ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕ ΓΕΝΕΤΙΚΟΥΣ ΑΛΓΟΡΙΘΜΟΥΣ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΥΠΟΛΟΓΙΣΤΙΚΗΣ ΟΡΑΣΗΣ

Η

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

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

από τον

Σάββα Δημητριάδη

ως μέρος των Υποχρεώσεων
για τη λήψη
του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Νοέμβριος 2010

ΑΦΙΕΡΩΣΗ

Στην Έφη, την Ελένη και τη Χαρά.

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Κωνσταντίνο Μπλέκα για την πολύτιμη βοήθεια και την καθοδήγηση που μου προσέφερε κατά την εκπόνηση αυτής της εργασίας.

Ένα μεγάλο ευχαριστώ στην οικογένειά μου για την συμπαράσταση και την υπομονή που έδειξαν κατά το χρονικό διάστημα των μεταπτυχιακών μου σπουδών.

Τέλος, ευχαριστώ όλους τους δικούς μου ανθρώπους για την κατανόηση που έδειξαν.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

ΠΕΡΙΛΗΨΗ

Σάββας Δημητριάδης του Χαραλάμπους και της Ασημούλας.
MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Νοέμβριος, 2010.
Επιλογή Χαρακτηριστικών με Γενετικούς Αλγορίθμους σε προβλήματα Υπολογιστικής Όρασης.
Επιβλέπων: Κωνσταντίνος Μπλέκας.

Αντικείμενο της εργασίας είναι η μελέτη του προβλήματος της επιλογής χαρακτηριστικών (feature selection) στο πρόβλημα της αναγνώρισης προσώπων (face recognition). Αντιμετωπίζοντας το πρόβλημα ως ένα πρόβλημα ταξινόμησης εικόνων (image classification) θεωρούμε ένα σύστημα αναγνώρισης εικόνων προσώπων που περιλαμβάνει δύο στάδια, την επιλογή κατάλληλων χαρακτηριστικών διάκρισης και το στάδιο ταξινόμησης με βάση αυτά τα χαρακτηριστικά. Η επιλογή του κατάλληλου υποσυνόλου χαρακτηριστικών αποτελεί κρίσιμο ζήτημα στην κατασκευή ενός μοντέλου ταξινόμησης προτύπων ιδίως σε προβλήματα όπου τα δεδομένα είναι υψηλής διάστασης.

Στην παρούσα εργασία οι εικόνες αναπαρίστανται ως διανύσματα διάστασης ίσης με το μέγεθος της εικόνας (πλήθος των pixels) και επομένως ως διανύσματα «υψηλής διάστασης». Για την ταξινόμηση χρησιμοποιήσαμε τον ταξινομητή SVM (Support Vector Machine). Για την κατάλληλη επιλογή των χαρακτηριστικών-εικονοστοιχείων (pixels) προτείνουμε ένα υβριδικό σχήμα Γενετικού αλγορίθμου (Genetic algorithm) και Προσομοιωμένης Ανόπτησης (Simulated Annealing). Εξαιτίας της πολυπλοκότητας του προβλήματος που εξετάζουμε, λόγω της μεγάλης διάστασης του χώρου των δεδομένων, η βασική ιδέα της μεθοδολογίας είναι ο διαχωρισμός της αναζήτησης σε δύο στάδια: ένα καθολικό και ένα τοπικό στάδιο.

Διαμερίζοντας αρχικά την εικόνα σε μικρότερες περιοχές – blocks, στο πρώτο στάδιο επιλέγεται το κατάλληλο υποσύνολο των ενεργών blocks χρησιμοποιώντας τους Γενετικούς Αλγορίθμους. Στη συνέχεια σε δεύτερο στάδιο, επιτελείται μια τοπική αναζήτηση εσωτερικά σε κάθε ένα από τα ενεργά blocks για να βρεθεί το κατάλληλο υποσύνολο των pixels κάθε ενεργού block. Τα δύο παραπάνω στάδια εκτελούνται επαναληπτικά.

Για την αναπαράσταση των υποσυνόλων των χαρακτηριστικών προτείνεται μια συνεχής κωδικοποίηση όπου για κάθε χαρακτηριστικό-pixel υπάρχει μια πιθανότητα να συμμετέχει στη διαδικασία αξιολόγησης των χαρακτηριστικών.

Η πειραματική μελέτη της μεθόδου που διεξήχθη σε πραγματικά σύνολα δεδομένων (εικόνες προσώπων) υψηλής διάστασης έδειξε ικανοποιητική συμπεριφορά και σημαντικά καλύτερα αποτελέσματα από υπάρχουσες μεθοδολογίες επιλογής χαρακτηριστικών.

EXTENDED ABSTRACT IN ENGLISH

Dimitriadis, Savvas, Ch. MSc, Computer Science Department, University of Ioannina, Greece. November, 2010.
Thesis Title: Feature selection using Genetic Algorithms for problems of Computer Vision.
Thesis Supervisor: Konstantinos Blekas.

Face recognition has been a very important research problem in the field of computer vision. We consider this problem as an image classification problem. In this study the face recognition system includes two main parts. In the first one, it selects a number of features and in the second, it trains a classifier using the selected features for the correct classification among different class instances. Choosing an appropriate set of features is critical when building pattern classification systems because it can yield significant benefits such as reduced risk of overfitting and reduced computational cost.

In this thesis, a 2D face image is represented as features (pixels) vector in high dimensional image space. For the classification we used the SVM (Support Vector Machine) classifier. To find the appropriate features (pixels) we used a hybrid Genetic algorithm and Simulated Annealing. Because of the complexity of the problem we propose to search the space of features in two stages: one global and one local. Initially we divided the image into smaller regions (blocks) and then we used Genetic algorithm in order to determine interesting regions. In continuation we used local search with Simulated Annealing to find the most useful features in the blocks.

We propose a real decoding representation for the features of subsets in order for each feature-pixel to have a probability to participate in the evaluation process.

The proposed method was tested on high dimensional faces datasets. The results of the experimental research illustrate significant performance in comparison with other feature selection techniques.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή του προβλήματος

Η υπολογιστική όραση (computer vision) έχει ως αντικείμενο την προσομοίωση της απόδοσης οπτικών εργασιών τις οποίες οι άνθρωποι είναι γνωστό ότι κάνουν με πολύ επιτυχημένο τρόπο. Το πρόβλημα της αναγνώρισης προσώπων αποτελεί ένα σημαντικό θέμα στο τομέα της υπολογιστικής όρασης. Η αναγνώριση προσώπων έχει ως στόχο την ταξινόμηση ενός συγκεκριμένου προσώπου ή αλλιώς την ανίχνευση ενός προσώπου, από μια συλλογή εικόνων ή μια ακολουθία εικόνων video στη σωστή κατηγορία. Οι άνθρωποι έχουν την ικανότητα με λίγη προσπάθεια να αναγνωρίζουν ένα πλήθος αντικειμένων σε εικόνες, παρά το γεγονός ότι η εικόνα των αντικειμένων μπορεί να μην είναι και τόσο ευδιάκριτη και τα αντικείμενα να μην προβάλλονται σωστά λόγω κακών συνθηκών φωτισμού και μεταβολών στην κλίμακα κ.α. Η ικανότητα αυτή του ανθρώπου ξεκινάει από την παιδική του ηλικία και σταδιακά μέσω των διαδικασιών μάθησης του εγκεφάλου, το οπτικό σύστημα εξελίσσεται φτάνοντας σε υψηλά επίπεδα.

Οι παράγοντες που δυσχεραίνουν την επίλυση του προβλήματος της αναγνώρισης προσώπων μπορούν να συνοψιστούν στα παρακάτω σημεία:

- Ποικιλία της εμφάνισης μέσα στην ίδια κατηγορία.

- Διαφορετικές εκφράσεις του ίδιου προσώπου.
- Μεταβολή του φωτισμού.
- Αλληλοεπικαλύψεις ή απόκρυψη από άλλα αντικείμενα όπως γυαλιά, καπέλα και άλλα αξεσουάρ.
- Αλλαγές στην εμφάνιση λόγω περιστροφής του προσώπου. Αυτό μπορεί να οφείλεται είτε σε στροφή του κεφαλιού είτε σε μετακίνηση της κάμερας κατά τη λήψη της εικόνας.

Η ταξινόμηση προσώπων με βάση τα παραπάνω προβλήματα αποτελεί μια σύνθετη διαδικασία εξαιτίας της απαίτησης επεξεργασίας μεγάλου όγκου πληροφορίας, σημαντικό τμήμα της οποίας αποτελούν άχρηστα ή περιττά δεδομένα. Οι τεχνικές αναγνώρισης προσώπων χωρίζονται σε δύο γενικές κατηγορίες: τις τεχνικές χαρακτηριστικών (features based techniques) και τις καθολικές προσεγγίσεις (global approaches). Στις καθολικές προσεγγίσεις ολόκληρη η εικόνα χρησιμοποιείται ως ένα διάνυσμα χαρακτηριστικών (feature vector) διάστασης ίσης με το μέγεθος της εικόνας, ενώ αντίθετα στις τεχνικές χαρακτηριστικών γίνεται εξαγωγή των χαρακτηριστικών σημείων του προσώπου (local features) και το διάνυσμα της εικόνας αποτελείται από αυτά τα χαρακτηριστικά [1].

Σε ένα σύστημα αναγνώρισης προσώπων η διαδικασία ταξινόμησης των εικόνων θα πρέπει να βασίζεται μόνο στα χαρακτηριστικά που περιέχουν σημαντική πληροφορία για την εικόνα και είναι απαραίτητα για την κατηγοριοποίησή της.
Η επιλογή του κατάλληλου υποσυνόλου χαρακτηριστικών αποτελεί κρίσιμο ζήτημα στην κατασκευή ενός μοντέλου ταξινόμησης προτύπων ιδίως σε προβλήματα όπου τα δεδομένα είναι υψηλής διάστασης. Ένα από τα σημαντικότερα προβλήματα της μηχανικής μάθησης και της αναγνώρισης προτύπων είναι το πρόβλημα της «κατάρας της διάστασης» (curse of dimensionality), στην περίπτωση που τα δεδομένα είναι λίγα αλλά υψηλής διάστασης. Συχνά το μοντέλο ταξινόμησης που παράγεται από τη διαδικασία εκπαίδευσης ενός ταξινομητή σε ένα σύνολο δεδομένων, ταιριάζει πολύ καλά στα δεδομένα εκπαίδευσης κατατάσσοντάς τα στις σωστές κατηγορίες με υψηλό βαθμό επιτυχίας αλλά όμως η ικανότητα γενίκευσης που έχει, δηλαδή να ταξινομεί άγνωστα μέχρι στιγμής παραδείγματα δεν είναι αποτελεσματική (σφάλμα

ταξινόμησης σε άγνωστα μέχρις στιγμής δεδομένα). Το φαινόμενο αυτό ονομάζεται υπερεκπαίδευση (overfitting). Μια βασική αιτία εμφάνισης αυτού του φαινόμενου είναι η απουσία ικανοποιητικού αριθμού παραδειγμάτων εκπαίδευσης σε συνδυασμό με τον αυξημένο αριθμό χαρακτηριστικών δηλαδή τη μεγάλη διάσταση των χαρακτηριστικών. Το πρόβλημα του περιορισμένου αριθμού δεδομένων εκπαίδευσης πολλές φορές δεν μπορεί να λυθεί αφού είναι ζήτημα του πεδίου εφαρμογής του προβλήματος που εξετάζεται κάθε φορά. Το πρόβλημα της μείωση της διάστασης μπορεί να αμβλυνθεί μέσω τεχνικών επιλογής χαρακτηριστικών.

Η μείωση της διάστασης σε διάφορα προβλήματα της υπολογιστικής όρασης αντιμετωπίζεται κατά κύριο λόγο με παραδοσιακές μεθόδους εξαγωγής χαρακτηριστικών (feature extraction) όπως για παράδειγμα τη μέθοδο PCA (Principal-Component Analysis) ή τη μέθοδο ICA (Independent-Component Analysis). Αυτές οι μέθοδοι παρουσιάζουν μειονεκτήματα γιατί δεν χρησιμοποιούν κάποιο αποτελεσματικό σχήμα επιλογής των καταλληλότερων υποσυνόλων χαρακτηριστικών, αγνοώντας τη σημασία της απομάκρυνσης των περιττών και άχρηστων χαρακτηριστικών. Το πρόβλημα αυτό γίνεται ιδιαίτερα έντονο όταν υπάρχουν λίγα δεδομένα και πολλά χαρακτηριστικά [2].

Το τελευταίο διάστημα έχει αναπτυχθεί ένα ιδιαίτερο ενδιαφέρον στην ανάπτυξη και χρήση τεχνικών επιλογής χαρακτηριστικών σε εφαρμογές υπολογιστικής όρασης όπως object detection [2], face detection [3,4], gender classification [5,6], vehicle detection [7], target detection [8], image retrieval [9].

1.2 Αντικείμενο της εργασίας

Η παρούσα εργασία εστιάζει στην μείωση της διάστασης στο πρόβλημα της αναγνώρισης προσώπων (face recognition) από το πεδίο της υπολογιστικής όρασης (computer vision). Το πρόβλημα της επιλογής χαρακτηριστικών μπορεί να χαρακτηριστεί και πρόβλημα βελτιστοποίησης αφού αναζητείται το βέλτιστο υποσύνολο χαρακτηριστικών σε σχέση με ένα προκαθορισμένο κριτήριο αξιολόγησης

των επιλεγμένων χαρακτηριστικών. Εξαιτίας της πολυπλοκότητας του προβλήματος που εξετάζεται στη παρούσα εργασία λόγω της μεγάλης διάστασης του χώρου των δεδομένων, η μεθοδολογία επιλογής χαρακτηριστικών που προτείνουμε είναι να διασπάσουμε τον πολυδιάστατο αρχικό χώρο του προβλήματος σε υποχώρους μικρότερης διάστασης και στη συνέχεια με στοχαστικές μεθόδους καθολικής βελτιστοποίησης, τους Γενετικούς αλγορίθμους και τη μέθοδο της Προσομοιωμένης Ανόπτησης (Simulated Annealing), να επεξεργαστούμε παράλληλα αυτούς τους διανυσματικούς υποχώρους.

Ο συνδυασμός αυτών των μεθόδων εμφανίζει ένα συγκριτικό πλεονέκτημα σε σχέση με άλλες μεθόδους επιλογής χαρακτηριστικών σε προβλήματα υψηλής διάστασης, αφού οι Γενετικοί αλγόριθμοι πραγματοποιούν ταυτόχρονη εξερεύνηση διαφορετικών συνδυασμών των υποσυνόλων των χαρακτηριστικών, ενώ η μέθοδος της Προσομοιωμένης Ανόπτησης (Simulated Annealing) ανακαλύπτει με γρήγορο και αποδοτικό τρόπο τους βέλτιστους συνδυασμούς αυτών των υποσυνόλων. Έτσι με αυτό τον τρόπο ανακαλύπτονται χαρακτηριστικά τα οποία αλληλεπιδρούν μεταξύ τους αφού έχουν μεγάλη πιθανότητα να βρεθούν στο ίδιο υποσύνολο χαρακτηριστικών εξαιτίας του τρόπου εξερεύνησης του Γενετικού αλγορίθμου και παράλληλα να εντοπιστούν από τη μέθοδο τοπικής αναζήτησης Simulated Annealing. Το γεγονός αυτό αποτελεί ένα κρίσιμο ζήτημα υπό την έννοια ότι σημαντική πληροφορική αξία για την ταξινόμηση μπορεί να περιέχεται σε υψηλής τάξης συσχετίσεις μεταξύ των χαρακτηριστικών (pixels) των εικόνων.

