακτηριστικά, έστω $\bar{W}$ και σ_W αντίστοιχα. Στη συνέχεια για κάθε υποψήφιο χαρακτηριστικό X_i, υπολογίζεται η κανονικοποιημένη συνάφεια $V'(X_i)$ και περιττή πληροφορία $W'(X_i)$:

$$V'(X_i) = \frac{V(X_i) - \overline{V}}{\sigma_V} \text{ και } W'(X_i) = \frac{W(X_i) - \overline{W}}{\sigma_W}$$

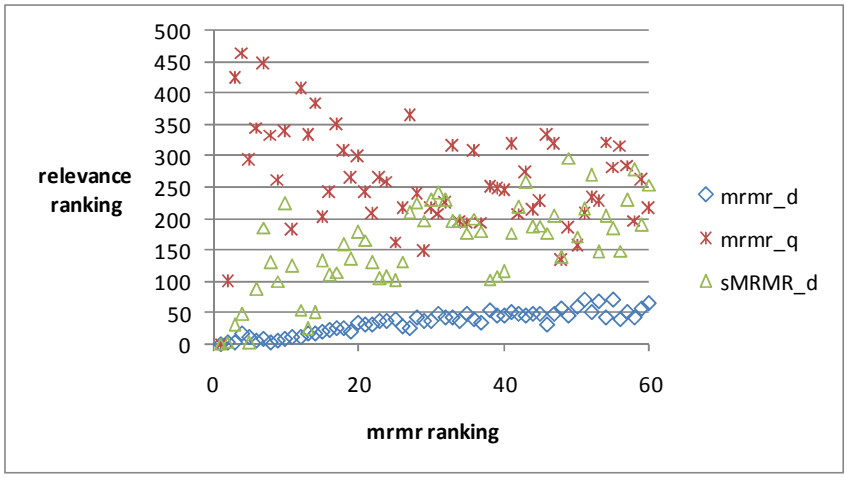
Με αυτήν την κανονικοποίηση, και τα δύο μεγέθη έχουν μέση τιμή μηδέν και τυπική απόκλιση ένα. Τελικά επιλέγεται το χαρακτηριστικό X^* που μεγιστοποιεί τη διαφορά:

$$X^* = \arg\max_{X_i \in F-S} \left[V'(X_i) - W'(X_i) \right] \qquad \text{Εξ. 4.14}$$

Στη συνέχεια εξετάζουμε τη διαφορά που προκύπτει στη συμπεριφορά του mRMR από αυτή την παραλλαγή. Χρησιμοποιούμε τα αρχικά mrmr_d (mrmr-difference) για να αναφερθούμε στη μέθοδο mRMR που μεγιστοποιεί τη διαφορά συνάφειας και περιττής πληροφορίας (V(S)-W(S)), τα αρχικά mrmr_q (mrmr-quotient) αντιστοιχούν στη μέθοδο βελτιστοποίησης του λόγου των δύο μεγεθών (V(S)/W(S)), ενώ για την κανονικοποιημένη παραλλαγή του αλγορίθμου χρησιμοποιούνται τα αρχικά sMRMR_d (scaled-mrmr-difference). Όπως φαίνεται και στο σχήμα 4.3 το μέγεθος της συνάφειας έχει μεγαλύτερη διακύμανση οπότε αναμένεται η μέθοδος mrmr_d να προτιμά συναφή χαρακτηριστικά ακόμα και αν αυτά περιέχουν σημαντική περιττή πληροφορία. Η mrmr_q καταφέρνει να δώσει περισσότερη έμφαση στη μείωση της περιττής πληροφορίας, αν και λόγω της διαίρεσης δεν είναι εύκολα κατανοητό πόσο βάρος δίνεται σε κάθε μέγεθος και πώς αυτό το βάρος επηρεάζεται από τις τιμές της συνάφειας και της περιττής πληροφορίας. Η κανονικοποίηση των δύο μεγεθών στη μέθοδο sMRMR_d αναμένεται να οδηγεί στην επιλογή λιγότερο συναφών αλλά και λιγότερο περιττών χαρακτηριστικών.



Σχήμα 4.4 Κατάταξη επιλεγμένων χαρακτηριστικών με βάση τη συνάφεια στο σύνολο δεδομένων Lymphoma



Σχήμα 4.5 Κατάταξη επιλεγμένων χαρακτηριστικών με βάση τη συνάφεια στο σύνολο δεδομένων Multiple features

Το σχήμα 4.4 επιβεβαιώνει τα παραπάνω. Στο σχήμα αυτό παρουσιάζεται ποια είναι κατά μέσο όρο η κατάταξη-ως προς τη συνάφεια με την κατηγορία-των 50 χαρακτηριστικών που επιλέχθηκαν από τις τρεις μεθόδους στο σύνολο δεδομένων Lymphoma που περιγράφεται στο 5^{ο} κεφάλαιο. Ο κατακόρυφος άξονας δείχνει την κατάταξη του χαρακτηριστικού με βάση τη συνάφεια με την κατηγορία και ο οριζόντιος δείχνει τον γύρο κατά τον οποίο επιλέχτηκε το χαρακτηριστικό. Για παράδειγμα, αν για τη μέθοδο sMRMR_d, η τιμή 20 ως προς τον οριζόντιο άξονα

αντιστοιχεί σε τιμή 250 ως προς τον κατακόρυφο άξονα, σημαίνει ότι το χαρακτηριστικό που επιλέγεται στον $20^{ο}$ γύρο από τη μέθοδο sMRMR_d έχει κατά μέσο όρο κατάταξη 250. Βλέπουμε ότι τα χαρακτηριστικά που επιλέγονται από τη μέθοδο mrmr_d έχουν κατά μέσο όρο κατάταξη μικρότερη από 100, δηλαδή τα 50 χαρακτηριστικά που επιλέγει η μέθοδος, προέρχονται συνήθως από το υποσύνολο των 100 πιο συναφών χαρακτηριστικών. Από την άλλη, η μέθοδος mrmr_q επιλέγει χαρακτηριστικά που η κατάταξη τους είναι κατά μέσο όρο μεγαλύτερη από 450. Αυτό ίσως σημαίνει ότι τα χαρακτηριστικά που επιλέγονται δεν είναι αρκετά συναφή. Τέλος η μέθοδος sMRMR_d φαίνεται να δίνει μια ισορροπία ανάμεσα στις δύο προηγούμενες παραλλαγές αφού επιλέγει χαρακτηριστικά από ένα μεγάλο εύρος της κατάταξης. Παρόμοια συμπεράσματα βγαίνουν εξετάζοντας τα χαρακτηριστικά που επιλέγονται από τις τρεις μεθόδους για το σύνολο δεδομένων Multiple features (επίσης περιγράφεται πιο αναλυτικά στο κεφάλαιο 5) στο σχήμα 4.5. Η διαφοροποίηση που προκύπτει στην απόδοση ελέγχεται πειραματικά στο κεφάλαιο 5.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΑ

5.1 Τα σύνολα δεδομένων
5.2 Αναλυτική περιγραφή μεθοδολογίας σύγκρισης
5.3 Σύγκριση mRMR και κανονικοποιημένου mRMR
5.4 Σύγκριση προσεγγίσεων για τον υπολογισμό της περιττής πληροφορίας

Στην ενότητα αυτή συγκρίνονται οι τροποποιήσεις που συζητούνται στο κεφάλαιο 4, με τις αυθεντικές εκδοχές του αλγορίθμου mRMR όπως προτείνονται στο [10]. Αρχικά, στην ενότητα 5.1 παρουσιάζονται τα σύνολα δεδομένων που χρησιμοποιήθηκαν για τη σύγκριση. Στην ενότητα 5.2 περιγράφονται τα πειράματα και οι παράμετροι που χρησιμοποιήθηκαν. Στην ενότητα 5.3 συγκρίνεται η αυθεντική εκδοχή του αλγορίθμου με την κανονικοποιημένη παραλλαγή που παρουσιάστηκε στην ενότητα 4.3. Τέλος, στην ενότητα 5.4, εξετάζονται οι τροποποιήσεις της ενότητας 4.1 που αφορούν στον υπολογισμό της περιττής πληροφορίας.