1.3 Διάρθρωση της εργασίας

Στο κεφάλαιο 2 της εργασίας γίνεται μια επισκόπηση του προβλήματος της μεγάλης διάστασης και αναλύονται κάποια σημαντικά ζητήματα σχετικά με την επιλογή χαρακτηριστικών. Περιγράφονται οι στρατηγικές αναζήτησης βέλτιστων υποσυνόλων χαρακτηριστικών καθώς και οι τεχνικές αξιολόγησης των υποσυνόλων αυτών.

Στο κεφάλαιο 3 παρουσιάζονται βασικά ζητήματα σχετικά με το θεωρητικό υπόβαθρο των Γενετικών αλγορίθμων, της μεθόδου Simulated Annealing (Προσομοιωμένη Ανόπτηση) και των ταξινομητών SVM (Support Vector Machines).

Στο κεφάλαιο 4 παρουσιάζεται η προτεινόμενη μεθοδολογία επιλογής χαρακτηριστικών, με αναφορά σε διάφορες προσεγγίσεις του μοντέλου υλοποίησης και στον υβριδικό αλγόριθμο επιλογής χαρακτηριστικών.

Στη συνέχεια, στο κεφάλαιο 5, αναλύονται τα πειραματικά αποτελέσματα σχετικά με την αξιολόγηση της προτεινόμενης μεθόδου σε πραγματικά σύνολα δεδομένων εικόνων προσώπου. Επίσης γίνεται συγκριτική αξιολόγηση των αποτελεσμάτων με άλλες τεχνικές επιλογής χαρακτηριστικών.

Τέλος στο κεφάλαιο 6, παρουσιάζονται τα συμπεράσματα σχετικά με την εφαρμογή της προτεινόμενης μεθόδου για την επιλογή χαρακτηριστικών στο πρόβλημα της ταξινόμησης εικόνων.

ΚΕΦΑΛΑΙΟ 2. ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

2.1 Εισαγωγή στο πρόβλημα

Σε πολλούς τομείς της υπολογιστικής όρασης όπως η αναγνώριση προσώπων (face recognition), συναντάει κανείς το φαινόμενο, τα στιγμιότυπα των δεδομένων του προβλήματος να προσδιορίζονται από ένα τεράστιο αριθμό χαρακτηριστικών δηλαδή να είναι υψηλής διάστασης. Θα πρέπει λοιπόν πριν την εφαρμογή οποιασδήποτε άλλης διαδικασίας, να προσδιορίσουμε ακριβώς εκείνα μόνο τα χαρακτηριστικά που περιγράφουν με επάρκεια τα δεδομένα μας και αποτελούν χρήσιμη πληροφορία για την αναπαράσταση των λύσεων του προβλήματος.

Το παραπάνω πρόβλημα της μεγάλης διάστασης συνδέεται άμεσα με το πρόβλημα της «κατάρας της διαστατικότητας» (curse of dimensionality), σύμφωνα με το οποίο η αύξηση του αριθμού των χαρακτηριστικών πρέπει να συνοδεύεται από εκθετική αύξηση του αριθμού των παραδειγμάτων (στιγμιότυπα του προβλήματος).

Στον πραγματικό κόσμο πολλές φορές τα σύνολα δεδομένων περιέχουν δεκάδες χαρακτηριστικά και άρα είμαστε αναγκασμένοι να διαθέτουμε και έναν μεγάλο αριθμό παραδειγμάτων. Αυτό δεν είναι εφικτό ούτε επιθυμητό τις περισσότερες φορές να γίνει. Στην πλειοψηφία των περιπτώσεων όπου εργαζόμαστε με σύνολα δεδομένων υψηλής διάστασης, πολλά χαρακτηριστικά είναι περιττά ή άσχετα με την επίλυση του προβλήματος και προκαλούν σύγχυση στους αλγόριθμους ταξινόμησης ή ομαδοποίησης μειώνοντας την απόδοσή τους, ενώ παράλληλα προκύπτουν προβλήματα τόσο στο σχεδιασμό όσο και στη διαχείριση του χρόνου εκτέλεσης και της κατανάλωσης υπολογιστικών πόρων.

Λύση στο παραπάνω πρόβλημα προσπαθούν να δώσουν οι τεχνικές μείωσης διαστάσης (Dimensionality Reduction – DR), οι οποίες προβάλουν ένα σύνολο από διανύσματα υψηλής διάστασης του χώρου R^n σε ένα νέο χώρο χαμηλότερης διάστασης R^k (συνήθως $k << n$) διατηρώντας ταυτόχρονα τις πιο σημαντικές ιδιότητες των δεδομένων. Στη βιβλιογραφία αναφέρονται δύο βασικές τεχνικές μείωσης της διάστασης: η μείωση της διάστασης μέσω μεθόδων επιλογής χαρακτηριστικών (feature selection) και η μείωση της διάστασης μέσω μεθόδων εξαγωγής χαρακτηριστικών (feature extraction). Συγκεκριμένα:

- Η επιλογή χαρακτηριστικών (feature selection) σε προβλήματα ταξινόμησης, ορίζεται ως η διαδικασία εντοπισμού των χαρακτηριστικών τα οποία είναι απαραίτητα για τη σωστή κατάταξη των παραδειγμάτων στις κατηγορίες.
- Η εξαγωγή χαρακτηριστικών (feature extraction), ορίζεται ως η διαδικασία δημιουργίας ενός υποσυνόλου χαρακτηριστικών μέσω μετασχηματισμών του αρχικού χώρου των δεδομένων.

Στον τομέα της υπολογιστικής όρασης και ειδικότερα στην αναγνώριση αντικειμένων η πιο δημοφιλής τεχνικής εξαγωγής χαρακτηριστικών είναι η μέθοδος Ανάλυσης Κυρίων Συνιστωσών (Principal Component Analysis - PCA). Αποτελεί μια μη επιβλεπόμενη τεχνική και ανήκει στην κατηγορία των μεθόδων γραμμικού μετασχη,ματισμού δεδομένων σε ένα χώρο χαμηλότερης διάστασης. Οι Sirovich και Kirby εισήγαγαν πρώτοι τη μέθοδο PCA στο πεδίο της αναγνώρισης εικόνων [10]. Οι

Turk και Pentland καθιέρωσαν την έννοια των eigenfaces στον τομέα της αναγνώρισης προσώπων [11]. Η μέθοδος PCA δεν λαμβάνει υπόψη σχέσεις εξάρτησης μεταξύ τριών ή περισσότερων pixels και το γεγονός αυτό δημιουργεί πρόβλημα, αφού σημαντική πληροφορική αξία μπορεί να περιέχεται σε υψηλής τάξης συσχετίσεις μεταξύ των χαρακτηριστικών (pixels). Το μειονέκτημα αυτό καλύπτει η μέθοδος KPCA (Kernel Principal Component Analysis) η οποία αποτελεί επέκταση της μεθόδου PCA. Μετασχηματίζει τον αρχικό χώρο με μη γραμμικό τρόπο στο χώρο των χαρακτηριστικών και προσδιορίζει τα κύρια συστατικά του χώρου αυτού. Επίσης άλλες τεχνικές εξαγωγής χαρακτηριστικών, όπως η Ανάλυση Γραμμικού Διαχωρισμού (linear discriminant analysis-LDA) και ICA (independent components analysis),χρησιμοποιούνται σε προβλήματα αναγνώρισης αντικειμένων με στόχο τη μείωση της διάστασης των δεδομένων [12,13].

Οι παραπάνω μέθοδοι σε γενικές γραμμές δεν χρησιμοποιούν κάποιο αποτελεσματικό σχήμα επιλογής ενός κατάλληλου υποσυνόλου χαρακτηριστικών και λειτουργούν περισσότερο σαν μέθοδοι ταξινόμησης και όχι σαν μέθοδοι που απορρίπτουν περιττά και άχρηστα χαρακτηριστικά. Ειδικότερα, παρουσιάζουν προβληματική συμπεριφορά όταν τα διαθέσιμα δεδομένα είναι περιορισμένα και ο αριθμός των χαρακτηριστικών αρκετά μεγάλος. Για το σκοπό αυτό συνήθως χρησιμοποιούνται σε συνδυασμό με κάποιες άλλες τεχνικές επιλογής χαρακτηριστικών όπως για παράδειγμα τους Γενετικούς Αλγορίθμους [2].

2.2 Επιλογή χαρακτηριστικών

Το πρόβλημα της επιλογής χαρακτηριστικών από μια άποψη μπορεί να χαρακτηριστεί και πρόβλημα βελτιστοποίησης αφού αναζητείται το βέλτιστο υποσύνολο χαρακτηριστικών σε σχέση με κάποιο κριτήριο αξιολόγησης, όπως η απόδοση ενός ταξινομητή ή μια άλλη συνάρτηση αξιολόγησης.
Αναλυτικότερα το πρόβλημα μπορεί να οριστεί ως εξής:

Δεδομένου ενός συνόλου διανυσμάτων $X_d = \{x_i \mid i = 1...d\}$ επέλεξε εκείνο το υποσύνολο $Y_m = \{x_{i1}, x_{i2}, ..., x_{im}\}$ με $m << d$ το οποίο βελτιστοποιεί μια συνάρτηση αξιολόγησης J.

$$\begin{bmatrix} x_1 \\ x_2 \\ . \\ . \\ x_d \end{bmatrix} \rightarrow \begin{bmatrix} x_{i1} \\ x_{i2} \\ . \\ . \\ x_{im} \end{bmatrix} \quad \{x_{i1}, x_{i2}, ..., x_{im}\} = \arg\max_{m, J_m}[J\{x_i \mid i = 1...d\}]$$

Όλα τα δυνατά υποσύνολα χαρακτηριστικών ορίζουν τον χώρο των υποψήφιων λύσεων. Με δεδομένο ότι ο αριθμός των υποσυνόλων αυτών αυξάνεται εκθετικά με το πλήθος των χαρακτηριστικών N και γίνεται ίσος με 2^N κάνει το πρόβλημα της αποτίμησης όλων αυτών των συνόλων πρακτικά ασύμφορο και σε μερικές περιπτώσεις αδύνατον να λυθεί. Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιούνται διάφορες τεχνικές αναζήτησης που βασίζονται σε διάφορους αλγόριθμους αναζήτησης ντετερμινιστικούς ή στοχαστικούς αντίστοιχα.

Η επιλογή ενός βέλτιστου υποσυνόλου χαρακτηριστικών απαιτεί δύο βασικά στοιχεία που πρέπει να καθοριστούν:

- τον τρόπο με τον οποίο θα γίνει η αναζήτηση των υποψηφίων λύσεων
- την αντικειμενική συνάρτηση αξιολόγησης βάση της οποίας γίνεται η επιλογή των βέλτιστων υποσυνόλων.

Όσο αφορά τον τρόπο αναζήτησης των υποψηφίων λύσεων υπάρχουν οι ακόλουθες 3 στρατηγικές αναζήτησης:

Exhaustive search (εξαντλητική αναζήτηση): Η εξαντλητική αναζήτηση είναι κύριο χαρακτηριστικό των αλγορίθμων αυτής της κατηγορίας και για το λόγο αυτό είναι απαγορευτική η χρήση τους στην επιλογή χαρακτηριστικών σε δεδομένα υψηλής διάστασης αφού αποτιμούν ένα πλήθος υποσυνόλων που αυξάνεται εκθετικά με τη διάσταση.

Heuristic (Ευρετικές): Χτίζουν προσθετικά ή αφαιρετικά τα υποψήφια υποσύνολα με σειριακό τρόπο (προσθέτουν ή αφαιρούν χαρακτηριστικά σειριακά) αλλά τείνουν να παγιδεύονται σε τοπικά ελάχιστα.

Randomized (στοχαστικές): Επιστρατεύουν το τυχαίο ψάξιμο τους για να αποφύγουν την παγίδευση σε τοπικά ελάχιστα. Εφαρμόζουν τυχαία δειγματοληψία ενός προεπιλεγμένου πλήθους Ν τυχαίων σημείων ομοιόμορφα κατανεμημένων στον χώρο αναζήτησης και αναζητούν το καλύτερο.

Όσο αφορά τη μεθοδολογία αξιολόγησης για την επιλογή των βέλτιστων υποσυνόλων των χαρακτηριστικών οι μέθοδοι επιλογής χαρακτηριστικών χωρίζονται σε τρεις κατηγορίες:
Filter: Η μέθοδος των φίλτρων πραγματοποιεί ανεξάρτητη αποτίμηση των υποσυνόλων των χαρακτηριστικών βασιζόμενη σε διάφορα στατιστικά μέτρα, όπως για παράδειγμα το μέτρο της αμοιβαίας πληροφορίας (mutual information).

Wrapper: Στην κατηγορία αυτή εντάσσονται όλοι οι αλγόριθμοι επιλογής χαρακτηριστικών που χρησιμοποιούν την ακρίβεια ταξινόμησης ως κριτήριο αξιολόγησης των υποσυνόλων.

Embedded: Στην κατηγορία αυτή συναντούμε μεθόδους που έχουν σχεδιαστεί με στόχο να δουλεύουν σε συνεργασία με ένα ταξινομητή συγκεκριμένου τύπου. Σε αντίθεση με τους wrappers που απλώς χρησιμοποιούν την έξοδο ενός ταξινομητή, οι μέθοδοι αυτές επιλέγουν χαρακτηριστικά ενσωματώνοντας περιορισμούς στο μοντέλο του ταξινομητή.

Στη ενότητα 2.3 γίνεται παρουσίαση κάποιων γνωστών μεθόδων αναζήτησης που βασίζονται στις στρατηγικές αναζήτησης που αναφέρθηκαν πιο πάνω.

2.3 Τεχνικές Αναζήτησης

Ο αλγόριθμος Forward selection είναι ένας απλός σειριακός greedy αλγόριθμος, ο οποίος ξεκινάει με ένα κενό υποσύνολο χαρακτηριστικών. Σε κάθε βήμα εξετάζονται όλα τα υποσύνολα που προκύπτουν από την προσθήκη ενός χαρακτηριστικού στο τρέχον υποσύνολο. Το χαρακτηριστικό που βελτιώνει σε συνδυασμό με τα ήδη υπάρχοντα χαρακτηριστικά την απόδοση του υποσυνόλου σύμφωνα πάντα με το κριτήριο αξιολόγησης προστίθεται στο υποσύνολο. Η επέκταση των υποσυνόλων μπορεί να σταματήσει με βάση κάποια συνθήκη τερματισμού εκτελώντας σταθερό αριθμό βημάτων ή με δυναμικό τρόπο όπως για παράδειγμα αν κανένα από τα παραγόμενα υποσύνολα δεν οδήγησε σε καλύτερη απόδοση. Ο αλγόριθμος forward selection έχει καλύτερη απόδοση όταν είναι περιορισμένο το πλήθος των βέλτιστων χαρακτηριστικών. Το βασικότερο μειονέκτημα του είναι ότι δεν μπορεί σε επόμενο βήμα να αφαιρέσει χαρακτηριστικά που έχουν ήδη ενσωματωθεί σε κάποιο υποσύνολο. Επιπλέον κινδυνεύει να παγιδευτεί σε τοπικά ελάχιστα.

Ο αλγόριθμος Backward elimination ακολουθεί παρόμοια προσέγγιση με τον προηγούμενο αλγόριθμο με τη μόνη διαφορά όμως ότι το αρχικό υποσύνολο περιέχει όλα τα χαρακτηριστικά και η διαδικασία επιλογής είναι αφαιρετική. Σε κάθε επανάληψη απομιμούνται όλα τα υποσύνολα που προκύπτουν με τη διαγραφή ενός χαρακτηριστικού από το τρέχον υποσύνολο. Τελικά διαγράφεται αυτό του οποίου η απουσία οδηγεί στη μεγαλύτερη απόδοση ως προς το κριτήριο αξιολόγησης. Ο αλγόριθμος έχει καλή απόδοση όταν το βέλτιστο υποσύνολο έχει πολλά χαρακτηριστικά, καθώς εξετάζει αρχικά τα μεγάλα υποσύνολα. Το βασικότερο μειονέκτημα του είναι ότι δεν μπορεί σε επόμενο βήμα να επανεκτιμήσει χαρακτηριστικά που έχουν ήδη απορριφθεί. Επιπλέον και αυτός κινδυνεύει να παγιδευτεί σε τοπικά ελάχιστα.

Ο αλγόριθμος Plus l – take away r αποτελεί μια προέκταση των forward selection και backward elimination. Σε μια τυπική επανάληψη του αλγορίθμου, ξεκινώντας από υποσύνολο μεγέθους N, προστίθενται L χαρακτηριστικά. 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. Τελικά γίνονται R διαγραφές, πριν η μέθοδος συνεχίσει και πάλι με την προσθήκη L χαρακτηριστικών. Η ιδέα είναι να αποφευχθούν τα προβλήματα των προηγούμενων αλγορίθμων forward selection και backward elimination οι οποίες δεν μπορούν να αφαιρέσουν ή να προσθέσουν αντίστοιχα χαρακτηριστικά που έχουν ληφθεί υπόψη στη δημιουργία των υποσυνόλων σε προηγούμενες επαναλήψεις. Το σκεπτικό είναι να λαμβάνονται υπόψη συσχετίσεις ανάμεσα στα χαρακτηριστικά που ενσωματώνονται ή αφαιρούνται από τα υποσύνολα σε επόμενες επαναλήψεις. Το βασικό μειονέκτημά του είναι η αρχικοποίηση των L και R σε βέλτιστες τιμές.

Ο αλγόριθμος Best first search κρατά όλες τις καταστάσεις στο χώρο αναζήτησης. Συγκεκριμένα όλες οι καταστάσεις (υποσύνολα χαρακτηριστικών) που προκύπτουν αξιολογούνται και τοποθετούνται σε μια λίστα. Στη συνέχεια σε κάθε βήμα εντοπίζεται η καλύτερη σύμφωνα με τη συνάρτηση αξιολόγησης και αυτή είναι που θα επεκταθεί. Οι καταστάσεις που θα προκύψουν από αυτήν την κατάσταση αξιολογούνται και αυτές με τη σειρά τους και μπαίνουν στη λίστα. της αξιολόγησής τους. Η διαδικασία τερματίζει αν μετά από N διαδοχικές επεκτάσεις δεν προκύψει κάποια κατάσταση που να βελτιώνει την απόδοση. Η μέθοδος αυτή έχει μεγάλη πολυπλοκότητα.

2.4 Μέθοδοι Αξιολόγησης Wrapper

Η φιλοσοφία της μεθοδολογίας wrapper απαιτεί το συνδυασμό ενός ταξινομητή με μια μέθοδο αναζήτησης βέλτιστων χαρακτηριστικών. Ο ταξινομητής αποτιμά κάθε φορά το υποέξεταση υποσύνολο χαρακτηριστικών και επιστρέφει την απόδοσή του σε αυτό το υποσύνολο. Η αξιολόγηση ενός υποψήφιου υποσυνόλου χαρακτηριστικών, έστω Μ, 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 Μ. 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 [14].

Όταν τα διαθέσιμα παραδείγματα είναι λίγα, χρησιμοποιείται συνήθως η τεχνική cross validation (διασταυρωμένη επικύρωση). Η μέθοδος της διασταυρωμένης επικύρωσης k-πτυχών (k-fold cross–validation) ακολουθεί τα παρακάτω βήματα :

- Διάσπαση των δεδομένων σε k ισομερή ξένα υποσύνολα.
- Επιλογή του $1^{ου}$ υποσυνόλου ως σύνολο επικύρωσης και των υπόλοιπων k-1 υποσυνόλων για την εκπαίδευση του ταξινομητή.
- Δημιουργία μοντέλου ταξινόμησης και εκτίμηση του μέτρου σφάλματος.
- Επιλογή του επόμενου υποσυνόλου για επικύρωση και χρήση των υπόλοιπων ως δεδομένα εκπαίδευσης.
- Επανάληψη της διαδικασίας μέχρι κάθε υποσύνολο να χρησιμοποιηθεί για επικύρωση.
- Υπολογισμός του μέσου όρου της ακρίβειας ταξινόμησης σε κάθε επανάληψη της διαδικασίας.

Ο μέσος όρος της ακρίβειας ταξινόμησης στα διαφορετικά σύνολα επικύρωσης είναι και το κριτήριο αξιολόγησης. Η αξιολόγηση των υποσυνόλων των χαρακτηριστικών με βάση την απόδοση του ταξινομητή απαιτεί την κατασκευή ενός μοντέλου ταξινόμησης για κάθε υποσύνολο χαρακτηριστικών που εξετάζεται και αυτό έχει ως συνέπεια το αυξημένο υπολογιστικό κόστος. Το μειονέκτημα αυτό μπορεί να ξεπεραστεί με τη χρήση τεχνολογιών παράλληλης επεξεργασίας.

2.5 Μέθοδοι Αξιολόγησης με Φίλτρα

Η μέθοδος των φίλτρων πραγματοποιεί ανεξάρτητη αποτίμηση των υποσυνόλων των χαρακτηριστικών βασιζόμενη σε διάφορα στατιστικά μέτρα. Στη γενική περίπτωση κάθε χαρακτηριστικό αξιολογείται με βάση τη συσχέτισή του με τις κατηγορίες. Όσο μεγαλύτερη συσχέτιση υπάρχει, τόσο πιο χρήσιμο θεωρείται το χαρακτηριστικό.

Αυτή η προσέγγιση παρουσιάζει κάποια μειονεκτήματα με κυριότερο την αποτυχία εντοπισμού συσχετίσεων ανάμεσα στα χαρακτηριστικά αφού κάθε χαρακτηριστικό αξιολογείται ξεχωριστά από τα υπόλοιπα. Για το σκοπό αυτό έχουν αναπτυχθεί και μέθοδοι οι οποίες αξιολογούν τα χαρακτηριστικά λαμβάνοντας υπόψη και την παρουσία άλλων χαρακτηριστικών στο ίδιο υποσύνολο [17].

2.5.1 Αμοιβαία Πληροφορία (mutual information)

Ένα μέτρο της εξάρτησης μεταξύ δύο τυχαίων μεταβλητών είναι η αμοιβαία πληροφορία (mutual information). Η αμοιβαία πληροφορία σχετίζεται με την εντροπία της πληροφορίας που είναι μέτρο της αβεβαιότητας για την τιμή μιας τυχαίας μεταβλητής.

Έστω ότι έχουμε:

- k διαφορετικές τιμές για μια κατηγορία Y
- pi την πιθανότητα εμφάνισης της τιμής i για την κατηγορία Y, τότε η εντροπία της πληροφορίας για την κατηγορία S υπολογίζεται από τον τύπο:

$$H(Y) = -\sum_{i=1}^{k} p_i \log_2(p_i)$$

Η πρόβλεψη της κατηγορίας χωρίς την γνώση κάποιου χαρακτηριστικού γίνεται καθαρά στην τύχη. Η εντροπία εκφράζει την αβεβαιότητα που υπάρχει στην πρόβλεψη αυτή. Ελάχιστη τιμή εντροπίας ίση με 0 φανερώνει με μέγιστη βεβαιότητα την εμφάνιση της τιμής i της κατηγορίας Y . Η μέγιστη τιμή εντροπίας αντίθετα, φανερώνει τη μέγιστη αβεβαιότητα για την πρόβλεψη της τιμής μιας κατηγορίας. Η μέγιστη τιμή εντροπίας επιτυγχάνεται, όταν οι πιθανότητες εμφάνισης όλων των τιμών για την κατηγορία S είναι ίσες δηλαδή $p_i=1/k$.

Εάν τώρα γνωρίζουμε ότι για ένα παράδειγμα που θέλουμε να ταξινομήσουμε και τις τιμές που λαμβάνει το χαρακτηριστικό X τότε η εντροπία είναι:

$$H(Y \mid X) = -\sum_{j=1}^{m} p(x_j) H(Y \mid X = x_j)$$

Όπου $x \in Values(X)$

$$I(X;Y) = H(Y) - H(Y \mid X)$$

Με λίγα λόγια, η διαφορά της πληροφορίας για την τιμή της κατηγορίας Y πριν τη γνώση των τιμών του χαρακτηριστικού X και της πληροφορίας για την κατηγορία μετά τη γνώση των τιμών του X φανερώνει το μέγεθος της χρησιμότητας του χαρακτηριστικού για την κατηγορία. Όσο μεγαλύτερη είναι η διαφορά αυτή τόσο πιο χρήσιμο είναι το χαρακτηριστικό αυτό. Όταν μειώνεται η πληροφοριακή εντροπία, αυξάνεται η πυκνότητα πληροφορίας και άρα η περιγραφή γίνεται περισσότερο συμπαγής. Από όλους τους δυνατούς διαχωρισμούς για όλες τις ιδιότητες του συνόλου S επιλέγεται αυτός που δίνει το μεγαλύτερο κέρδος πληροφορίας.

2.5.2 Αλγόριθμος RELIEF-F

Ο αλγόριθμος RELIEF-F [15] αποτελεί μια εξέλιξη του αλγόριθμου RELIEF που εφάρμοσαν αρχικά οι Kira και Rendell. Ο αλγόριθμος RELIEF-F ανήκει στην κατηγορία των φίλτρων εξετάζοντας συσχετίσεις των τιμών των μεταβλητών ανάμεσα σε διαφορετικά στιγμιότυπα του προβλήματος. Μπορεί να διαχειριστεί και περιπτώσεις multi-class προβλημάτων επιλογής χαρακτηριστικών.
Τα βήματα του αλγορίθμου περιγράφονται ως εξής:

Algorithm ReliefF
Input: for each training instance a vector of attribute values and the class value
Output: the vector W of estimations of the qualities of attributes

1. set all weights $W[A] := 0.0$;
2. **for** i := 1 **to** m **do begin**
3. randomly select an instance R_i;
4. find k nearest hits H_j;
5. **for** each class $C \neq class(R_i)$ **do**
6. from class C find k nearest misses $M_j(C)$;
7. **for** $A := 1$ **to** a **do**
8. $W[A] := W[A] - \sum_{j=1}^{k} \text{diff}(A, R_i, H_j)/(m \cdot k) +$
9. $\sum_{C \neq class(R_i)} [\frac{P(C)}{1-P(class(R_i))} \sum_{j=1}^{k} \text{diff}(A, R_i, M_j(C))]/(m \cdot k)$;
10. **end**;

Η κύρια ιδέα του αλγορίθμου Relief εντοπίζεται στον υπολογισμό της διαχωριστικής ικανότητας των χαρακτηριστικών βάση των τιμών τους σε κοντινά στιγμιότυπα του συνόλου εκπαίδευσης. Το σκεπτικό είναι ότι ένα χαρακτηριστικό με μεγάλη διακριτική ικανότητα, πρέπει να λαμβάνει διαφορετικές τιμές μεταξύ παραδειγμάτων που ανήκουν σε διαφορετικές κατηγορίες και πρέπει να έχει την ίδια τιμή για παραδείγματα της ίδιας κατηγορίας.

Για ένα τυχαία επιλεγμένο στιγμιότυπο Ri, η μέθοδος ψάχνει τους κοντινότερους γείτονές του από τα στοιχεία του συνόλου εκπαίδευσης ως εξής:

- k παραδείγματα του συνόλου εκπαίδευσης που είναι κοντά στο Ri και ανήκουν στην ίδια κατηγορία με αυτό. Τα k αυτά παραδείγματα ονομάζονται κοντινότερες επιτυχίες (nearest hits) και,
- k παραδείγματα από διαφορετική κατηγορία τα οποία βρίσκονται σε ελάχιστη απόσταση από το Ri και ονομάζονται κοντινότερες αποτυχίες (nearest miss).

Τα m και k είναι παράμετροι που καθορίζονται από το χρήστη. Με m συμβολίζουμε το πόσες φορές θα αντλήσουμε παραδείγματα από ένα σύνολο εκπαίδευσης (ένα παράδειγμα κάθε φορά) και με k τον αριθμό τον αριθμό των κοντινότερων γειτόνων του επιλεγμένου παραδείγματος Ri. Με a συμβολίζουμε τον αριθμό των χαρακτηριστικών.

Η κύρια φιλοσοφία του αλγορίθμου συμπυκνώνεται στις γραμμές 8-9 και είναι η εξής: Αν το παράδειγμα R_i και τα k κοντινότερα παραδείγματα $H_j \in class(R_i)$ μοιράζονται διαφορετικές τιμές για ένα χαρακτηριστικό A τότε το χαρακτηριστικό A διαχωρίζει παραδείγματα που ανήκουν στην ίδια κλάση και αυτό δεν είναι επιθυμητό από τον αλγόριθμο και έτσι μειώνεται η ποσότητα W[A].

Αντίθετα αν το παράδειγμα R_i και τα τα k κοντινότερα παραδείγματα $M_j(C), \acute{o}\pi o\upsilon\, C \notin class(R_i)$ μοιράζονται διαφορετικές τιμές για ένα χαρακτηριστικό A τότε το χαρακτηριστικό A διαχωρίζει παραδείγματα που ανήκουν σε διαφορετικές κατηγορίες και αυτό είναι επιθυμητό από τον αλγόριθμο και έτσι μεταβάλλεται θετικά η ποσότητα W[A]. Επομένως, μία μεγάλη τιμή W[A] σηματοδοτεί ένα χαρακτηριστικό με καλή διακριτική ικανότητα.