5.1. Τα σύνολα δεδομένων

Οι παραλλαγές που εξετάστηκαν στο κεφάλαιο 4, συγκρίνονται με βάση την ακρίβεια ταξινόμησης που επιτυγχάνεται όταν χρησιμοποιούνται τα αντίστοιχα επιλεγμένα υποσύνολα χαρακτηριστικών. Η σύγκριση γίνεται σε 6 διαφορετικά σύνολα δεδομένων. Από αυτά, 4 προέρχονται από το πεδίο της βιοπληροφορικής και χρησιμοποιούνται επίσης στο [10]. Τα παραδείγματα περιέχουν τους συντελεστές έκφρασης γονιδίων (συνήθως αρκετών χιλιάδων) και αντιστοιχούν σε κύτταρα με διαφορετικές ιδιότητες, π.χ. μπορεί να αναπαριστούν υγιή και μη υγιή κύτταρα ασθενών ή άλλες φορές να προέρχονται από ιστούς προσβεβλημένους από διαφορετικά είδη της ίδιας ασθένειας. Σκοπός είναι, να δημιουργηθεί ένας αξιόπιστος ταξινομητής που θα κάνει αυτόματη διάγνωση για νέα δείγματα με βάση τους

συντελεστές έκφρασης των γονιδίων. Η επιλογή χαρακτηριστικών, εκτός από απαραίτητο βήμα επεξεργασίας για την αποφυγή της υπερεκπαίδευσης (αφού η διάσταση είναι τεράστια και τα παραδείγματα πολύ λίγα), δίνει την ευκαιρία να ανακαλυφθούν γονίδια που σχετίζονται με κάθε νόσο.

Εκτός από τα σύνολα δεδομένων βιοπληροφορικής, χρησιμοποιήθηκαν ένα σύνολο χειρόγραφων χαρακτήρων και ένα σύνολο διαφημίσεων στο διαδίκτυο. Αναλυτικά τα σύνολα δεδομένων που χρησιμοποιήθηκαν είναι τα εξής:

Colon-cancer

Αυτό το σύνολο δεδομένων παρουσιάζεται στην εργασία [2]. Μετρήθηκε το επίπεδο έκφρασης 6500 γονιδίων από 62 δείγματα, εκ των οποίων τα 22 προέρχονται από υγιή κύτταρα ενώ τα υπόλοιπα 40 προέρχονται από καρκινικά κύτταρα. Οι συγγραφείς έχουν ήδη κάνει μία προεπεξεργασία και από το αρχικό σύνολο γονιδίων έχουν διατηρηθεί μόνο 2000. Το σύνολο δεδομένων των 2000 γονιδίων είναι διαθέσιμο στο διαδίκτυο [8].

Lymphoma

Αυτό το σύνολο δεδομένων παρουσιάζεται στην εργασία [1] και περιλαμβάνει 96 δείγματα που ανήκουν σε 9 κατηγορίες που αντιστοιχούν σε διαφορετικά υποείδη λεμφώματος διάχυτου από μεγάλα Β-κύτταρα (diffuse large B-cell lymphoma). Από το αρχικό πλήθος γονιδίων διατηρούνται από τους συγγραφείς 4026. Κάθε γονίδιο διακριτοποιείται σε τρεις τιμές με τον τρόπο που αναφέρθηκε στην ενότητα 3.3. Το διακριτοποιημένο σύνολο δεδομένων είναι διαθέσιμο στο διαδίκτυο [24].

Lung-cancer

Αυτό το σύνολο δεδομένων περιγράφεται στο [14] και αποτελείται απο 73 παραδείγματα που ανήκουν σε 7 διαφορετικές κατηγορίες που αντιστοιχούν σε διαφορετικούς τύπους αδενοκαρκινώματος του πνεύμονα. Το σύνολο δεδομένων που χρησιμοποιήθηκε αποτελείται από 325 χαρακτηριστικά διακριτοποιημένα σε τρεις καταστάσεις και είναι διαθέσιμο στο διαδίκτυο [24].

Leukemia

Αυτό το σύνολο δεδομένων [15] περιέχει δύο κατηγορίες που αντιστοιχούν στους τύπους λευχαιμίας ALL (acute lymphoblastic leukemia) και AML (acute myeloid leukemia). Υπάρχουν 47 δείγματα της πρώτης και 25 δείγματα της δεύτερης κατηγορίας που περιγράφονται από 7129 γονίδια. Τα 72 δείγματα χωρίστηκαν από τους συγγραφείς σε δύο υποσύνολα από 38 και 34 δείγματα που χρησιμοποιήθηκαν ως σύνολα εκπαίδευσης και ελέγχου αντίστοιχα. Τα ίδια σύνολα ελέγχου και εκπαίδευσης χρησιμοποιούνται και σε αυτή την εργασία και είναι διαθέσιμα στο διαδίκτυο [8].

Multiple-features

Αυτό είναι ένα σύνολο δεδομένων χειρόγραφων ψηφίων. Αποτελείται από 2000 παραδείγματα που ανήκουν σε 10 διαφορετικές κατηγορίες, 200 σε κάθε μία, που αντιστοιχούν στα ψηφία ‘0’ έως ‘9’. Τα ψηφία περιγράφονται από 649 συνεχή χαρακτηριστικά τα οποία διακριτοποιήθηκαν σε 3 καταστάσεις με βάση τη διαδικασία που περιγράφηκε στην ενότητα 3.3. Το σύνολο δεδομένων είναι διαθέσιμο στο διαδίκτυο [3].

Internet-advertisments

Το σύνολο αυτό περιέχει την περιγραφή εικόνων από ιστοσελίδες από τις οποίες κάποιες αντιστοιχούν σε διαφημίσεις. Η περιγραφή βασίζεται σε γεωμετρικά χαρακτηριστικά της εικόνας όπως το πλάτος και το ύψος, αλλά και σε δυαδικά χαρακτηριστικά των οποίων η τιμή εξαρτάται από το αν συγκεκριμένες λέξεις βρίσκονται στο url της εικόνας. Στο σύνολο υπάρχουν 3279 παραδείγματα εκ των οποίων τα 459 αντιστοιχούν σε διαφημίσεις. Τα χαρακτηριστικά είναι 1559. Τα συνεχή χαρακτηριστικά διακριτοποιούνται σε τρεις τιμές με βάση τη διαδικασία της ενότητας 3.3. Το σύνολο δεδομένων είναι διαθέσιμο στο διαδίκτυο [3].

Πίνακας 5.1 Σύνοψη των συνόλων δεδομένων

Σύνολο δεδομένων	Αριθμός παραδειγμάτων	Αριθμός χαρακτηριστικών	Αριθμός κατηγοριών
Colon-cancer	62	2000	2
Leukemia	72	7129	2
Lymphoma	96	4026	9
Lung-cancer	73	325	7
Multiple features	2000	649	10
Internet Advertisments	3279	1559	2

5.2. Αναλυτική περιγραφή μεθοδολογίας σύγκρισης

5.2.1. Αποτίμηση ακρίβειας ταξινόμησης

Για την αποτίμηση μίας μεθόδου, το σύνολο δεδομένων διασπάται σε δύο υποσύνολα παραδειγμάτων, με το πρώτο να χρησιμοποιείται ως σύνολο εκπαίδευσης και το άλλο ως σύνολο ελέγχου. Η διαδικασία επαναλαμβάνεται k φορές, δημιουργούνται δηλαδή k ζεύγη συνόλων εκπαίδευσης και ελέγχου. Οι διασπάσεις γίνονται με βάση τη διαδικασία cross validation [22], δηλαδή με τρόπο τέτοιο ώστε κάθε παράδειγμα του αρχικού συνόλου δεδομένων να τοποθετηθεί στο σύνολο ελέγχου ακριβώς μία φορά. Στα προβλήματα βιοπληροφορικής δημιουργούνται διασπάσεις σύμφωνα με τη διαδικασία “leave-one-out cross-validation” (LOOCV) [22], δηλαδή το σύνολο ελέγχου αποτελείται κάθε φορά από ένα παράδειγμα. Ο λόγος για αυτήν την επιλογή είναι ότι στα προβλήματα βιοπληροφορικής ο αριθμός των παραδειγμάτων είναι μικρός και δεν αφήνει περιθώριο για το σχηματισμό ενός μεγάλου συνόλου ελέγχου καθώς τότε προκύπτει υπερβολικά μικρό σύνολο εκπαίδευσης. Τα σύνολα δεδομένων Multiple Features και Internet Ads, για τα οποία πολλά παραδείγματα είναι διαθέσιμα, διασπώνται σε 3 ζεύγη συνόλων εκπαίδευσης και ελέγχου (3-fold cross-validation). Η τελική εκτίμηση για την ακρίβεια ταξινόμησης υπολογίζεται ως η μέση τιμή της ακρίβειας που επιτεύχθηκε στα διαφορετικά σύνολα ελέγχου.

Η επιλογή χαρακτηριστικών γίνεται λαμβάνοντας υπόψιν μόνο το σύνολο εκπαίδευσης. Αυτό σημαίνει ότι για κάθε σύνολο εκπαίδευσης που προήλθε από τις προηγούμενες διασπάσεις, ένα ξεχωριστό υποσύνολο χαρακτηριστικών επιλέγεται. Η διαδικασία αυτή διαφέρει από τη διαδικασία που ακολουθείται στο [10], όπου επιλέγεται ένα σύνολο χαρακτηριστικών με βάση όλα τα παραδείγματα και στη συνέχεια αποτιμάται με χρήση leave-one-out cross-validation. Θεωρούμε ότι η αποτίμηση μιας μεθόδου με αυτόν τον τρόπο δεν είναι ακριβής γιατί σε κάθε περίπτωση η διαδικασία της επιλογής χαρακτηριστικών λαμβάνει υπόψιν τα παραδείγματα ελέγχου.

5.2.2. Ο ταξινομητής και η επιλογή παραμέτρων

Για την ταξινόμηση χρησιμοποιείται ο ταξινομητής SVM (support vector machine) [7],[30]. Χρησιμοποιείται η δημοφιλής υλοποίηση LIBSVM που είναι διαθέσιμη στο διαδίκτυο [8]. Ένα χαρακτηριστικό του ταξινομητή SVM είναι ότι μπορεί να υλοποιηθεί με πολλές διαφορετικές συναρτήσεις πυρήνα (kernel functions). Στα πειράματα χρησιμοποιήθηκε γραμμικός πυρήνας (linear kernel) καθώς και ο πυρήνας RBF. Ο ταξινομητής SVM απαιτεί τον καθορισμό μίας παραμέτρου C που είναι ένας όρος ποινής που εκφράζει την ανεκτικότητα σε εσφαλμένες κατατάξεις. Αν επιπλέον χρησιμοποιείται ο πυρήνας RBF, απαιτείται και ο καθορισμός μιας παραμέτρου g που αντιστοιχεί στο εύρος του πυρήνα.

Για την επιλογή των παραμέτρων C και g ένα σύνολο διαφορετικών συνδυασμών δοκιμάζεται και αποτιμάται με χρήση k-fold cross validation. Αυτή η διαδικασία δεν έχει σχέση με τον εξωτερικό βρόχο cross validation που περιγράφηκε στην ενότητα 5.2.1 και δημιουργεί διασπάσεις των διαθέσιμων δεδομένων σε σύνολα εκπαίδευσης και ελέγχου. Η διαδικασία που περιγράφεται εδώ, καλείται εσωτερικά για κάθε σύνολο εκπαίδευσης που προέκυψε από τις διασπάσεις, και συγκρίνει διαφορετικές παραμέτρους C, g μέσω cross validation. Για καλύτερη κατανόηση, η διαδικασία που χρησιμοποιείται για την αποτίμηση ενός υποσυνόλου χαρακτηριστικών περιγράφεται από τον ψευδοκώδικα 5.1.

Αλγόριθμος 5.1 Διαδικασία αποτίμησης υποσυνόλου χαρακτηριστικών

Είσοδος: σύνολο δεδομένων D, υποσύνολο χαρακτηριστικών S

Έξοδος: ακρίβεια πρόβλεψης subset_accuracy

1) Έστω D_S η προβολή του D στο υποσύνολο χαρακτηριστικών S.
2) Διάσπασε το D_S σε k ζεύγη συνόλων εκπαίδευσης και ελέγχου, έστω $D_{trai}(1),\ldots,D_{trai}(k)$ και $D_{test}(1),\ldots,D_{test}(k)$ αντίστοιχα
3) Για i από 1 έως k,
4) Αξιολόγησε κάθε διαφορετικό συνδυασμό παραμέτρων C,g με χρήση cross-validation στο $D_{trai}(i)$ και επέλεξε τον καλύτερο, έστω C',g'
5) Εκπαίδευσε το SVM με είσοδο το $D_{trai}(i)$ και παραμέτρους C',g'
6) Μέτρησε την ακρίβεια πρόβλεψης στο $D_{test}(i)$, έστω acc
7) overall_acc = overall_acc + acc
8) Τέλος
9) subset_accuracy = overall_acc/k

Στην $4^{η}$ γραμμή του αλγορίθμου εκτελείται ο εσωτερικός βρόχος cross-validation για κάθε υποψήφιο συνδυασμό παραμέτρων C και g. Υπονοείται ότι έχουν καθοριστεί οι πιθανοί συνδυασμοί που θα εξεταστούν, καθώς και ο αριθμός των folds που θα χρησιμοποιηθούν για την cross validation διαδικασία. Στον πίνακα 5.2 συνοψίζονται οι παράμετροι αυτοί όπως χρησιμοποιήθηκαν για τα διάφορα σύνολα δεδομένων.

Πίνακας 5.2 Παράμετροι που χρησιμοποιήθηκαν κατά την αποτίμηση υποσυνόλων και την εκπαίδευση του SVM

Σύνολο δεδομένων	Αριθμός διασπάσεων	Συνάρτηση Πυρήνα	Τιμές παραμέτρου C	Τιμές παραμέτρου g	Αριθμός διασπάσεων για εσωτερικό βρόχο CV
Colon-cancer	62(LOOCV)	Γραμμική	$2^{-5},2^{-4},...,2^{9},2^{10}$	-	5
Leukemia	1	Γραμμική	$2^{-5},2^{-4},...,2^{9},2^{10}$	-	5
Lymphoma	96(LOOCV)	Γραμμική	$2^{-5},2^{-4},...,2^{9},2^{10}$	-	5
Lung-cancer	73(LOOCV)	Γραμμική	$2^{-5},2^{-4},...,2^{9},2^{10}$	-	5
Multiple features	3	RBF	$2^{-3},2^{-1},...,2^{7},2^{9}$	$2^{-3},2^{-5},2^{-7},2^{-9},2^{-11}$	3
Internet Advertisments	3	Γραμμική	$2^{-5},2^{-4},...,2^{9},2^{10}$	-	3

5.2.3. Σύγκριση αποτελεσμάτων

Σε αντίθεση με άλλους αλγορίθμους, ο mRMR απαιτεί τον καθορισμό του μεγέθους του υποσυνόλου που θα επιλεγεί. Η σύγκριση δύο παραλλαγών του mRMR χρησιμοποιώντας υποσύνολα συγκεκριμένου μεγέθους, επιλεγμένου με αυθαίρετο τρόπο, δεν μπορεί να είναι αντικειμενική. Είναι πολύ πιθανό μία μέθοδος που φαίνεται να είναι καλύτερη όταν χρησιμοποιούνται 50 χαρακτηριστικά, να είναι πολύ χειρότερη αν χρησιμοποιούνται 20 χαρακτηριστικά. Για το λόγο αυτό οι μέθοδοι συγκρίνονται με βάση υποσύνολα πολλών διαφορετικών μεγεθών. Αν για παράδειγμα μία μέθοδος επιλέγει το υποσύνολο $S_k=\{X_1,X_2,...,X_k\}$, αποτιμώνται όλα τα εμφωλευμένα υποσύνολα του S_k μεγέθους 1 έως k. Το εμφωλευμένο υποσύνολο μεγέθους i είναι αυτό που αποτελείται από τα χαρακτηριστικά $X_1,X_2,...,X_i$. Στην προηγούμενη πρόταση υποθέτουμε ότι τα χαρακτηριστικά έχουν επιλεγεί με τη σειρά που εμφανίζονται μέσα στο S_k.