Η συνάρτηση diff (γραμμές 8-9) ορίζει τη διαφορά στις τιμές ενός χαρακτηριστικού Α ανάμεσα σε δύο παραδείγματα I1 και I2. Αναλυτικότερα:

Για ονομαστικές μεταβλητές έχει τη μορφή:

$$diff(A, I_1, I_2) = \begin{cases} 0; value(A, I_1) = value(A, I_2) \\ 1; \quad διαφορετικά \end{cases}$$

Για αριθμητικές μεταβλητές έχει τη μορφή:

$$diff(A, I_1, I_2) = \frac{|value(A, I_1) - value(A, I_2)|}{\max(A) - \min(A)}$$

Επίσης η συνάρτηση diff χρησιμοποιείται επίσης για τον εύρεση των k κοντινότερων παραδειγμάτων του επιλεγμένου παραδείγματος R_i. Η συνολική απόσταση ενός παραδείγματος I από το R_i υπολογίζεται ως το άθροισμα των αποστάσεων για όλες τα χαρακτηριστικά (μεταβλητές).

$$dist(R_i, I) = \sum_{A=1}^{a} diff(A, R_i, I)$$

2.6 Embedded Μέθοδοι Αξιολόγησης

Στην κατηγορία αυτή συναντούμε μεθόδους παρόμοιας περίπου φιλοσοφίας με τους wrapper. Η διαφορά τους έχει να κάνει με τον τρόπο αξιολόγησης των χαρακτηριστικών. Ενώ οι wrapper απλώς χρησιμοποιούν την έξοδο ενός ταξινομητή δηλαδή την απόδοσή του, οι embedded μέθοδοι αξιολογούν με βάση το πως επηρεάζονται κάποιες παράμετροι που εμπλέκονται στη διαδικασία εκπαίδευσης του ταξινομητή. Χαρακτηριστικό παράδειγμα embedded μεθόδου αποτελεί ο αλγόριθμος SVM-RFE (Recursive Feature Elimination using SVM classifiers – Αναδρομική Εξάλειψη Χαρακτηριστικών με χρήση ταξινομητών SVM) ο οποίος εκτελεί μια backward elimination διαδικασία (αναδρομική διαδικασία εξάλειψης), 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 [16,17]. Το χαρακτηριστικό απορρίπτεται και ο ταξινομητής εκπαιδεύεται λαμβάνοντας υπόψη τα εναπομείναντα χαρακτηριστικά. Με τον ίδιο τρόπο το λιγότερο σημαντικό χαρακτηριστικό διαγράφεται και η διαδικασία συνεχίζει ωσότου απομείνει ένας προκαθορισμένος αριθμός χαρακτηριστικών.

ΚΕΦΑΛΑΙΟ 3. ΜΕΘΟΔΟΙ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ ΣΤΗΝ ΕΠΙΛΟΓΗ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

3.1 Εισαγωγή

Το πρόβλημα της επιλογής χαρακτηριστικών όπως αναφέρθηκε στο Κεφάλαιο 2 μπορεί να χαρακτηριστεί και πρόβλημα βελτιστοποίησης αφού αναζητείται το βέλτιστο υποσύνολο χαρακτηριστικών σε σχέση με ένα προκαθορισμένο κριτήριο αξιολόγησης των επιλεγμένων χαρακτηριστικών. Εξαιτίας της πολυπλοκότητας του προβλήματος που εξετάζεται στη παρούσα εργασία λόγω της μεγάλης διάστασης του χώρου των δεδομένων, η μεθοδολογία που προτείνουμε βασίζεται σε δύο στοχαστικές μεθόδους καθολικής βελτιστοποίησης, τους Γενετικούς αλγορίθμους (Genetic algorithms) και τη μέθοδο της Προσομοιωμένης Ανόπτησης (Simulated Annealing).

Ο συνδυασμός αυτών των μεθόδων επιλέχθηκε γιατί οι Γενετικοί αλγόριθμοι παρέχουν το πλεονέκτημα της ταυτόχρονης εξερεύνησης διαφορετικών συνδυασμών

των υποσυνόλων των χαρακτηριστικών, ενώ η μέθοδος της Προσομοιωμένης Ανόπτησης (Simulated Annealing) μας βοηθάει να ανακαλύψουμε με γρήγορο και αποδοτικό τρόπο τα βέλτιστα χαρακτηριστικά αυτών των υποσυνόλων. Έτσι, με αυτό τον τρόπο χαρακτηριστικά τα οποία αλληλεπιδρούν μεταξύ τους έχουν μια μεγάλη πιθανότητα να βρεθούν στο ίδιο υποσύνολο χαρακτηριστικών εξαιτίας του τρόπου εξερεύνησης των Γενετικών αλγορίθμων και παράλληλα να εντοπιστούν από τη μέθοδο τοπικής αναζήτησης Simulated Annealing.

Με βάση τη wrapper μεθοδολογία η οποία αναλύθηκε στο Κεφάλαιο 2 απαιτείται και η χρήσης ενός ταξινομητή μαζί με τις παραπάνω μεθόδους αναζήτησης. Ο ταξινομητής που επιλέχθηκε είναι ο SVM (Support Vector Machines) ο οποίος αποτελεί μια από τις πιο αξιόπιστες μεθόδους για δύσκολα προβλήματα ταξινόμησης. Στη συνέχεια του κεφαλαίου γίνεται αναφορά στο θεωρητικό υπόβαθρο των μεθόδων της προτεινόμενης wrapper μεθοδολογίας ξεκινώντας από τους Γενετικούς αλγόριθμους.

3.2 Γενετικοί Αλγόριθμοι

3.2.1 Εισαγωγή-Βιολογική προσέγγιση

Οι Γενετικοί αλγόριθμοι (ΓΑ) είναι μέλη της κατηγορίας των εξελικτικών αλγορίθμων. Οι εξελικτικοί αλγόριθμοι βασίζονται στην αρχή της φυσικής εξέλιξης δηλαδή της επιβίωσης των ισχυρότερων. Σε αντίθεση με τις κλασικές μεθόδους αναζήτησης, δε χρησιμοποιούν μοναδικό σημείο αναζήτησης (single-point search), αλλά ένα πληθυσμό σημείων που ονομάζονται άτομα (individuals). Κάθε άτομο αντιπροσωπεύει μία πιθανή λύση για το σχεδιαστικό πρόβλημα. Στους αλγόριθμους αυτούς, ο πληθυσμός εξελίσσεται προς συνεχώς καλύτερες περιοχές του χώρου αναζήτησης, χρησιμοποιώντας τεχνικές όπως είναι η επιλογή, η διασταύρωση και η μετάλλαξη.

Η μοντελοποίηση των Γενετικών αλγόριθμων είναι εμπνευσμένη από τη θεωρία της εξέλιξης των ειδών και βασίζεται στους μηχανισμούς της βιολογικής αναπαραγωγής που συναντώνται στη φύση. Η θεωρία αυτή όπως διατυπώθηκε και αναπτύχθηκε από το Δαρβίνο κατάφερε μετά από την πάροδο αρκετών χρόνων να γίνει αποδεκτή δίνοντας μια νέα γνώση στον τρόπο με τον οποίο εξελίσσονται οι ζωντανοί οργανισμοί. Ο Κάρολος Ροβέρτος Δαρβίνος (1809-1882) ήταν Βρετανός φυσιοδίφης, συλλέκτης και γεωλόγος, ο οποίος έμεινε διάσημος στην ιστορία, ως ο θεμελιωτής της θεωρίας της εξέλιξης, καθώς και ως εισηγητής του μηχανισμού της φυσικής επιλογής (natural selection). Σε ένα συνεχώς μεταβαλλόμενο περιβάλλον με δυσκολίες, εμπόδια και περιορισμένους πόρους οι ζωντανοί οργανισμοί καλούνται να επιζήσουν και να πολλαπλασιαστούν. Έχοντας ως μοναδικό στόχο την επιβίωση και την αναπαραγωγή οι οργανισμοί εξελίσσονται τόσο σε ατομικό αλλά και συνεργατικό επίπεδο. Σύμφωνα με τη θεωρία του Δαρβίνου, η εξέλιξη είναι αποτέλεσμα του αγώνα των ειδών για επιβίωση.

Η μετέπειτα διατύπωση από τον Mendel των νόμων της κληρονομικότητας, της ανακάλυψης της έννοιας των γονιδίων τα οποία ναι μεν διατηρούν τα χαρακτηριστικά των προγόνων τους, αλλά έχουν τη δυνατότητα να μεταλλάσσονται, έδωσε την δυνατότητα να καταλάβουμε καλύτερα τη αρχική θεωρία του Δαρβίνου. Ο Γκρέγκορ Γιόχαν Μέντελ (1822-1884) ήταν ένας Αυστριακός μοναχός, γνωστός για τις μελέτες που πραγματοποίησε σχετικά με τους μηχανισμούς της κληρονομικότητας χαρακτηριστικών στα φυτά. Συχνά αναφέρεται και ως ο «πατέρας της Γενετικής», λόγω της σημασίας που είχαν οι νόμοι της Μεντελικής κληρονομικότητας για τη μελέτη της κληρονομικότητας στα υπόλοιπα είδη, συμπεριλαμβανομένου και του ανθρώπου.

Οι ΓΑ βασίζουν τη λειτουργία τους στην παραπάνω εξελικτική διαδικασία και στηρίζουν την εύρεση λύσεων στην επιβίωση του ικανότερου, των χρωμοσωμικών ανακατατάξεων και των γονιδιακών αλλαγών. Κάθε μεταβολή που προκαλείται σε ένα πληθυσμό προσαρμόζεται καλύτερα ή χειρότερα στις συνθήκες του περιβάλλοντος. Αν προσαρμόζεται καλύτερα τότε οι πρόγονοι εξαφανίζονται και τη

θέση τους παίρνουν οι επίγονοι. Αν η μεταβολή προσαρμόζεται χειρότερα, τότε οδηγεί σε θάνατο (απόρριψη).

Η ορολογία για την περιγραφή των δομικών τους στοιχείων είναι δανεισμένη από το χώρο της γενετικής. Συγκεκριμένα οι ΓΑ αναφέρονται στην έννοια του πληθυσμού ο οποίος αποτελείται από άτομα (individuals) ή γενότυπους. Κάθε άτομο αποτελείται από χρωμοσώματα. Στους ΓΑ αλγόριθμους τα άτομα αποτελούνται από ένα χρωμόσωμα και οι δύο έννοιες συνήθως ταυτίζονται Τα άτομα αποτελούν τις υποψήφιες λύσεις για το πρόβλημα που εξετάζεται. Η δομή των ατόμων περιγράφεται σαν μια διατεταγμένη γραμμική ακολουθία γονιδίων τα οποία ορίζουν τα γενετικά χαρακτηριστικά. Μέσω των γονιδίων γίνεται η μεταβίβαση των χαρακτηριστικών ενός ατόμου στους απογόνους του. Για την αναπαράσταση της πληροφορίας των ατόμων χρησιμοποιείται κατά κύρια βάση η δυαδική αναπαράσταση. Το χρωμόσωμα αποτελεί ουσιαστικά μια υποψήφια λύση του προβλήματος και κωδικοποιείται τις περισσότερες φορές ως μια ακολουθία από δυαδικά ψηφία. Τα γονίδια αντιστοιχούν είτε σε μεμονωμένα δυαδικά ψηφία είτε σε μια ομάδα δυαδικών ψηφίων αν ορίζεται έτσι η κωδικοποίηση των μεταβλητών του προβλήματος. Η απεικόνιση της υποψήφιας λύσης, γίνεται μέσω του φαινότυπου ο οποίος αποτελεί την πραγματική αναπαράσταση της υποψήφιας λύσης στο χώρο αναζήτησης και προκύπτει μέσω της αποκωδικοποίησης της δυαδικής συμβολοσειράς του ατόμου. Ο φαινότυπος ουσιαστικά δείχνει αυτό που βλέπουμε ως αποτέλεσμα και με βάση την τιμή του γίνεται η αξιολόγηση των υποψήφιων λύσεων.

3.2.2 Επιλογή χαρακτηριστικών και Γενετικοί Αλγόριθμοι

Οι ΓΑ αποτελούν μια στοχαστική μέθοδο καθολικής βελτιστοποίησης με δυνατότητες παράλληλης επεξεργασίας ενός πληθυσμού υποψήφιων λύσεων (σημεία αναζήτησης). Χρησιμοποιούν πιθανοθεωρητικούς/στοχαστικούς κανόνες αναζήτησης όχι με την αυστηρή έννοια αφού η αναπαραγωγή των λύσεων εκτός της τυχαιότητας βασίζεται και στα αποτελέσματα της αξιολόγησης των λύσεων δηλαδή το πόσο καλή είναι μια λύση και με βάση και αυτό το κριτήριο θα παραμένει και θα εξελιχθεί μέσα

σε ένα πληθυσμό. Στο πρόβλημα της επιλογής χαρακτηριστικών τα μέλη του πληθυσμού είναι τα υποσύνολα χαρακτηριστικών (υποψήφιες λύσεις) που αναπαρίστανται από δυαδικά διανύσματα. Η επεξεργασία των υποψηφίων λύσεων λαμβάνει χώρα σε διακριτές φάσεις που ονομάζονται γενιές. Σε κάθε γενιά τα υποψήφια υποσύνολα χαρακτηριστικών αξιολογούνται βάση μιας αντικειμενικής συνάρτησης αξιολόγησης και οι πιο εύρωστες λύσεις επιλέγονται για αναπαραγωγή. Η αναπαραγωγή πραγματοποιείται διασταυρώνοντας χαρακτηριστικά (features) από διαφορετικούς υποσύνολα χαρακτηριστικών, ώστε να παραχθούν νέα υποσύνολα (children). Τα νέα υποσύνολα εισέρχονται στη συνέχεια στον πληθυσμό και η διαδικασία επαναλαμβάνεται ακολουθώντας ένα εξελικτικό μοτίβο (Δαρβινικό περιβάλλον).

Η «καθολικά βέλτιστη» μέθοδος βελτιστοποίησης και κατά συνέπεια και μια μέθοδος επιλογής χαρακτηριστικών θα πρέπει να συνδυάζει τα δύο ακόλουθα θεμελιώδη χαρακτηριστικά επίδοσης (Duan et al., 1992):

- αποτελεσματικότητα (effectiveness), δηλαδή υψηλή αξιοπιστία εντοπισμού (ή προσέγγισης) του καθολικού ακροτάτου της συνάρτησης ή αντίστοιχα του καθολικά βέλτιστου υποσυνόλων χαρακτηριστικών και,
- αποδοτικότητα (efficiency), δηλαδή υψηλή ταχύτητα σύγκλισης (εγγυημένος εντοπισμός του καθολικά βέλτιστου με εύλογο πλήθος δοκιμών).

Τα χαρακτηριστικά αυτά πολλές φορές είναι αντικρουόμενα αφού τεχνικές συστηματικής αναζήτησης προσεγγίζουν το καθολικό βέλτιστο με ακρίβεια, αλλά ταυτόχρονα απαιτούν και υψηλό υπολογιστικό κόστος, ενώ οι γρήγορες τεχνικές άμεσης αναζήτησης εγκλωβίζονται εύκολα σε τοπικά ακρότατα.

Τα εξελικτικά σχήματα βελτιστοποίησης και κατά συνέπεια και οι ΓΑ δείχνουν να υπερτερούν, στις περισσότερες κατηγορίες προβλημάτων βελτιστοποίησης. Μειονέκτημά τους, είναι η ύπαρξη αρκετών παραμέτρων εισόδου, που επηρεάζουν σημαντικά την επίδοση των αλγορίθμων (π.χ. μέγεθος πληθυσμού), ο ορισμός των οποίων απαιτεί πολλαπλούς πειραματικούς ελέγχους με διαφορετικές συνθήκες εκκίνησης.