Από τη σύγκριση με βάση τα εμφωλευμένα υποσύνολα, συχνά δεν προκύπτει σαφής υπεροχή υπέρ μίας μεθόδου. Αυτό συμβαίνει για παράδειγμα αν μία μέθοδος επιτυγχάνει αρκετά καλή ακρίβεια ταξινόμησης με 20 χαρακτηριστικά, και μία δεύτερη μέθοδος επιτυγχάνει ελάχιστα καλύτερη ακρίβεια ταξινόμησης αλλά χρησιμοποιώντας 30 χαρακτηριστικά. Εφόσον ζητείται η ελαχιστοποίηση του

αριθμού των χαρακτηριστικών και ταυτόχρονα η μεγιστοποίηση της ακρίβειας ταξινόμησης, δεν υπάρχει ξεκάθαρος νικητής ανάμεσα στις μεθόδους.

5.3. Σύγκριση mRMR και κανονικοποιημένου mRMR

Στην ενότητα αυτή συγκρίνονται οι δύο εκδοχές του mRMR (όπως παρουσιάστηκαν στην ενότητα 3.2) με την παραλλαγή της ενότητας 4.3 που κανονικοποιεί τα μεγέθη της συνάφειας και της περιττής πληροφορίας. Όπως και στην ενότητα 4.3 χρησιμοποιούμε τα αρχικά *mrmr_d* και *mrmr_q* για να αναφερθούμε στη μέθοδο mRMR που μεγιστοποιεί τη διαφορά και το λόγο αντίστοιχα μεταξύ συνάφειας και περιττής πληροφορίας. Για την κανονικοποιημένη παραλλαγή του αλγορίθμου χρησιμοποιούνται τα αρχικά sMRMR_d (scaled-mrmr-difference).

Τα επιλεγμένα υποσύνολα από κάθε μέθοδο αποτιμώνται με τη διαδικασία cross-validation. Στα σύνολα δεδομένων βιοπληροφορικής αποτιμήθηκαν όλα τα εμφωλευμένα υποσύνολα που περιέχουν από 1 έως 50 χαρακτηριστικά. Στα σύνολα δεδομένων Multiple features και Internet-Ads, που είναι πολύ μεγαλύτερα σε μέγεθος, αποτιμήθηκαν τα εμφωλευμένα υποσύνολα με άρτιο αριθμό χαρακτηριστικών και μέγιστο μέγεθος 60. Τα υποσύνολα περιττού μεγέθους δεν αποτιμήθηκαν προκειμένου να συντομευτεί η διαδικασία, που είναι χρονοβόρα παρά το γεγονός ότι χρησιμοποιείται 3-fold cross validation.

Οι πίνακες 5.4 έως 5.7 παρουσιάζουν μια σύνοψη του αριθμού των λανθασμένων προβλέψεων που έγιναν κατά τη διαδικασία leave-one-out cross validation χρησιμοποιώντας τα υποσύνολα κάθε μεθόδου. Η δεύτερη στήλη παρουσιάζει τον ελάχιστο αριθμό λαθών που έγινε χρησιμοποιώντας εμφωλευμένα υποσύνολα με 1 έως 5 χαρακτηριστικά. Στην τρίτη στήλη εμφανίζεται ο ελάχιστος αριθμός λαθών που έγινε χρησιμοποιώντας εμφωλευμένα υποσύνολα μεγέθους 6 έως 10 κ.ο.κ. Στον πίνακα 5.3 δίνεται η ακρίβεια ταξινόμησης που επιτυγχάνεται χρησιμοποιώντας όλα τα χαρακτηριστικά στα διάφορα σύνολα δεδομένων ώστε να είναι αντιληπτό αν και κατά πόσο η επιλογή χαρακτηριστικών βελτιώνει την απόδοση.

Πίνακας 5.3 Ακρίβεια ταξινόμησης που επιτυγχάνεται χρησιμοποιώντας όλα τα χαρακτηριστικά

Σύνολο δεδομένων	Ακρίβεια ταξινόμησης/Λάθη
Colon-cancer	83.87 %(10 λάθη)
Leukemia	82.35% (6 λάθη)
Lymphoma	95.83% (4 λάθη)
Lung-cancer	87.67% (9 λάθη)
Multiple features	98.6%
Internet Advertisments	97.46%

Ξεκινώντας από το σύνολο δεδομένων colon-cancer (πίνακας 5.4), θα λέγαμε ότι δεν υπάρχει εμφανής διαφορά στην απόδοση των τριών μεθόδων στο σύνολο αυτό. Και οι τρεις μέθοδοι οδηγούν σε βελτίωση της απόδοσης σε σχέση με την περίπτωση που χρησιμοποιούνται όλα τα χαρακτηριστικά. Συγκεκριμένα καταφέρνουν στην καλύτερη περίπτωση να κάνουν 7 λάθη αντί για 10 στις συνολικά 62 προβλέψεις. Η μέθοδος sMRMR_d πετυχαίνει πρώτη φορά αυτή την απόδοση χρησιμοποιώντας 10 χαρακτηριστικά, ενώ 12 χρειάζεται η μέθοδος mrmr_d και 19 η μέθοδος mrmr_q. Στο σύνολο δεδομένων Lymphoma (πίνακας 5.5) η μέθοδος mrmr_q φαίνεται να είναι η χειρότερη καθώς κάνει κατά κανόνα περισσότερα λάθη για υποσύνολα ίδιου μεγέθους και καταφέρνει στην καλύτερη περίπτωση να κάνει 7 λάθη τη στιγμή που οι μέθοδοι mrmr_d και sMRMR_d κάνουν 5 και 4 λάθη αντίστοιχα. Η μέθοδος mrmr_d κάνει 5 λάθη χρησιμοποιώντας 26 χαρακτηριστικά, σε σχέση με την sMRMR_d που κάνει 4 λάθη με 28 χαρακτηριστικά. Αν και η mrmr_d δεν κάνει σε καμία περίπτωση λιγότερα από 5 λάθη, καταφέρνει με μόλις 12 χαρακτηριστικά να κάνει σχετικά λίγα λάθη, δηλαδή 7. Σε αυτή την περίπτωση δεν μπορεί να θεωρηθεί ότι μία μέθοδος παράγει καλύτερα αποτελέσματα από την άλλη. Στο σύνολο δεδομένων lung-cancer (πίνακας 5.6), οι μέθοδοι sMRMR_d και mrmr_q υπερέχουν καθαρά έναντι της mrmr_d που καταφέρνει στην καλύτερη περίπτωση να κάνει 11 λάθη ενώ οι δύο πρώτες καταφέρνουν να κάνουν μόνο 6 λάθη. Τα υποσύνολα της μεθόδου sMRMR_d οδηγούν κατά κανόνα σε λιγότερα λάθη από τα υποσύνολα της μεθόδου mrmr_q. Στο σύνολο δεδομένων Leukemia (πίνακας 5.7) η μέθοδος mrmr_d υπερέχει, κάνοντας μόνο μία λάθος πρόβλεψη στο σύνολο ελέγχου το οποίο περιέχει 34 παραδείγματα. Η μέθοδος sMRMR_d δεν κάνει σε καμία περίπτωση λιγότερα από 3 λάθη, αλλά υπερέχει της μεθόδου mrmr_q για υποσύνολα με μέγεθος 1 έως 30.