Οι σημαντικότεροι λόγοι χρησιμοποίησης των ΓΑ στο πρόβλημα της επιλογής χαρακτηριστικών είναι οι ακόλουθοι:

- Πραγματοποιούν ταυτόχρονη εξερεύνηση σε πολλά διαφορετικά υποσύνολα (σημεία του χώρου αναζήτησης) και επομένως η εξαγωγή βέλτιστων λύσεων προκύπτει μέσα από ένα πλήθος διαφορετικών υποσυνόλων χαρακτηριστικών.
- Για την εύρεση ενός καθολικά βέλτιστου υποσυνόλου χαρακτηριστικών οι ΓΑ ξεφεύγουν ως ένα βαθμό από τον εγκλωβισμό τους σε τοπικά βέλτιστα υποσύνολα χαρακτηριστικών. Αυτό γίνεται συνδυάζοντας υποσύνολα χαρακτηριστικών και δημιουργώντας παράλληλα νέα μέσω των διαδικασιών αναπαραγωγής του πληθυσμού. Έτσι προηγούμενες λύσεις που στην ουσία ενδεχομένως να είχαν εγκλωβίσει τον αλγόριθμο σε περιοχές τοπικών βέλτιστων απομακρύνονται.
- Δουλεύουν με μια κωδικοποίηση των δεδομένων του προβλήματος και όχι με τα ίδια τα δεδομένα.
- Απαιτούν τη γνώση μόνο της αντικειμενικής συνάρτησης αξιολόγησης των υποσύνολων των χαρακτηριστικών.
- Μπορούν να συνδυαστούν με κάποια τοπική μέθοδος αναζήτησης αυξάνοντας την αποτελεσματικότητα και την αποδοτικότητα τους σε δύσκολα υπολογιστικά προβλήματα όπου η διάσταση του χώρου αναζήτησης είναι μεγάλη.

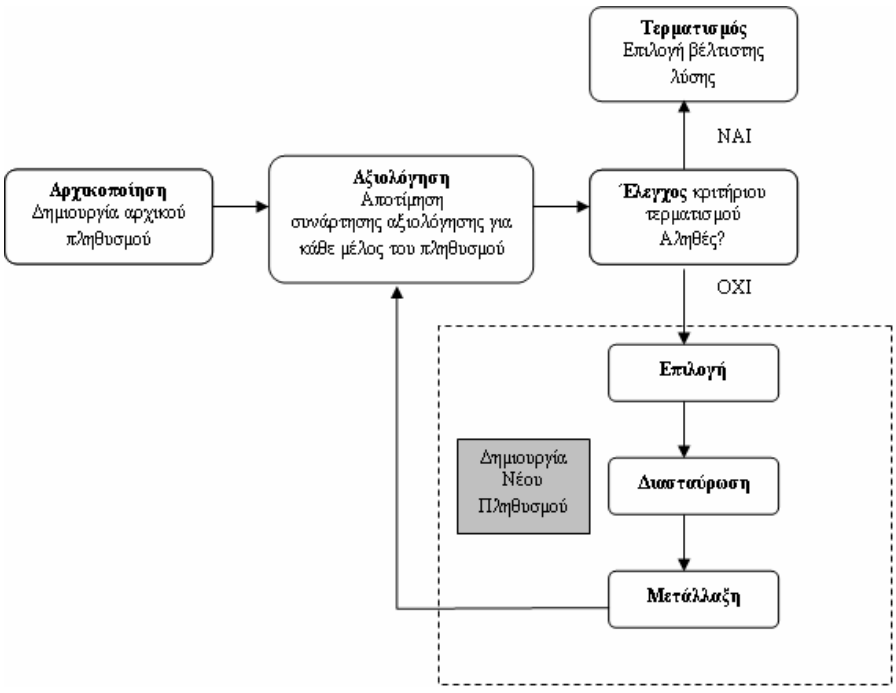
Οι ΓΑ ως μέθοδοι επιλογής χαρακτηριστικών χρησιμοποιήθηκαν για πρώτη φορά το 1989 [18] και τα τελευταία χρόνια χρησιμοποιούνται ευρέως σε προβλήματα μείωσης της διάστασης στο χώρο της Μηχανικής Μάθησης. Στον τομέα της υπολογιστικής όρασης χρησιμοποιούνται συνήθως σε συνδυασμό με μεθόδους εξαγωγής χαρακτηριστικών όπως η PCA, όπου ο γενετικός αλγόριθμος επιλέγει το βέλτιστο υποσύνολο ιδιοδιανυσμάτων που προέκυψαν από τη μέθοδο PCA [1,19].

3.2.3 Δομή και Λειτουργία Γενετικού Αλγορίθμου

Σε γενικές γραμμές η λειτουργία ενός ΓΑ καθορίζεται από τα ακόλουθα απλά βήματα:

1. Τυχαία αρχικοποίηση των μελών του αρχικού πληθυσμού.
2. Αξιολόγηση κάθε μέλους με βάση τη δοθείσα συνάρτηση αξιολόγησης καταλληλότητας.
3. Εφαρμογή της διαδικασίας της επιλογής για κάθε μέλος του τρέχοντος πληθυσμού. Άτομα με μεγάλη τιμή στη συνάρτηση καταλληλότητας έχουν μεγαλύτερη πιθανότητα να επιλεγούν.
4. Εφαρμογή της διαδικασίας της διασταύρωσης για κάποια μέλη του πληθυσμού με βάση την καθορισμένη πιθανότητα διασταύρωσης.
5. Εφαρμογή της διαδικασίας της μετάλλαξης για κάποια γονίδια των μελών του πληθυσμού με βάση την καθορισμένη πιθανότητα μετάλλαξης.
6. Επιστροφή στο βήμα 2 και έλεγχος κριτηρίου τερματισμού. Εάν δεν ικανοποιείται το κριτήριο τερματισμού η διαδικασία επαναλαμβάνεται από το βήμα 2.

Στο Σχήμα 3.1 αποτυπώνονται διαγραμματικά τα παραπάνω βήματα:



Σχήμα 3.1 Αναπαράσταση λειτουργίας ΓΑ.

Στη συνέχεια περιγράφουμε τα βασικά συστατικά της δομής ενός απλού ΓΑ.

Κωδικοποίηση

Η κωδικοποίηση αποτελεί μια σημαντική και απαραίτητη διαδικασία για έναν ΓΑ αφού ουσιαστικά συνδέει τον αλγόριθμο με τα πραγματικά δεδομένα του προβλήματος. Υπάρχουν διάφοροι τρόποι κωδικοποίησης οι οποίοι καθορίζονται από το είδος του προβλήματος που θέλουμε να λύσουμε. Ο πιο δημοφιλής και πιο απλός τρόπος, είναι η κωδικοποίηση των υποψήφιων λύσεων μέσω ενός διανύσματος δυαδικών ψηφίων (binary bits). Στη δυαδική κωδικοποίηση κάθε παράμετρος μιας υποψήφιας λύσης κωδικοποιείται μέσω ενός γονιδίου, χρησιμοποιώντας ένα δυαδικό ψηφίο (0 ή 1). Ένας άλλος τρόπος κωδικοποίησης είναι με χρήση πραγματικών αριθμών όπου κάθε υποψήφια λύση κωδικοποιείται μέσω ενός διανύσματος πραγματικών αριθμών.

Αρχικοποίηση

Η αρχικοποίηση των λύσεων σε μια σειρά προβλημάτων βελτιστοποίησης γίνεται με τυχαίο τρόπο με χρήση μιας γεννήτριας ισοκατανεμημένων αριθμών 0 ή 1. Τα αντίστοιχα γονίδια κάθε υποψηφίας λύσης σχηματίζουν ένα χρωμόσωμα/άτομο το οποίο αποτελεί μια υποψήφια λύση για τον ΓΑ. Το πλήθος των χρωμοσωμάτων που αντιπροσωπεύει αυτόνομες λύσεις του υποεξέταση προβλήματος σχηματίζει τον αρχικό πληθυσμό. Στην παρούσα εργασία και στα πλαίσια του υποεξέταση προβλήματος διαπιστώσαμε πως η αρχικοποίηση των τιμών των δυαδικών ψηφίων των χρωμοσωμάτων παίζει έναν ιδιαίτερο και σημαντικό ρόλο στην εύρεση της βέλτιστης λύσης, προσανατολίζοντας την αναζήτηση σε λύσεις με το μικρότερο δυνατό αριθμό επιλεγμένων μεταβλητών. Κατά την εφαρμογή ενός απλού ΓΑ για την επιλογή χαρακτηριστικών σε δεδομένα υψηλής διάστασης μεροληπτούμε κατά την αρχικοποίηση υπέρ των δυαδικών ψηφίων που φέρνουν την τιμή 0, δηλαδή αρχικά τα επιλεγμένα χαρακτηριστικά είναι ελάχιστα. Αυτό κρίνεται αναγκαίο, γιατί σε διαφορετική περίπτωση, ο αλγόριθμος δεν θα σύγκλινε ποτέ σε μια βέλτιστη λύση η οποία αποτελεί συνδυασμό της απόδοσης του ταξινομητή και της διάστασης της βέλτιστης λύσης. Για παράδειγμα σε ένα χώρο διάστασης 10300 χαρακτηριστικών με

μια δίκαιη κατανομή 0 και 1 κατά τη φάση της αρχικοποίησης ο αλγόριθμος δεν θα έδινε λύση 150 χαρακτηριστικών.

Συνάρτηση Καταλληλότητας (fitness function)

Δέχεται ως είσοδο την αποκωδικοποιημένη τιμή ενός χρωμοσώματος και επιστρέφει έναν αριθμό που δηλώνει το βαθμό καταλληλότητας (fitness rate), του χρωμοσώματος. Αποτελεί μέτρο της ποιότητας κάθε υποψήφιας λύσης. Οι τιμές της συνάρτησης αξιολόγησης των ατόμων ορίζουν και τις αντίστοιχες πιθανότητες επιβίωσης οι οποίες θα χρησιμοποιηθούν στη φάση της επιλογής του πληθυσμού που θα αποτελέσει την επόμενη γενιά. Σε πολλές περιπτώσεις η τιμή της συνάρτησης αξιολόγησης αποτελεί και μέρος της συνθήκης τερματισμού του αλγορίθμου.

Γενετικοί τελεστές (genetic operators)

Οι γενετικοί τελεστές χρησιμοποιούνται από τον αλγόριθμο για τη δημιουργία των γενεών σε κάθε επανάληψη. Οι κυριότεροι γενετικοί τελεστές περιγράφονται ακολούθως.

Επιλογή (selection)

Η διαδικασία της επιλογής καθορίζει ποια από τα άτομα του τρέχοντος πληθυσμού θα συνεχίσουν την εξελεγκτική διαδικασία περνώντας στη φάση της αναπαραγωγής για να κληροδοτήσουν στην επόμενη γενιά κάποια η όλα τα χαρακτηριστικά τους.

Για την διαδικασία της επιλογής υπάρχουν διάφορες τεχνικές. Δημοφιλέστερες είναι η τεχνική της Ρουλέτας (Roulette Wheel Selection) και η ταξινομημένη επιλογή (Rank Selection). Η μέθοδος του τροχού της τύχης (Ρουλέτα) αποτελεί μια στοχαστική διαδικασία δειγματοληψίας με αντικατάσταση και υλοποιείται ως εξής:

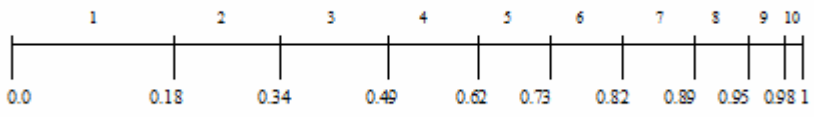
Η επιλογή των ατόμων αυτών γίνεται με βάση την τιμή της καταλληλότητας τους. Όσο μεγαλύτερη είναι η τιμή της καταλληλότητας ενός ατόμου σε σχέση με την τιμή των άλλων μελών τόσο αυξάνεται η πιθανότητα να επιλεγεί το άτομο αυτό μία ή περισσότερες φορές. Οι τιμές καταλληλότητας του πληθυσμού κανονικοποιημένες στο διάστημα [0,1] παριστάνονται από ισάριθμα διαδοχικά ευθύγραμμα τμήματα των

οποίων το μήκος είναι ανάλογο της τιμής της καταλληλότητας (απόδοσης) του κάθε ατόμου. Τα ευθύγραμμα τμήματα έχουν συνολικό μέγεθος 1.Έτσι σε ένα μέλος του πληθυσμού με μεγάλη απόδοση θα αντιστοιχεί και μεγαλύτερο ευθύγραμμο τμήμα από ένα μέλος με μικρότερη απόδοση. Αυτό σημαίνει ότι θα έχει και μεγαλύτερη πιθανότητα να επιλεγεί. Στη συνέχει παράγεται ένας τυχαίος αριθμός στο διάστημα [0,1] και το ευθύγραμμο τμήμα στο οποίο ανήκει μας φανερώνει και το αντίστοιχο μέλος του πληθυσμού που θα επιλεγεί. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να επιλεγεί ο επιθυμητός αριθμός ατόμων που θα συμμετάσχουν στην αναπαραγωγική διαδικασία. Ακολουθεί ένα παράδειγμα υλοποίησης της μεθόδου Ρουλέτας. Έστω ο ακόλουθος πίνακας του Σχήματος 3.2 ο οποίος περιγράφει τις καταλληλότητες και τις αντίστοιχες πιθανότητες επιλογής των ατόμων του πληθυσμού.

Αριθμός μέλους	1	2	3	4	5	6	7	8	9	10
Τιμή καταλληλότητας	2.0	1.8	1.6	1.4	1.2	1.0	0.8	0.6	0.4	0.2
Πιθανότητα επιλογής	0.18	0.16	0.15	0.13	0.11	0.09	0.07	0.06	0.03	0.02

Σχήμα 3.2. Πίνακας καταλληλοτήτων γενετικού πληθυσμού.

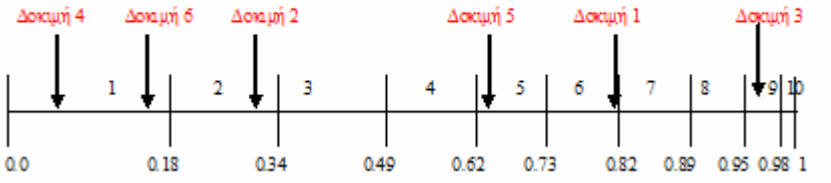
Με βάση τον πίνακα αποτυπώνουμε σχηματικά το μέγεθος των πιθανοτήτων επιλογής των ατόμων του πληθυσμού. Στο Σχήμα 3.3 φαίνεται ότι όσο μεγαλύτερη είναι η πιθανότητα επιλογής για κάποιο άτομο τόσο μεγαλύτερο μήκος καταλαμβάνει στο διάνυσμα της ρουλέτας .



Σχήμα 3.3 Διάνυσμα αναπαράστασης ρουλέτας.

Για να επιλέξουμε το πλήθος των μελών τραβάμε αντίστοιχους τυχαίους αριθμούς. Έστω ότι θέλουμε να επιλέξουμε 6 άτομα τότε χρησιμοποιούμε με τη σειρά τους ακολούθους τυχαίους αριθμούς: 0.81, 0.32, 0.96, 0.01, 0.65, 0.17.