Εκτός των παραπάνω παρατηρούμε ότι σε όλα τα σύνολα δεδομένων βιοπληροφορικής εκτός από το Lymphoma η επιλογή χαρακτηριστικών βελτιώνει την απόδοση.

Πίνακας 5.4 Αριθμός λαθών στο σύνολο δεδομένων Colon-cancer

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	10	9	7	7	7	7	7	7	7	7
mrmr_q	17	10	8	7	7	7	7	7	7	7
sMRMR_d	10	7	7	7	7	7	8	7	7	7

Πίνακας 5.5 Αριθμός λαθών στο σύνολο δεδομένων Lymphoma

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	26	12	7	11	6	5	5	5	5	5
mrmr_q	37	30	14	14	15	12	11	12	8	7
sMRMR_d	30	15	14	9	6	4	6	6	4	4

Πίνακας 5.6 Αριθμός λαθών στο σύνολο δεδομένων Lung-cancer

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	27	13	15	14	12	16	14	11	12	12
mrmr_q	36	29	19	14	13	12	10	8	7	6
sMRMR_d	30	20	17	14	12	12	8	7	6	6

Πίνακας 5.7 Αριθμός λαθών στο σύνολο δεδομένων Leukemia

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	4	5	4	4	3	3	2	2	1	1
mrmr_q	8	8	5	6	6	5	4	3	3	4
sMRMR_d	4	3	4	4	3	4	4	4	4	3

Στον πίνακα 5.8 συνοψίζονται τα αποτελέσματα για το σύνολο δεδομένων Multiple features. Η ακρίβεια ταξινόμησης που επιτυγχάνει η παραλλαγή sMRMR_d είναι

κατά κανόνα καλύτερη από την ακρίβεια που πετυχαίνουν οι μέθοδοι mrmr_d και mrmr_q με υποσύνολα αντίστοιχου μεγέθους. Η μόνη περίπτωση που δεν ισχύει αυτό είναι για υποσύνολα μεγέθους 11 έως 20 χαρακτηριστικών όπου η μέθοδος mrmr_d δίνει καλύτερη ακρίβεια. Η μέθοδος mrmr_q δίνει χειρότερα αποτελέσματα από την mrmr_d για μικρό αριθμό χαρακτηριστικών, αλλά η απόδοση της είναι καλύτερη όταν ο αριθμός των χαρακτηριστικών αυξάνεται. Στο σύνολο δεδομένων Internet-ads (πίνακας 5.9) οι τρεις μέθοδοι έχουν παρόμοια απόδοση με την mrmr_q να εμφανίζεται ελαφρώς καλύτερη και την sMRMR_d να ακολουθεί. Η ακρίβεια ταξινόμησης 96.16% που επιτυγχάνεται με μόλις 28 χαρακτηριστικά από τη μέθοδο sMRMR_d δεν επιτυγχάνεται από κανένα υποσύνολο από αυτά που δίνει η mrmr_d.

Πίνακας 5.8 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Multiple Features

	1-10	11-20	21-30	31-40	41-50	51-60
mrmr_d	91.55	97.25	97.85	98.2	98.2	98.4
mrmr_q	85.5	94.65	97	98.25	98.6	98.55
sMRMR_d	93.15	96.9	98	98.6	98.7	98.75

Πίνακας 5.9 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Internet-Ads

	1-10	11-20	21-30	31-40	41-50	51-60
mrmr_d	95.12	95.76	95.85	95.97	96.04	96.07
mrmr_q	95.15	96.25	96.4	96.37	96.61	96.65
sMRMR_d	95.67	95.73	96.16	96.28	96.49	96.65

Από την παραπάνω σύγκριση φαίνεται ότι υπάρχουν σύνολα δεδομένων στα οποία αποδίδει καλύτερα η μέθοδος mrmr_d και άλλα στα οποία αποδίδει καλύτερα η μέθοδος mrmr_q. Η μέθοδος sMRMR_d φαίνεται σε κάθε σύνολο δεδομένων να είναι εξ ίσου καλή ή και να υπερέχει της καλύτερης από τις άλλες δύο. Ένα στοιχείο που προκύπτει είναι ότι η μέθοδος mrmr_q υστερεί έναντι των άλλων δύο για υποσύνολα μικρού μεγέθους γεγονός που ίσως οφείλεται στο ότι τα χαρακτηριστικά που επιλέγει, ειδικά κατά τους πρώτους γύρους, δεν είναι αρκετά συναφή. Επίσης παρατηρείται ότι η καλύτερη απόδοση για τη μέθοδο mrmr_d είναι χειρότερη από την

καλύτερη απόδοση της μεθόδου sMRMR_d, σε κάθε σύνολο δεδομένων εκτός από αυτό της λευχαιμίας. Το γεγονός ότι η mrmr_d δίνει έμφαση στην επιλογή πιο συναφών χαρακτηριστικών, παρόλο που αυτά κατά κανόνα περιέχουν περισσότερη περιττή πληροφορία [10], μάλλον ευθύνεται για το ότι η προσθήκη επιπλέον χαρακτηριστικών δεν συνοδεύεται από αρκετά μεγάλη αύξηση της απόδοσης.

5.4. Σύγκριση προσεγγίσεων για τον υπολογισμό της περιττής πληροφορίας

Στην ενότητα αυτή συγκρίνονται πειραματικά οι παραλλαγές που διατυπώθηκαν στην ενότητα 4.1 και αφορούν στον υπολογισμό της περιττής πληροφορίας που περιέχει ένα χαρακτηριστικό. Αναφερόμαστε στην παραλλαγή του mRMR που χρησιμοποιεί την max-redundancy προσέγγιση (Εξ. 4.2) ως mrmr_max και στην παραλλαγή που χρησιμοποιεί την weighted-redundancy προσέγγιση (Εξ. 4.4) ως mrmr_w.

Η σύγκριση έγινε με τον τρόπο που παρουσιάστηκε στην ενότητα 5.3. Στο σύνολο δεδομένων colon-cancer (πίνακας 5.10) η παραλλαγή mrmr_max επιτυγχάνει τον ελάχιστο αριθμό λαθών, δηλαδή 6 χρησιμοποιώντας 14 χαρακτηριστικά. Για υποσύνολα άλλου μεγέθους όμως δεν καταφέρνει σε καμία περίπτωση να κάνει λιγότερα από 8 λάθη και η απόδοση της επιδεινώνεται όσο προστίθενται περισσότερα χαρακτηριστικά. Οι mrmr_d και mrmr_w κάνουν 7 λάθη στην καλύτερη περίπτωση, με την πρώτη να χρειάζεται λιγότερα χαρακτηριστικά για να πετύχει αυτήν την απόδοση. Στο σύνολο δεδομένων Lymphoma (πίνακας 5.11) η παραλλαγή mrmr_max είναι χειρότερη από τις άλλες δύο. Όλες οι μέθοδοι οδηγούν στην καλύτερη περίπτωση σε 5 λάθη, με τη μέθοδο mrmr_w να είναι αυτή που το καταφέρνει πιο νωρίς. Στο σύνολο δεδομένων lung-cancer (πίνακας 5.12) η μέθοδος mrmr_d φαίνεται να υπερέχει ελαφρώς για μικρού μεγέθους υποσύνολα αλλά οι άλλες δύο μέθοδοι οδηγούν σε μικρότερο αριθμό λαθών όταν πλέον χρησιμοποιούνται πολλά χαρακτηριστικά (26 έως 50). Στο σύνολο δεδομένων Leukemia (πίνακας 5.13) η μέθοδος mrmr_d δίνει το υποσύνολο που οδηγεί στην καλύτερη επίδοση με μόλις 1 λάθος. Όμως η παραλλαγή mrmr_max καταφέρνει με 7 χαρακτηριστικά να κάνει 2 λάθη ενώ η mrmr_w με μόλις 3 χαρακτηριστικά πετυχαίνει 3 λάθη. Στο σύνολο δεδομένων Multiple Features (πίνακας 5.14) η μέθοδος mrmr_max δίνει την καλύτερη ακρίβεια, δηλαδή 98.5%, χρησιμοποιώντας 50 χαρακτηριστικά ενώ η

mrmr_w πετυχαίνει ακρίβεια 98.4%, επίσης με 50 χαρακτηριστικά. Για μικρότερο μέγεθος υποσυνόλων δεν ξεχωρίζει καμία από τις τρεις μεθόδους. Στο σύνολο Internet Ads (πίνακας 5.15), η μέθοδος mrmr_w επιτυγχάνει την καλύτερη απόδοση 96.52% και ακολουθεί η mrmr_max με 96.46% ενώ η mrmr_d σε καμία περίπτωση δεν ξεπερνά το 96.07%.