Η διαδικασία επιλογής σύμφωνα με τον μηχανισμό της ρουλέτας αποτυπώνεται στο Σχήμα 3.4 ως εξής:

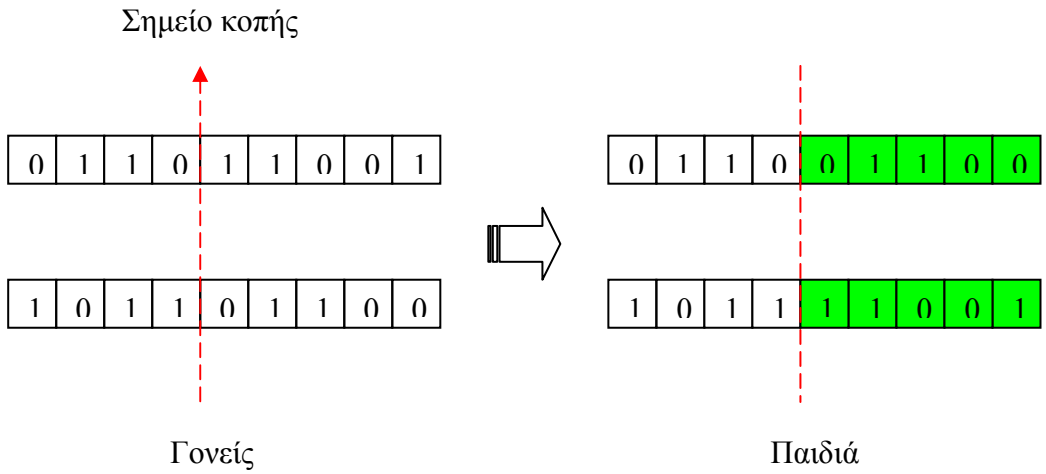


Σχήμα 3.4 Διαδικασία επιλογής της ρουλέτας.

Μετά την επιλογή τα άτομα που προκύπτουν και θα αποτελέσουν τη δεξαμενή ζευγαρώματος είναι τα: 6, 2, 9, 1, 5, 1. Η διάταξη των ατόμων βασίζεται στην σειρά με βάση την οποία επιλέχθηκαν. Παρατηρούμε ότι το άτομο 1 κατάφερε να επιλεγεί 2 φορές αφού είχε και τη μεγαλύτερη πιθανότητα με βάση την τιμή της καταλληλότητας του.

Διασταύρωση (Crossover)

Η διασταύρωση αποτελεί μια σημαντική λειτουργία των ΓΑ. Ο πληθυσμός που προέκυψε μετά τη διαδικασία επιλογής αποτελεί τη δεξαμενή ζευγαρώματος από όπου επιλέγονται τα άτομα που θα συνδυαστούν ώστε να δημιουργηθούν νέα άτομα. Στόχος της διασταύρωσης είναι ο νέος πληθυσμός που θα προκύψει μετά την εφαρμογή της να είναι καλύτερος από τον προηγούμενο και τα άτομα-απόγονοι να φέρουν τα καλύτερα χαρακτηριστικά των γονέων τους. Υπάρχουν διάφοροι τρόποι διασταύρωσης. Ο πιο απλός τρόπος είναι η διασταύρωση ενός σημείου (Single Point Crossover). Σύμφωνα με αυτόν τον τρόπο για κάθε ζεύγος χρωμοσωμάτων που επιλέχθηκε για διασταύρωση παράγεται τυχαία ένας ακέραιος k από το διάστημα [1,m-1] όπου m το μήκος του δυαδικού ψηφίου σε χρωμοσώματα. Ο αριθμός k προσδιορίζει το σημείο κοπής ή αλλιώς σημείο διασταύρωσης των χρωμοσωμάτων. Από το σημείο αυτό, μέχρι και το τέλος του μήκους των χρωμοσωμάτων γίνεται ανταλλαγή των γονιδίων τους. Η διαδικασία περιγράφεται στο Σχήμα 3.5.



Σχήμα 3.5 Διασταύρωση ενός σημείου.

Άλλοι τρόποι διασταύρωσης είναι η διασταύρωση πολλών σημείων (Multi Point Crossover) και η ομοιόμορφη διασταύρωση (Uniform Crossover). Στη διασταύρωση πολλών σημείων για κάθε ζεύγος χρωμοσωμάτων που επιλέχθηκε για διασταύρωση παράγεται τυχαία n ακέραιοι k_i από το διάστημα [1,m-1] όπου m το μήκος σε δυαδικά ψηφία των χρωμοσωμάτων. Οι αριθμοί k_i προσδιορίζουν τα σημεία κοπής ή αλλιώς σημεία διασταύρωσης των χρωμοσωμάτων. Τα γονίδια των γονέων μεταξύ των διαδοχικών σημείων διασταύρωσης ανταλλάσσονται για να παράγουν τα παιδιά τους. Στη ομοιόμορφη διασταύρωση οι δύο γονείς θα ανταλλάξουν το γενετικό υλικό με βάση μια μάσκα από δυαδικά ψηφία. Η μάσκα αυτή είναι μήκους m όσο το μήκος των χρωμοσωμάτων και επιλέγεται με τυχαίο τρόπο. Ανάλογα με την τιμή του δυαδικού ψηφίου στην κάθε θέση της μάσκας καθορίζεται για κάθε παιδί από ποιόν γονέα θα προέρχεται το γενετικό υλικό στην αντίστοιχη θέση. Για παράδειγμα αν το δυαδικό ψηφίο στην 1η θέση της μάσκας έχει τιμή 1 τότε το $1^{ο}$ παιδί θα πάρει την τιμή της αντίστοιχης θέση του $1^{ου}$ γονέα αλλιώς αν είναι 0 τότε θα θα πάρει την τιμή της αντίστοιχης θέση του $2^{ου}$ γονέα. Για το $2^{ο}$ παιδί ισχύει το αντίστροφο, δηλαδή από όποιον γονέα πάρει γενετικό υλικό το 1ο παιδί, το $2^{ο}$ παιδί θα πάρει από τον άλλον γονέα.

Μετάλλαξη

Η μετάλλαξη λαμβάνει χώρα μετά τη διασταύρωση και όπως συμβαίνει και στους φυσικούς οργανισμούς πραγματοποιείται αραιά γιατί σε διαφορετική περίπτωση ο αλγόριθμος εκφυλίζεται σε εντελώς τυχαία αναζήτηση. Όταν συμβαίνει, εξασφαλίζει ότι κανένα σημείο του χώρου αναζήτησης δεν αποκλείεται από την διαδικασία αναζήτησης και διορθώνει τυχόν απώλεια γενετικών πληροφοριών στο στάδιο της επιλογής και διασταύρωσης. Υπάρχουν αρκετές μέθοδοι μετάλλαξης. Δύο από αυτές είναι οι ακόλουθες:

Μετάλλαξη ενός σημείου: Απλή αντιστροφή ενός τυχαίου δυαδικού ψηφίου του γονέα, από 0 σε 1 ή το αντίστροφο και δημιουργία ενός παιδιού. Η επιλογή γίνεται τυχαία με μια μικρή προκαθορισμένη πιθανότητα την λεγόμενη πιθανότητα μετάλλαξης.

Μετάλλαξη πολλών σημείων: Παρόμοια φιλοσοφία με την μετάλλαξη ενός σημείου με τη διαφορά ότι επιλέγονται πάνω από σημεία μετάλλαξης για έναν γονέα.

Παράμετροι Γενετικών αλγορίθμων

Οι ΓΑ εξαρτώνται από ένα πλήθος παραμέτρων το οποίο προσαρμόζεται ανάλογα με το πρόβλημα που επιλύεται κάθε φορά. Το πρόβλημα είναι ότι η ανάθεση τιμών σε αυτές τις παραμέτρους δεν καθορίζεται από κάποιον κανόνα αλλά είναι αποτέλεσμα πειραματικών μελετών για κάθε είδος προβλήματος. Οι βασικές παράμετροι είναι η πιθανότητα διασταύρωσης (crossover probability), η πιθανότητα μετάλλαξης (mutation probability) και το μέγεθος του πληθυσμού (population size).

Η **πιθανότητα διασταύρωσης (pc)** καθορίζει τη συχνότητα της διασταύρωσης, δηλαδή πόσα μέλη του πληθυσμού που επεξεργάζεται ο ΓΑ θα διασταυρωθούν. Η πιθανότητα αυτή ποικίλει ανάλογα με το είδος του προβλήματος. Πιθανότητα διασταύρωσης ίση με 1 σημαίνει όλα τα μέλη του πληθυσμού θα διασταυρωθούν μεταξύ τους, ενώ αν είναι 0, τότε οι απόγονοι θα είναι πιστά αντίγραφα των γονέων εκτός αν συμβούν αλλαγές στη φάση της μετάλλαξης.

Η **πιθανότητα μετάλλαξης (pm)** καθορίζει το πόσο συχνά τα γονίδια των χρωμοσωμάτων θα αλλάζουν κατάσταση (από 0 σε 1 ή το αντίστροφο). Η τιμή της

πιθανότητας μετάλλαξης θα πρέπει να είναι 1/n όπου n ο αριθμός των μεταβλητών παραμέτρων του υποεξέταση προβλήματος. Σε γενικές γραμμές το ποσοστό της μετάλλαξης θα πρέπει να είναι χαμηλό, αλλιώς ο αλγόριθμος εγκλωβίζεται σε ένα βρόγχο τυχαίας αναζήτησης. Διάφορες μελέτες έχουν γίνει για τη σωστή προσέγγιση του ποσοστού της μετάλλαξης. Ο Holland υποστήριξε ότι η μετάλλαξη είναι δευτερογενής τελεστής ενώ ο Goldberg προτείνει να αντιστρέφεται ένα στα χίλια δυαδικά ψηφία κατά μέσο όρο σε κάθε επανάληψη. Η μετάλλαξη αντιμετωπίζει τα δυαδικά ψηφία όλων των μελών του πληθυσμού σαν μια ενωμένη συμβολοσειρά και η αναφορά του Goldberg μιλάει για το σύνολο των δυαδικών ψηφίων του πληθυσμού.

Το **μέγεθος του πληθυσμού** δηλώνει τον αριθμό των υποψηφίων λύσεων κάθε γενιάς. Ο καθορισμός του μεγέθους αυτού είναι συνάρτηση του είδους του προβλήματος που θα επιλυθεί και των διαθέσιμων υπολογιστικών πόρων. Το μέγεθος του πληθυσμού είναι κρίσιμο για την εύρεση της βέλτιστης λύσης Οι μικροί πληθυσμοί συγκλίνουν πιο γρήγορα σε τοπικά βέλτιστα αλλά εγκλωβίζονται σε αυτά, ενώ οι μεγάλοι πληθυσμοί είναι πολύ πιθανόν να μην εγκλωβίσουν τον αλγόριθμο σε τοπικό βέλτιστα αλλά θέλουν περισσότερο υπολογιστικό χρόνο και πόρους για την εξεύρεση της λύσης.

3.2.4 Θεωρητική θεμελίωση Γενετικών Αλγορίθμων

Οι Γενετικοί αλγόριθμοι αν και είναι απλοί στην εφαρμογή τους η συμπεριφορά τους είναι δύσκολο να κατανοηθεί και να εξηγηθεί. Η θεωρητική θεμελίωση των Γενετικών αλγορίθμων βασίζεται στην αναπαράσταση των λύσεων ως δυαδικές συμβολοσειρές και στην έννοια του σχήματος το οποίο αποτελεί ένα πρότυπο που επιτρέπει τον προσδιορισμό της ομοιότητας μεταξύ των χρωμοσωμάτων.

Ο Holland [20] πρώτος υπέθεσε ότι οι ΓΑ λειτουργούν: ανακαλύπτοντας, δίνοντας έμφαση και ανασυνδυάζοντας «καλά δομικά στοιχεία» λύσεων, με ένα τρόπο υψηλού παραλληλισμού. Για να τυποποιήσει τα λεγόμενα καλά δομικά στοιχεία εισήγαγε την έννοια του σχήματος. Ένα σχήμα κατασκευάζεται εισάγοντας ένα αδιάφορο σύμβολο

(don't care symbol) * στο αλφάβητο των γονιδίων (0,1) και αναπαριστά όλες τις συμβολοσειρές (ένα υπερεπίπεδο-επίπεδο ή άλλο υποσύνολο του χώρου αναζήτησης) οι οποίες ταιριάζουν σε όλες τις θέσεις εκτός από αυτές με το αδιάφορο σύμβολο [21].

Ο Goldberg περιγράφει ότι ένας ΓΑ αναζητεί απόδοση κοντά στο βέλτιστο, τοποθετώντας δίπλα-δίπλα μικρού μήκους, χαμηλής τάξης και υψηλής απόδοσης σχήματα, που ονομάζονται δομικά στοιχεία της λύσης [22]. Τα δομικά αυτά στοιχεία συνδυάζονται μεταξύ τους μέσω της διασταύρωσης και παράγουν νέα χρωμοσώματα με υψηλή απόδοση. Με αυτόν τον τρόπο μειώνεται η πολυπλοκότητα του προβλήματος γιατί αντί να παράγουμε λύσεις δοκιμάζοντας όλους τους δυνατούς συνδυασμούς των λύσεων της προηγούμενης γενιάς,κατασκευάζουμε λύσεις συνδυάζοντας τις καλύτερες επιμέρους λύσεις παλαιότερων γενεών.

Η παραπάνω υπόθεση ονομάστηκε υπόθεση δομικών τμημάτων (BBH=Building Block Hypothesis). Αξίζει να αναφερθεί ότι η επεξεργασία των δομικών στοιχείων γίνεται παράλληλα. Η ιδιότητα αυτή του «εγγενούς παραλληλισμού» (implicit parallelism) θεωρείται εξαιρετικά σημαντική, καθώς ο ΓΑ ουσιαστικά υπολογίζει την μέση τιμή της συνάρτησης καταλληλότητας για ένα πολύ μεγαλύτερο πλήθος σχημάτων με το ίδιο υπολογιστικό κόστος. Με αυτόν τον τρόπο ο ΓΑ δειγματοληπτεί αποδοτικά μεγαλύτερο τμήμα του χώρου λύσεων, αποφεύγοντας την παγίδευση σε τοπικά βέλτιστα [22].

Η παραπάνω θεωρία είναι καθαρά εμπειρική και δεν εμπεριέχει κάποια φορμαλιστική μαθηματική ανάλυση της συμπεριφοράς των ΓΑ. Στην πραγματικότητα, οι ΓΑ. δεν έχουν ακόμη αναλυθεί μαθηματικά και αυτό είναι το μεγαλύτερό τους μειονέκτημα. Παρόλα αυτά, παρουσιάζουν υψηλή αποδοτικότητα σε ένα ευρύ φάσμα προβλημάτων βελτιστοποίησης και υποστηρίζουν με επιτυχία,τη λειτουργία πολλών συστημάτων σε τομείς όπως, η Επεξεργασία Εικόνας (Image Processing), η σχεδίαση με χρήση Η/Υ (Computer Aided Desing –CAD), η Οικονομία, οι Τηλεπικοινωνίες, η Τεχνολογία Λογισμικού, ο Χρονοπρογραμματισμός, τα Τεχνολογικά και Μηχανολογικά Συστήματα [21].

3.3 Simulated Annealing (Προσομοιωμένη Ανόπτηση)

Μια άλλη στοχαστική μέθοδος αναζήτησης που χρησιμοποιείται σε προβλήματα βελτιστοποίησης είναι και ο αλγόριθμος της Προσομοιωμένης Ανόπτησης (Simulated Annealing). Στην παρούσα εργασία χρησιμοποιήθηκε ως τοπική μέθοδος αναζήτησης των βέλτιστων χαρακτηριστικών.