Πίνακας 5.10 Αριθμός λαθών στο σύνολο δεδομένων Colon-cancer

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	10	9	7	7	7	7	7	7	7	7
mrmr_max	13	8	6	8	9	10	11	11	13	12
mrmr_w	10	8	8	7	7	8	8	7	7	8

Πίνακας 5.11 Αριθμός λαθών στο σύνολο δεδομένων Lymphoma

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	26	12	7	11	6	5	5	5	5	5
mrmr_max	29	16	13	13	9	7	6	6	5	5
mrmr_w	29	15	7	7	5	5	6	6	7	7

Πίνακας 5.12 Αριθμός λαθών στο σύνολο δεδομένων Lung-cancer

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	27	13	15	14	12	16	14	11	12	12
mrmr_max	28	20	17	17	14	14	13	12	10	11
mrmr_w	26	17	16	16	15	13	13	11	11	10

Πίνακας 5.13 Αριθμός λαθών στο σύνολο δεδομένων Leukemia

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
mrmr_d	4	5	4	4	3	3	2	2	1	1
mrmr_max	6	2	4	3	3	3	3	3	3	2
mrmr_w	3	4	4	3	3	3	2	2	2	2

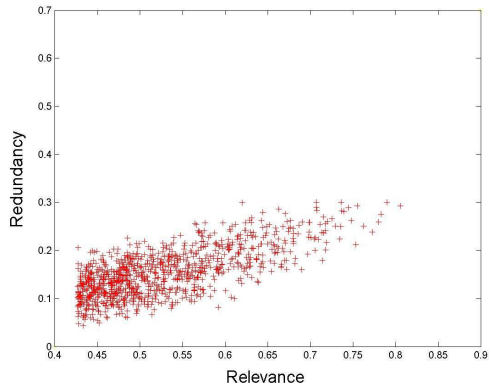
Πίνακας 5.14 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Multiple Features

	1-10	11-20	21-30	31-40	41-50	51-60
mrmr_d	91.55	97.25	97.85	98.2	98.2	98.4
mrmr_max	91.8	96.85	98.15	98.05	98.5	98.5
mrmr_w	93.95	97	97.85	98.1	98.4	98.3

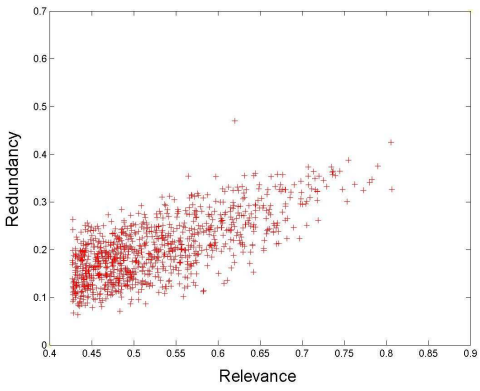
Πίνακας 5.15 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Internet-Ads

	1-10	11-20	21-30	31-40	41-50	51-60
mrmr_d	95.12	95.76	95.85	95.97	96.04	96.07
mrmr_max	95.52	96	96.34	96.31	96.46	96.46
mrmr_w	95.49	95.88	96.07	96.43	96.52	96.37

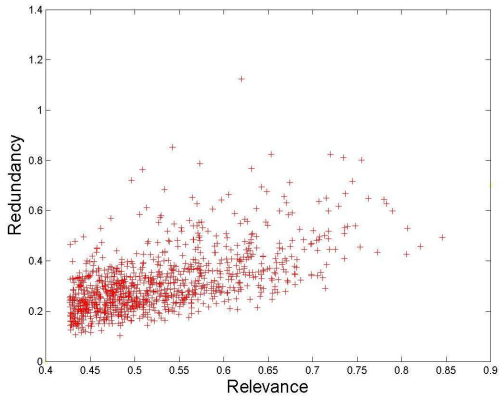
Τα παραπάνω αποτελέσματα δεν επιτρέπουν την εξαγωγή ασφαλών συμπερασμάτων αφού οι διαφορές είναι μικρές και δεν υπάρχει κάποια μέθοδος που σταθερά να ξεπερνά τις άλλες. Μία εκ των υστέρων διαπίστωση είναι ότι η περιττή πληροφορία υπολογισμένη με τις προσεγγίσεις weighted-redundancy και max-redundancy μπορεί να έχει μεγαλύτερη διακύμανση από αυτή που έχει όταν υπολογίζεται με την προσέγγιση mean-redundancy. Αυτό φαίνεται για παράδειγμα στο σχήμα 5.1. Στη σύγκριση των προσεγγίσεων εμπλέκεται το πρόβλημα της άνισης διακύμανσης που αναλύθηκε στην ενότητα 4.3 και 5.3. Η mean-redundancy προσέγγιση θα μπορούσε να είναι η καλύτερη για την εκτίμηση της περιττής πληροφορίας, παρόλα αυτά οι παραλλαγές mrmr_max και mrmr_w υπερέχουν έναντι της mrmr_d, γιατί αντιμετωπίζουν λιγότερο έντονα το πρόβλημα της άνισης διακύμανσης. Για να ελεγχθεί η παραπάνω υπόθεση, οι τρεις διαφορετικές προσεγγίσεις χρησιμοποιούνται σε συνδυασμό με την κανονικοποιημένη εκδοχή του mRMR. Αναφερόμαστε στο συνδυασμό του κανονικοποιημένου mRMR με τις προσεγγίσεις max-redundancy και weighted-redundancy ως sMRMR_max και sMRMR_w.



a) Mean-redundancy προσέγγιση



b) weighted-redundancy προσέγγιση



c) max-redundancy προσέγγιση

Σχήμα 5.1 Συνάφεια και περιττή πληροφορία υποψηφίων χαρακτηριστικών κατά τη δέκατη επανάληψη του mRMR με βάση τις τρεις διαφορετικές προσεγγίσεις (σύνολο δεδομένων Lymphoma).

Τα αποτελέσματα της σύγκρισης παρουσιάζονται στους πίνακες 5.16 έως 5.21. Στο σύνολο δεδομένων colon-cancer δεν υπάρχει σημαντική διαφορά μεταξύ των μεθόδων. Στο σύνολο δεδομένων Lymphoma ωστόσο, η μέθοδος sMRMR_d πετυχαίνει χρησιμοποιώντας λιγότερα χαρακτηριστικά καλύτερη απόδοση. Η διαφορά είναι ακόμα μεγαλύτερη υπέρ της sMRMR_d στο σύνολο δεδομένων lung-cancer. Η μέθοδος sMRMR_d υπερέχει επίσης στο σύνολο δεδομένων Multiple-features ενώ δεν υπάρχει σημαντική διαφορά στο σύνολο Internet-Ads. Οι δύο παραλλαγές sMRMR_w και sMRMR_max δίνουν καλύτερα αποτελέσματα μόνο στο σύνολο δεδομένων Leukemia. Τα αποτελέσματα δείχνουν ότι έχοντας διορθώσει το πρόβλημα της άνισης διακύμανσης συνάφειας και περιττής πληροφορίας, η

προσέγγιση mean-redundancy είναι πιο συνεπής στο να παράγει καλά υποσύνολα χαρακτηριστικών.