Το όνομα και η φιλοσοφία του αλγορίθμου προέρχεται από την διαδικασία της ανόπτησης, μια τεχνική που χρησιμοποιείται για τη σκλήρυνση των μετάλλων και του γυαλιού με ελεγχόμενη τη θέρμανση και την ψύξη. Συγκεκριμένα το γυαλί αρχικά θερμαίνεται σε υψηλή θερμοκρασία (heating up) και στη συνέχεια γίνεται βαθμιαία-σταδιακή ψύξη (cooling it down), ώστε να σχηματίζονται μεγάλοι συμπαγείς κρύσταλλοι και το υλικό να βρίσκεται σε κατάσταση ελάχιστης ενέργειας. Η μείωση της θερμοκρασίας γίνεται σταδιακά για να μην προκληθούν ατέλειες στο υλικό. Σε υψηλές θερμοκρασίες τα άτομα μέσα στο υλικό έχουν υψηλές ενέργειες και επομένως έχουν περισσότερη ελευθερία κινήσεων. Ενώ η θερμοκρασία μειώνεται, μειώνονται παράλληλα και οι ατομικές ενέργειες. Ένας κρύσταλλος με ομαλή δομή επιτυγχάνεται στην κατάσταση όπου το σύστημα έχει ελάχιστη ενέργεια.

Η αντιστοιχία ανάμεσα στα συνδυαστικά προβλήματα βελτιστοποίησης και στα προβλήματα της σκλήρυνσης των υλικών έγινε από τους Kirkpatrick et al, οι οποίοι πρότειναν τον αλγόριθμο της προσομοιωμένης ανόπτησης, εκτιμώντας την αναλογία μεταξύ του φορτίου ενέργειας ενός υλικού και το κόστος μιας λύσης ενός συνδυαστικού προβλήματος βελτιστοποίησης [23]. Με βάση αυτήν την αναλογία, οι καταστάσεις του στερεού, αντιπροσωπεύουν υποψήφιες λύσεις του προβλήματος βελτιστοποίησης. Οι ενέργειες των καταστάσεων αντιστοιχούν στην αξία των λύσεων με βάση κάποια συνάρτηση αξιολόγησης. Η ελάχιστη ενέργεια ή αντίστοιχα η μέγιστη ενέργεια, ανάλογα με τον ορισμό του προβλήματος, αντιστοιχεί στην βέλτιστη λύση του προβλήματος. Επίσης, όπως στην περίπτωση των υλικών, η θερμοκρασία λειτουργεί ως μια παράμετρος ελέγχου της διαδικασίας επιτρέποντας στο σύστημα να οργανωθεί σε καταστάσεις μειωμένης ενέργειας, έτσι και στον αλγόριθμο Simulated Annealing ορίζεται μια αντίστοιχη παράμετρος και ένα

χρονοδιάγραμμα μείωσης της τιμής της έτσι ώστε να εξερευνηθούν καταστάσεις που οδηγούν σε βέλτιστη λύση.

Υπάρχουν πολλές παραλλαγές του αλγορίθμου της προσομοιωμένης ανόπτησης. Τα γενικά βήματα παρουσιάζονται ακολούθως.

function SIMULATED-ANNEALING($problem, schedule$)
returns a solution state
  inputs: $problem$, a problem
    $schedule$, a mapping from time to "temperature"
  local variables: $current$, a node $next$, a node T, the temperature
  $current \leftarrow$ MAKENODE(RANDOMSTATE[$problem$])
  for $t \leftarrow 1$ **to** ∞ **do**
    $T \leftarrow schedule[t]$
    if $T = 0$ **then return** $current$
    $next \leftarrow$ a randomly selected successor of $current$
    $\Delta E \leftarrow$ VALUE[$next$]– VALUE[$current$]
    if $\Delta E > 0$ **then** $current \leftarrow next$
    else $current \leftarrow next$ only with probability $e^{\Delta E/T}$

Ο αλγόριθμος ξεκινώντας από μια αρχική λύση επαναληπτικά εκτελεί στοχαστική αναζήτηση διαλέγοντας μια τυχαία κατάσταση στη γειτονία της τρέχουσας κατάστασης. Η λύση της νέας κατάστασης γίνεται αποδεκτή με πιθανότητα ίση με 1 αν βελτιώνει την λύση της τρέχουσας κατάστασης. Σε διαφορετική περίπτωση ο αλγόριθμος κάνει δεκτή την νέα λύση με πιθανότητα μικρότερη του 1 η οποία καθορίζεται από τον τύπο $e^{\Delta E/T}$ (Metropoli's criterion).

Αναλυτικότερα η πιθανότητα να δεχτεί ο αλγόριθμος μια νέα κατάσταση ορίζεται ως:

$$p = \begin{Bmatrix} 1 \, if \ \Delta E > 0 \\ e^{\Delta E / T} \end{Bmatrix}$$

Η παραπάνω πιθανότητα στο σκέλος που αφορά την αποδοχή μια νέας λύσης χειρότερης από αυτήν που έχει βρει μέχρι τώρα, μειώνεται εκθετικά από δύο παράγοντες. Ο $1^{ος}$ παράγοντας έχει να κάνει με το πόσο κακή είναι η λύση της νέας κατάστασης όπως αυτό προσδιορίζεται από την ποσότητα ΔΕ που δηλώνει τη

διαφορά της ενέργειας τη χρονική στιγμή t+1 από την ενέργεια τη χρονική στιγμή t. Ο $2^{ος}$ παράγοντας έχει να κάνει με την μείωση της θερμοκρασίας T. Οι λεγόμενες χειρότερες λύσεις γίνονται πιο συχνά αποδεκτές στον αρχικό χρόνο εκτέλεσης του αλγορίθμου όταν η παράμετρος (θερμοκρασία) T είναι υψηλή ενώ, καθώς μειώνεται η παράμετρος T τότε μειώνονται οι πιθανότητες να γίνουν αποδεκτές. Σε υψηλή θερμοκρασία η αναζήτηση είναι σχεδόν τυχαία, ενώ σε χαμηλή θερμοκρασία η αναζήτηση χαρακτηρίζεται ως «σχεδόν λαίμαργη» (almost greedy). Στη θερμοκρασία T=0, η αναζήτηση γίνεται απολύτως «λαίμαργη», δηλαδή μόνο καλές κινήσεις γίνονται αποδεκτές.

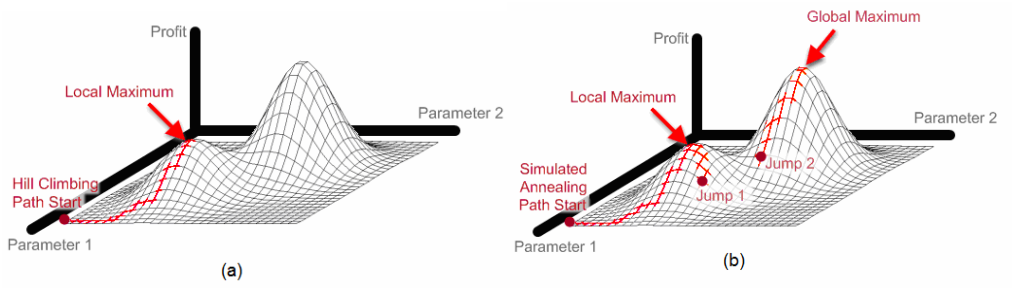
Το σημείο κλειδί της επιτυχίας του αλγορίθμου είναι το γεγονός ότι καταφέρνει να απεγκλωβίζεται από τοπικά βέλτιστα, αφού αποδέχεται νέες λύσεις οι οποίες δεν είναι άμεσα καλύτερες από προηγούμενες.

Προκειμένου να λειτουργήσει ικανοποιητικά ο αλγόριθμος θα πρέπει να καθοριστούν διάφοροι παράμετροι όπως: η αρχική θερμοκρασία, το πρόγραμμα ψύχρανσης και το κριτήριο τερματισμού της διαδικασίας. Επιπλέον θα πρέπει να ληφθούν αποφάσεις σχετικά με την συνάρτηση κόστους και τον τρόπο επιλογής της γειτονιάς μιας κατάστασης. Όσον αφορά τον τρόπο με τον οποίο ο αλγόριθμος μεταβαίνει από μια λύση σε μια άλλη, αυτό θα πρέπει να γίνεται με μικρά βήματα επιλέγοντας γειτονικές λύσεις μικρού μεγέθους. Η θερμοκρασία έναρξης της διαδικασίας θα πρέπει να είναι αρκετά υψηλή έτσι ώστε να επιτρέπει την μετάβαση σε όλες τις γειτονικές λύσεις. Ωστόσο σε ορισμένα προβλήματα δεν είναι δυνατόν να προσδιοριστεί εκ των προτέρων κατάλληλη τιμή αρχικής θερμοκρασίας και απαιτούνται αρκετοί πειραματισμοί για την εύρεσή της. Για το πρόγραμμα ψύχρανσης μπορεί να χρησιμοποιηθούν τα εξής σχήματα:

- $T = aT$, με $0.8 < a < 0.995$.
- $T = T+bT$, με το b κοντά στο 0.
- $T = c/\log(1+k)$, k είναι ο αριθμός της επανάληψης και c μία σταθερά.

Όπως προκύπτει από τα παραπάνω αλγόριθμος της Προσομοιωμένης Ανόπτησης αποτελεί μια στοχαστική μέθοδο με κύριο στόχο την αποφυγή εγκλωβισμού της αναζήτησης σε τοπικά βέλτιστα. Για να αντιληφθούμε καλύτερα αυτό το πλεονέκτημα της μεθόδου συγκρίνουμε τη μέθοδο με τον αλγόριθμο της αναρρίχησης

λόφων (Hill Climbing) σε μια υποθετική αναζήτηση βέλτιστης λύσης. Όπως φαίνεται στο Σχήμα 3.6 (a) ο Hill Climbing μπορεί να εγκλωβιστεί σε μια κατάσταση της οποίας οι γειτονικές καταστάσεις είναι χειρότερες λύσεις για το πρόβλημα. Η συγκεκριμένη κατάσταση όμως δεν είναι η καθολικά βέλτιστη. Αυτό οφείλεται στο γεγονός ότι δεν γίνονται κινήσεις προς τα κάτω δηλαδή προς καταστάσεις με χειρότερο κόστος. Αντίθετα ο Simulated Annealing όπως φαίνεται στο Σχήμα 3.6 (b), προσπαθεί με «αναταράξεις» του χώρου αναζήτησης να ξεφύγει από τοπικά βέλτιστα κάνοντας κινήσεις προς τα κάτω δηλαδή δέχεται χειρότερες λύσεις. Με το 2^{o} jump απεγκλωβίστηκε και βρήκε τελικά το ολικό βέλτιστο.



Σχήμα 3.6. (a) Εγκλωβισμός του Hill Climbing σε τοπικό βέλτιστο. (b) Αποφυγή τοπικού βέλτιστου με στοχαστικά βήματα προς τα κάτω από τη μέθοδο Simulated Annealing. Το 2^{o} jump απεγκλώβισε την αναζήτηση και έφτασε σε ολική βέλτιστη λύση.

3.4 Ταξινομητής SVM (Support Vector Machine)

Ο ταξινομητής SVM στα πλαίσια της παρούσης εργασίας χρησιμοποιείται στον υπολογισμό της συνάρτησης καταλληλότητας (fitness function) των υποψηφίων υποσυνόλων χαρακτηριστικών αποτιμώντας κάθε φορά το υποέξεταση υποσύνολο χαρακτηριστικών και επιστρέφοντας την απόδοσή του σε αυτό. Μια εναλλακτική μέθοδος ταξινόμησης για τον υπολογισμό της συνάρτησης καταλληλότητας, θα μπορούσαν να είναι και τα Νευρωνικά Δίκτυα (Neural Networks), τα οποία χρησιμοποιούνται ευρύτατα σε προβλήματα ταξινόμησης. Οι ταξινομητές SVMs και

τα Νευρωνικά δίκτυα χρησιμοποιούνται και σε προβλήματα υπολογιστικής όρασης, όπως αυτό της αναγνώρισης προσώπων [24,25]. Στη συνέχεια ακολουθεί μια συνοπτική περιγραφή των SVMs.

Οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines), είναι ένα σύνολο μεθόδων μάθησης με επίβλεψη (supervised learning) που αναπτύχθηκαν από τον Vapnik και χρησιμοποιούνται ευρύτατα σε προβλήματα ταξινόμησης και παλινδρόμησης παρουσιάζοντας υψηλή απόδοση σε χώρους δεδομένων υψηλής διάστασης. Οι ταξινομητές SVMs μπορεί να χρησιμοποιηθούν για γραμμικά και μη γραμμικά δεδομένα. Στόχος ενός ταξινομητή βασιζόμενου σε διανύσματα υποστήριξης είναι να «κατασκευάσει» το βέλτιστο διαχωριστικό υπερεπίπεδο σε ένα πολυδιάστατο χώρο χαρακτηριστικών (feature space). Με τον όρο «βέλτιστο» εννοούμε στην περίπτωση γραμμικά διαχωρίσιμων κατηγοριών το υπερεπίπεδο που θα κατασκευαστεί να είναι αυτό με το μέγιστο περιθώριο (ή απόσταση) από τα πιο κοντινά παραδείγματα στον πολυδιάστατο χώρο των χαρακτηριστικών. Η μέθοδος SVM βρίσκει αυτό το υπερεπίπεδο χρησιμοποιώντας:

- Διανύσματα υποστήριξης (support vectors), τα οποία αποτελούν τα σημαντικότερα στιγμιότυπα εκπαίδευσης.
- Όρια – margins, που καθορίζονται από τα διανύσματα υποστήριξης.

Στην περίπτωση όπου η γραμμική διαχωρισιμότητα στο χώρο των χαρακτηριστικών δεν ισχύει τότε ταξινομητής με το μέγιστο περιθώριο πρέπει να τροποποιηθεί για να αντιμετωπιστεί το πρόβλημα. Σε αυτή την περίπτωση, μιλάμε για έναν ταξινομητή εύκαμπτου περιθωρίου (soft margin).

Θεωρούμε ότι έχουμε διαθέσιμα δεδομένα εκπαίδευσης για 2 κατηγορίες:

$$(x_1, y_1)(x_2, y_2)...(x_l, y_l), \quad x_i \in R^l, \; y_i \in \{-1,+1\}$$

Η συνάρτηση απόφασης που χρησιμοποιείται για το διαχωρισμό δύο κατηγοριών είναι η:

$$f(x) = \text{sgn}(\sum_{i=1}^{l} a_i y_i K(x, x_i) + b)$$

όπου $K(x_i, x_j)$ είναι μια συνάρτηση πυρήνα (kernel) και το πρόσημο της συνάρτησης $f(x)$ προσδιορίζει σε ποια κατηγορία ανήκει το x.

Η κατασκευή ενός βέλτιστου υπερεπιπέδου είναι ισοδύναμη με την εύρεση όλων των μη μηδενικών α_i. Τα σημεία x_i για τα οποία ισχύει $a_i > 0$ αποτελούν τα διανύσματα υποστήριξης (support vectors) και είναι τα σημαντικότερα δείγματα του συνόλου εκπαίδευσης.

Οι βασικές συναρτήσεις kernel που χρησιμοποιούνται σε προβλήματα αναγνώρισης προτύπων είναι εξής:

- Γραμμικός πυρήνας (linear): $K(x_i, x_j) = x_i x_j$
- Ακτινικής συνάρτησης βάσης (radial basefunction–RBF): $K(x_i, x_j) = e^{-\gamma \|x_i - x_j\|^2}, \gamma > 0$
- Πολυωνυμικός πυρήνας (polynomial): $K(x_i, x_j) = (\gamma x_i \bullet x_j + r)^d, \gamma > 0$
- Σιγμοειδής (Sigmoid): $K(x_i, x_j) = \tanh(\gamma x_i \bullet x_j + r)$

Τα γ, r και d αποτελούν τις παραμέτρους kernel.