Χωρίς ο λόγος να είναι εύκολα αντιληπτός, φαίνεται πως η mean-redundancy προσέγγιση είναι καταλληλότερη για την εκτίμηση της ομοιότητας ενός χαρακτηριστικού με τα ήδη επιλεγμένα. Ένας λόγος για τον οποίο μπορεί να συμβαίνει αυτό είναι ο εξής. Έστω ότι W(X) είναι η πραγματική περιττή πληροφορία που περιέχει το χαρακτηριστικό X και W'(X) η περιττή πληροφορία όπως υπολογίζεται με βάση κάποια προσέγγιση. Δεν είναι ιδιαίτερα σημαντικό το W'(X) να προσεγγίζει ικανοποιητικά το W(X), είναι όμως σημαντικό για δύο χαρακτηριστικά X_1 και X_2 για τα οποία ισχύει ότι $W(X_1)<W(X_2)$, να ισχύει επίσης ότι $W'(X_1)<W'(X_2)$. Με άλλα λόγια έχει σημασία η κατάταξη των χαρακτηριστικών ως προς την προσέγγιση να είναι όσο το δυνατόν πιο κοντά στην κατάταξη ως προς την πραγματική περιττή πληροφορία ακόμα και αν τα μεγέθη μεταβάλλονται. Ίσως λοιπόν η mean-redundancy προσέγγιση διατηρεί καλύτερα την κατάταξη των χαρακτηριστικών σε σχέση με τις άλλες δύο προσεγγίσεις. Η επαλήθευση της υπόθεσης αυτής χρήζει περισσότερης μελέτης.

Πίνακας 5.16 Αριθμός λαθών στο σύνολο δεδομένων Colon-cancer

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
sMRMR_d	10	7	7	7	7	7	8	7	7	7
sMRMR_max	9	9	8	7	7	8	8	7	7	8
sMRMR_w	10	8	7	7	7	7	7	8	7	9

Πίνακας 5.17 Αριθμός λαθών στο σύνολο δεδομένων Lymphoma

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
sMRMR_d	30	15	14	9	6	4	6	6	4	4
sMRMR_max	27	22	18	11	9	7	6	7	8	8
sMRMR_w	28	18	14	10	8	6	6	7	8	5

Πίνακας 5.18 Αριθμός λαθών στο σύνολο δεδομένων Lung cancer

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
sMRMR_d	30	20	17	14	12	12	8	7	6	6
sMRMR_max	32	20	16	15	13	13	14	10	11	10
sMRMR_w	30	16	17	14	13	13	11	11	8	9

Πίνακας 5.19 Αριθμός λαθών στο σύνολο δεδομένων Leukemia

	1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40	41-45	46-50
sMRMR_d	4	3	4	4	3	4	4	4	4	3
sMRMR_max	3	3	3	4	3	3	3	2	1	1
sMRMR_w	4	4	3	2	2	3	2	2	2	3

Πίνακας 5.20 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Multiple Features

	1-10	11-20	21-30	31-40	41-50	51-60
sMRMR_d	93.15	96.9	98	98.6	98.7	98.75
sMRMR_max	91.15	96.7	97.8	98.25	98.3	98.4
sMRMR_w	92.5	97.45	98.05	97.95	98.3	98.4

Πίνακας 5.21 Ακρίβεια ταξινόμησης στο σύνολο δεδομένων Internet-Ads

	1-10	11-20	21-30	31-40	41-50	51-60
sMRMR_d	95.67	95.73	96.16	96.28	96.49	96.65
sMRMR_max	95.49	95.7	96.07	96.25	96.52	96.28
sMRMR_w	95.61	95.82	96.07	96.22	96.55	96.46

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΜΕΛΛΟΝΤΙΚΗ ΕΡΓΑΣΙΑ

Στην εργασία αυτή μελετήθηκε το πρόβλημα της επιλογής χαρακτηριστικών για προβλήματα ταξινόμησης. Αφού παρουσιάστηκαν οι βασικές κατηγορίες μεθόδων, δόθηκε ιδιαίτερη έμφαση στις μεθόδους που στοχεύουν στη μείωση της περιττής πληροφορίας. Αναλυτικότερα εξετάστηκε ο αλγόριθμος mRMR για τον οποίο προτάθηκαν τροποποιήσεις που αφορούν την προσέγγιση της περιττής πληροφορίας, τη συνάρτηση αξιολόγησης που χρησιμοποιείται καθώς και τη μέθοδο βελτιστοποίησης της συνάρτησης αυτής. Όσον αφορά στην προσέγγιση της περιττής πληροφορίας, είδαμε ότι όταν το πρόβλημα της άνισης διακύμανσης συνάφειας και περιττής πληροφορίας διορθώνεται με χρήση κανονικοποίησης (ενότητα 4.3), η προσέγγιση mean-redundancy υπερτερεί έναντι των προσεγγίσεων max-redundancy και weighted-redundancy. Ο λόγος για τον οποίο συμβαίνει αυτό είναι ένα θέμα που απαιτεί περαιτέρω μελέτη.

Η κανονικοποιημένη εκδοχή του αλγορίθμου (ενότητα 4.3) αντιμετωπίζει ως εξ ίσου σημαντικούς τους στόχου αύξησης της συνάφειας και μείωσης της περιττής πληροφορίας. Η προσέγγιση mrmr_d τείνει να επιλέγει τα πιο συναφή χαρακτηριστικά. Η προσέγγιση mrmr_q τείνει να επιλέγει λιγότερο συναφή αλλά και λιγότερο περιττά χαρακτηριστικά. Η χρήση της κανονικοποιημένης εκδοχής sMRMR_d δίνει μια ισορροπία ανάμεσα στις δύο άλλες εκδοχές και τα πειράματα δείχνουν ότι η επίτευξη αυτής της ισορροπίας αποφέρει καλύτερα αποτελέσματα.

Η χρήση μεθόδων καθολικής βελτιστοποίησης δεν προσφέρει σημαντική βελτίωση της λύσης που βρίσκεται από τον mRMR με forward άπληστη αναζήτηση. Συναρτήσεις αξιολόγησης που δίνουν μεγαλύτερη έμφαση στη μείωση της περιττής πληροφορίας μπορούν να ωφεληθούν περισσότερο από χρήση τέτοιων μεθόδων.

Ένα σημαντικό ερώτημα που αφορά στη μελλοντική εργασία, είναι πόση έμφαση πρέπει να δίνεται στη βελτιστοποίηση των δύο στόχων, αύξησης της συνάφειας και μείωσης της περιττής πληροφορίας. Τα πειράματα δείχνουν ότι άλλοτε η μέθοδος mrmr_d και άλλοτε η μέθοδος mrmr_q δίνουν καλύτερα αποτελέσματα. Η sMRMR_d επιτυγχάνει μια ισορροπία ανάμεσα στους δύο στόχους και έτσι βελτιώνει τα αποτελέσματα. Ίσως όμως ακόμα καλύτερα αποτελέσματα μπορούν να επιτευχθούν αν βρεθεί κάποιος τρόπος να καθορίζεται για το εκάστοτε σύνολο δεδομένων πόσο βάρος πρέπει να δίνεται στη βελτιστοποίηση κάθε στόχου. Αυτό θα μπορούσε να γίνει είτε με χρήση wrapper μεθοδολογίας είτε, δυσκολότερα ίσως, με χρήση στατιστικής.

Ένα άλλο σημαντικό ερώτημα είναι αν υπάρχει κάποιος τρόπος να καθορίζεται ο αριθμός χαρακτηριστικών που χρειάζεται να επιλεγούν. Η χρήση ενός συνόλου επικύρωσης για την αποτίμηση των εμφωλευμένων υποσυνόλων είναι ένας τρόπος. Είναι όμως σημαντικό να μελετηθεί πότε τα αποτελέσματα που παρατηρούνται μπορούν να θεωρούνται αξιόπιστα π.χ. ίσως ένα εμφωλευμένο υποσύνολο με πολύ καλή απόδοση δεν πρέπει θεωρείται αρκετά καλό αν το αμέσως μικρότερο και το αμέσως μεγαλύτερο εμφωλευμένο υποσύνολο έχουν κακή ή μέτρια απόδοση.

Ένα ερώτημα που παρουσιάζει ενδιαφέρον είναι αν από το συνδυασμό διαφορετικών συναρτήσεων αξιολόγησης, ανάλογων με αυτή που χρησιμοποιεί ο mRMR και το κριτήριο CMIM [13], μπορεί να προκύψει μία πιο αξιόπιστη συνάρτηση αξιολόγησης. Η ιδέα είναι ότι μία ομάδα συναρτήσεων αξιολόγησης μπορεί να δουλεύει καλά σε περιπτώσεις όπου κάποια από τις συναρτήσεις που συνιστούν το συνδυασμό αποτυγχάνει.

Τέλος, το ερώτημα για το αν μπορεί να υπάρξει καλύτερη εκτίμηση της περιττής πληροφορίας από αυτή που δίνει η προσέγγιση mean-redundancy παραμένει. Νέες παραλλαγές μπορούν να δοκιμαστούν ενώ με πειραματικό τρόπο θα μπορούσε να ελεγχθεί η ισχύς της υπόθεσης της ενότητα 5.4, ότι δηλαδή η κατάταξη των χαρακτηριστικών με βάση την προσέγγιση mean-redundancy είναι αρκετά κοντά στην κατάταξη με βάση την πραγματική περιττή πληροφορία.

ΑΝΑΦΟΡΕΣ

[1] A. A. Alizadeh, M. B. Eisen, E. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, J. C. Boldrick, H. Sabet, T. Tran, X. Yu, J. I. Powell, L. Yang, G. E. Marti, T. Moore, J. Hudson, L. Lu, D. B. Lewis, R. Tibshirani, G. Sherlock, W. C. Chan, T. C. Greiner, D. D. Weisenburger, J. O. Armitage, R. Warnke, R. Levy, W. Wilson, M. R. Grever, J. C. Byrd, D. Botstein, P. O. Brown and L. M. Staudt. "Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling". Nature, Vol. 403(6769), pp. 503-511, February 2000.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack and A. J. Levine. "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays". Proc Natl Acad Sci USA, Vol. 96(12), pp. 6745-6750, June 1999.

[3] A. Asuncion and D.J. Newman. "UCI Machine Learning Repository", 2007. [Online]. Available at http://www.ics.uci.edu/~mlearn/MLRepository.html.

[4] R. Battiti. "Using mutual information for selecting features in supervised neural net learning". In IEEE Transactions on Neural Networks, Vol. 5, 1994.

[5] C. M. Bishop. "Pattern recognition and machine learning". Springer, 2006

[6] A. Blum and P. Langley. "Selection of relevant features and examples in machine learning". Artificial Intelligence, Vol. 97(1-2), pp. 245-271, December 1997.

[7] B. Boser, I. Guyon and V. Vapnik. “A training algorithm for optimal margin classifiers”. In Proceedings of the Fifth Annual Workshop on Computational Learning Theory, pp. 144–152, 1992.

[8] C. C. Chang and C. J. Lin. “LIBSVM : a library for support vector machines”, 2001. [Online]. Available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[9] T. M. Cover and J. A. Thomas. “Elements of Information Theory”. Wiley-Interscience, 1991.

[10] C. Ding and H. Peng. “Minimum redundancy feature selection from microarray gene expression data”. Journal of Bioinformatics and Computational Biology,Vol. 3(2), pp.185-205, 2005.

[11] D. Dougherty, R. Kohavi and M. Sahami. “Supervised and unsupervised discretisation of continuous features”. In Machine Learning: Proceedings of the Twelvth International Conference, 1995.

[12] S. Dudoit, J. Fridlyand and T. Speed. “Comparison of discrimination methods for the classification of tumors using gene expression data”. Journal of the American Statistical Association, Vol. 97(457), pp. 77-87, March 2002.

[13] F. Fleuret. “Fast Binary Feature Selection with Conditional Mutual Information”. The Journal of Machine Learning Research, Vol. 5, pp. 1531-1555, 2004.

[14] M. E. Garber, O. G. Troyanskaya, K. Schluens, S. Petersen, Z. Thaesler, M. Pacyna-Gengelbach, M. van de Rijn, G. D. Rosen, C. M. Perou, R. I. Whyte, R. B. Altman, P. O. Brown, D. Botstein and I. Petersen. “Diversity of gene expression in adenocarcinoma of the lung”. Proceedings of the National Academy of Sciences of the United States of America, Vol. 98(24), pp. 13 784-13 789, November 2001.

[15] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, C. Bloomfield, and E. Lander. “Molecular

classification of cancer: Class discovery and class prediction by gene expression monitoring". Science, Vol. 286(5439), pp. 531-537, October 1999.

[16] I. Guyon and A. Elisseeff. "An introduction to variable and feature selection". Journal of Machine Learning Research, Vol. 3, pp1157-1182, March 2003.

[17] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. "Gene selection for cancer classification using support vector machines". Machine Learning, Vol. 46(1-3), pp. 389-422, 2002.

[18] M. A. Hall. "Correlation-based feature selection for discrete and numeric class machine learning". In ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning. pp. 359-366, 2000.

[19] M. A. Hall. "Correlation based feature selection for machine learning". PhD Thesis, Department of Computer Science, University of Waikato, Hamilton, New Zeland, 1998.

[20] R. Kohavi and D. Sommerfield. "Feature subset selection using the wrapper method: Overfitting and dynamic search space topology". In Proceedings of the First International Conference on Knowledge Discovery and Data Mining, AAAI Press, 1995.

[21] R. Kohavi and G. H. John. "Wrappers for feature subset selection". Artificial Intelligence, Vol. 97(1-2), pp. 273-324, December 1997.

[22] R. Kohavi. "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the 14th International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1995.

[23] D. Koller and M. Sahami. "Toward optimal feature selection". International Conference on Machine Learning, pp. 284-292, July 1996.

[24] “mRMR feature Selection Site”, 2009. [Online]. Available at http://penglab.janelia.org/proj/mRMR.

[25] P. Pavlidis, J. Weston, J. Cai and W. N. Grundy. “Gene functional classification from heterogeneous data”. Proceedings of the Fifth International Conference on Computational Molecular Biology, pp. 242-248, April 2001.

[26] H. Peng, F. Long and C. Ding. “Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy”. IEEE Trans Pattern Anal Mach Intell, vol. 27(8), pp. 1226-1238, August 2005.

[27] P. Pudil, J. Novovicova and J. Kittler. “Floating search methods in feature selection”. Pattern Recognition Letters, Vol. 15(11), pp. 1119-1125, 1994.

[28] J. Reunanen. “Overfitting in making comparisons between variable selection methods”. Journal of Machine Learning Research, Vol. 3, pp. 371-1382, 2003.

[29] P. N. Tan, M. Steinbach, V. Kumar. “Introduction to data mining”. Addison Wesley, 2005.

[30] V. N. Vapnik. “Statistical Learning Theory”. Wiley, New York, 1998.

[31] J. Weston, S. Mukherjee, O. Chapelle, M. Pontil, and V. Vapnik, “Feature selection for svms”. In Advances in Neural Information Processing Systems, Vol. 13, pp. 668-674, 2000.

[32] Z. Zhao and H. Liu. “Searching for interacting features”. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI-07), 2007.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ο Οδυσσέας Πετρόχειλος γεννήθηκε στην Ξάνθη το 1983. Το 2001 αποφοίτησε από το λύκειο και εισήχθη στο τμήμα πληροφορικής του πανεπιστημίου Ιωαννίνων από το οποίο αποφοίτησε με βαθμό «Λίαν Καλώς» το 2006. Παρακολούθησε το πρόγραμμα μεταπτυχιακών σπουδών του ίδιου τμήματος από τον Οκτώβριο του 2006 και αποφοίτησε τον Ιούλιο του 2009 αποκτώντας δίπλωμα με ειδίκευση στις «Τεχνολογίες-Εφαρμογές». Τα ερευνητικά του ενδιαφέροντα εστιάζονται στον τομέα της μηχανικής μάθησης και πιο συγκεκριμένα στα προβλήματα της ταξινόμησης και της επιλογής χαρακτηριστικών.