Για την περίπτωση προβλημάτων ταξινόμησης με περισσότερες κατηγορίες υπάρχουν δύο βασικές προσεγγίσεις : η One Versus All και η One Versus One.

Στην προσέγγιση One Versus All για ένα πρόβλημα ταξινόμησης L κατηγοριών εκπαιδεύονται L SVMs. Κάθε SVM διαχωρίζει μια κατηγορία από τις υπόλοιπες. Η ταξινόμηση ενός νέου δείγματος, είναι απόφαση του ταξινομητή που δίνει τη μεγαλύτερη τιμή σαν έξοδο.

Στην προσέγγιση One Versus One για ένα αντίστοιχο πρόβλημα ταξινόμησης L κατηγοριών εκπαιδεύονται $\frac{L(L-1)}{2}$ SVMs. Κάθε SVM διαχωρίζει ένα ζεύγος κατηγοριών. Για την ταξινόμηση ενός νέου δείγματος γίνεται σύγκριση κάθε κατηγορίας με καθεμία από τις υπόλοιπες L-1 ξεχωριστά. Για κάθε σύγκριση η

επικρατούσα κατηγορία παίρνει μια ψήφο και το δείγμα ταξινομείται στην κατηγορία που έχει συγκεντρώσει τις περισσότερες ψήφους [26].

ΚΕΦΑΛΑΙΟ 4. ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΣ ΕΠΙΛΟΓΗΣ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ

4.1 Εισαγωγή στη μέθοδο επιλογής χαρακτηριστικών

Οι Γενετικοί αλγόριθμοι (ΓΑ) αποτελούν μια μέθοδο επιλογής χαρακτηριστικών με ικανοποιητική απόδοση. Σε πολυδιάστατους χώρους αναζήτησης (face recognition) οι ΓΑ και η μέθοδος της Προσομοιωμένης Ανόπτησης παρουσιάζουν προβλήματα υπολογιστικού κόστους, αφού ο πληθυσμός των υποψηφίων λύσεων που πρέπει να εξεταστούν, εξαρτάται άμεσα από τον αριθμό των χαρακτηριστικών.

Αναφέρθηκε ήδη, ότι κάθε χαρακτηριστικό μιας υποψήφιας λύσης αναπαρίσταται μέσω ενός δυαδικού ψηφίου (0 ή 1). Τα δυαδικά ψηφία κάθε υποψηφίας λύσης σχηματίζουν ένα χρωμόσωμα το οποίο αποτελεί μια υποψήφια λύση για τον ΓΑ. Το πλήθος των χρωμοσωμάτων που αντιπροσωπεύει αυτόνομα υποψήφια βέλτιστα υποσύνολα χαρακτηριστικών σχηματίζει τον αρχικό πληθυσμό. Έστω για παράδειγμα δουλεύουμε σε ένα χώρο διάστασης 10300 χαρακτηριστικών. Με μια δίκαια κατανομή των 0 και 1 στη φάση της αρχικοποίησης θα είχαμε χρωμοσώματα με 5000

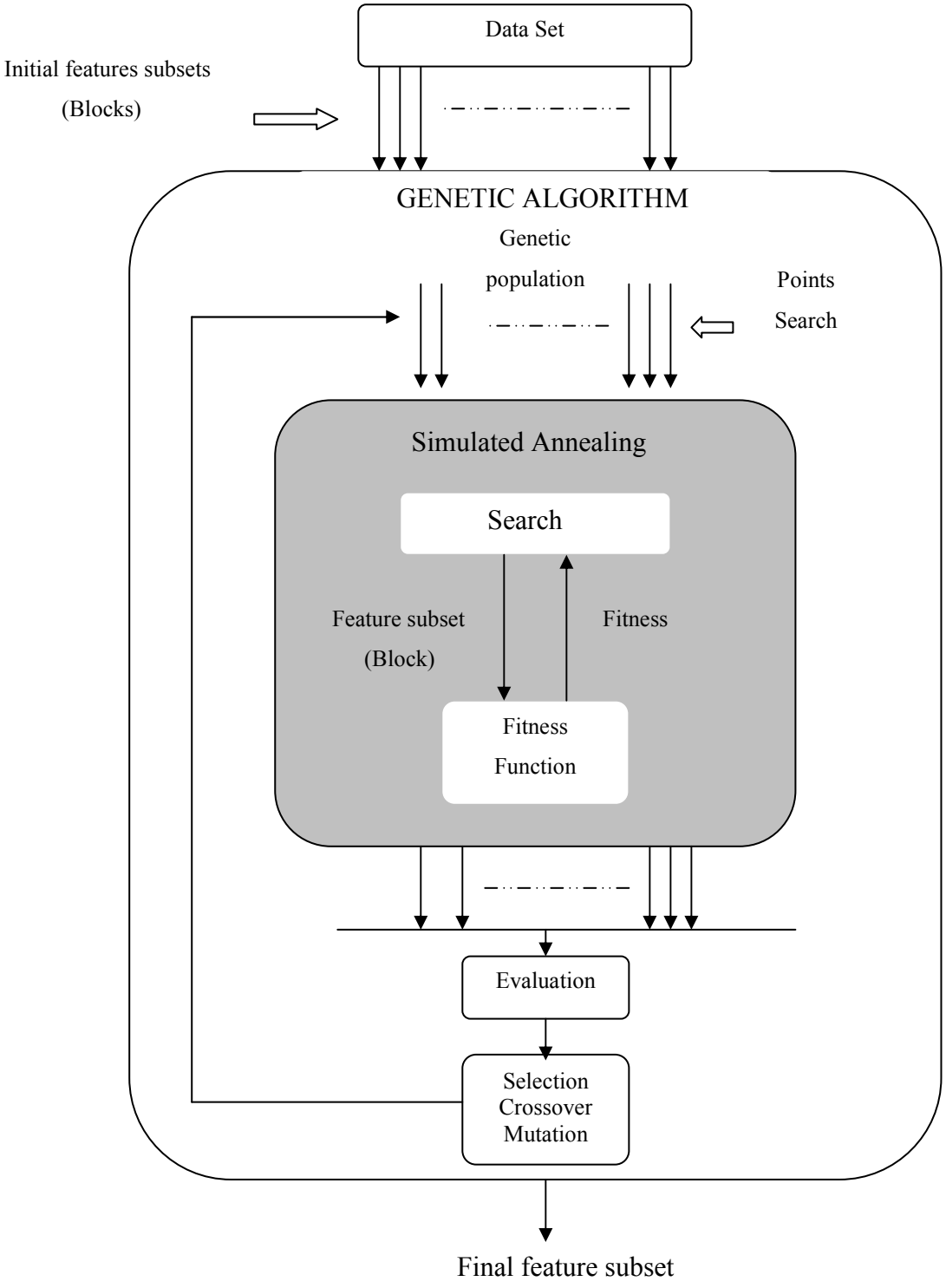
περίπου χαρακτηριστικά επιλεγμένα. Το ερώτημα που τίθεται είναι: αν το βέλτιστο υποσύνολο χαρακτηριστικών αποτελείται από 170 χαρακτηριστικά, με ποιο τρόπο ο απλός γενετικός αλγόριθμος θα έφτανε σε αυτό; Είναι πολύ πιθανόν, η σύγκλιση προς τη βέλτιστη λύση να ήταν πρακτικά αδύνατη.

Σε αυτήν την εργασία προτείνουμε μια υβριδική μέθοδο, ώστε να ξεπεραστούν τα προβλήματα του ΓΑ εξαιτίας της πολυπλοκότητας προβλημάτων από το χώρο της υπολογιστικής όρασης. Η γενική ιδέα της μεθοδολογίας που προτείνουμε σχετικά με την αντιμετώπιση του προβλήματος της επιλογής χαρακτηριστικών σε δεδομένα υψηλής διάστασης, είναι να διασπάσουμε τον πολυδιάστατο αρχικό χώρο του προβλήματος σε υποχώρους μικρότερης διάστασης και στη συνέχεια με το συνδυασμό ενός τροποποιημένου ΓΑ και της μεθόδου Simulated Annealing να επεξεργαστούμε παράλληλα αυτούς τους διανυσματικούς υποχώρους. Η συνεισφορά των ΓΑ έχει να κάνει με την παραλληλοποίηση της όλης διαδικασίας επιλογής χαρακτηριστικών, την διαμόρφωση και το συνδυασμό λύσεων χαμηλής διάστασης. Αντίστοιχα η μέθοδος Simulated Annealing με τη σειρά της επεξεργάζεται αντιπροσωπευτικά τμήματα των διανυσματικών υποχώρων και με βάση τους συνδυασμούς υποψήφιων λύσεων που δέχεται από το ΓΑ βρίσκει με αποδοτικό τρόπο την βέλτιστη λύση.

Στο Σχήμα 4.1 παρουσιάζεται το γενικό σχήμα αναπαράστασης της προτεινόμενης υβριδικής μεθόδου (Γενετικός αλγόριθμος & Simulated Annealing). Το αρχικό υποσύνολο χαρακτηριστικών χωρίζεται σε υποσύνολα χαρακτηριστικών (blocks), τα οποία αποτελούν εισόδους για το ΓΑ. Η αναζήτηση χωρίζεται σε δύο στάδια:

- Καθολικό: Αναζήτηση βέλτιστων συνδυασμών από blocks.
- Τοπικό: Αναζήτηση των σημαντικότερων χαρακτηριστικών (pixels) για κάθε block.

Τα μέλη του γενετικού πληθυσμού περνούν ως αρχικά σημεία στην τοπική μέθοδο αναζήτησης. Για τον τρόπο υλοποίησης της υβριδικής μεθόδου ακολουθήθηκαν δύο διαφορετικές στρατηγικές οι οποίες αναλύονται στη ενότητα στην ενότητα 4.5.



Σχήμα 4.1 Γενικό σχήμα υβριδικής μεθόδου

(Γενετικός αλγόριθμος & Simulated Annealing)

4.2 Ανάλυση/Επεξεργασία εικόνων

Η Ανάλυση της εικόνας λαμβάνει ως είσοδο ολόκληρη την περιοχή της εικόνας του αντικειμένου και εντοπίζει μη επικαλυπτόμενες περιοχές από pixels οι οποίες ορίζουν τα blocks της εικόνας. Στη συνέχεια γίνεται ο διαχωρισμός των pixels κάθε block από την υπόλοιπη εικόνα. Ο διαχωρισμός αυτός αποτελεί ένα σημαντικό στάδιο στην μεθοδολογία επίλυσης του προβλήματος γιατί παράγει υποπεριοχές της εικόνας με διάσταση μικρότερης της αρχικής και διαφορετικού πληροφοριακού περιεχομένου.

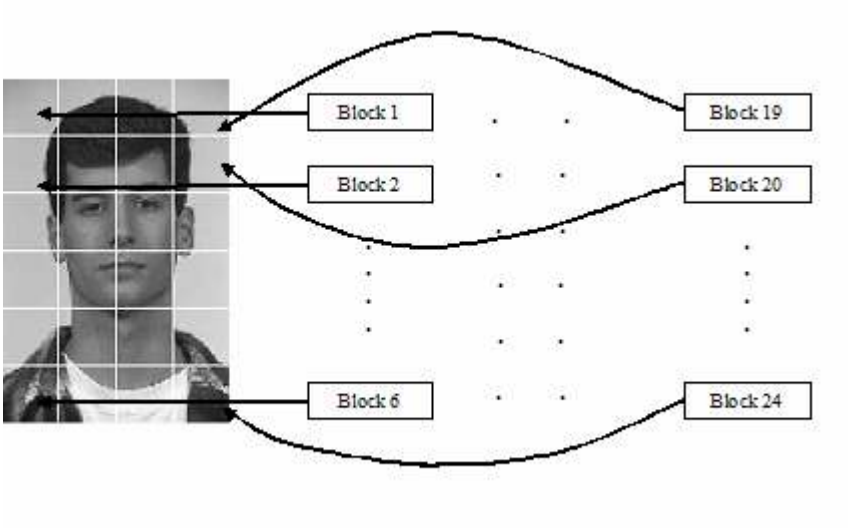
Αναλυτικότερα, κάθε εικόνα προσώπου διάστασης pxq pixels αναπαρίσταται από ένα διάνυσμα x_i διάστασης N, όπου N= pxq . Τα διανύσματα αποθηκεύονται σε έναν πίνακα nxN όπου n το πλήθος των εικόνων του συνόλου των παραδειγμάτων.

Στη συνέχεια ακολουθεί ο διαχωρισμός της εικόνας σε k μη επικαλυπτόμενα blocks μήκους q= N/k όπου N το πλήθος των pixels της κάθε εικόνας:

$$I = \{b_1, b_2, b_3, b_4, b_5, ..., b_k\}$$

Για κάθε block ορίζουμε το block $s_i(b_j)$ ως την ομάδα των q pixels του χώρου που ορίζει το block b_j στην εικόνα. Συνολικά παράγονται k blocks όπου k ο αριθμός των blocks και αποθηκεύονται σε ένα πίνακα k x q.

Στην Σχήμα 4.2 παρουσιάζεται ένα παράδειγμα διαχωρισμού της εικόνας σε 24 blocks.



Σχήμα 4.2 Διαχωρισμός της εικόνας σε k=24 μη επικαλυπτόμενα blocks.

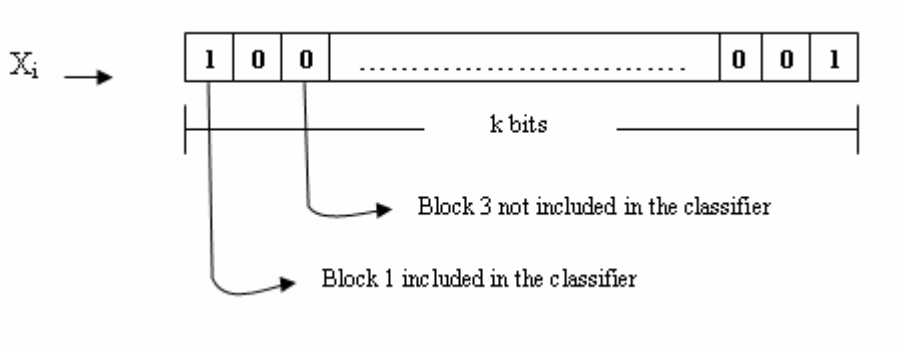
4.3 Βασικά στοιχεία υλοποίησης

Αναπαράσταση χρωμοσωμάτων Γενετικού πληθυσμού

Για την αναπαράσταση των μελών (χρωμοσωμάτων) του γενετικού πληθυσμού (υποψήφιων λύσεων στον γενετικό αλγόριθμο) επιλέχθηκε η δυαδική κωδικοποίηση.

Κάθε μέλος περιέχει μια ομάδα από blocks ή και ενδεχομένως όλα τα blocks ανάλογα με την αρχικοποίηση και αναπαρίσταται με δυαδική κωδικοποίηση ως ένα binary-string $X_i = (b_j, b_{j+1}, b_{j+2}, ..b_k)$ με j=(1,2,3,…k) όπου k ο αριθμός των blocks.

Η δυαδική τιμή κάθε θέσης του string δηλώνει αν τα features του $block_j$ θα λάβουν μέρος στην διαδικασία της αξιολόγησης του χρωμοσώματος X_i. Ένα παράδειγμα χρωμοσώματος φαίνεται στο Σχήμα 4.3.



Σχήμα 4.3 Αναπαράσταση χρωμοσώματος γενετικού πληθυσμού.

Δυαδική κωδικοποίηση για τα υποσύνολα χαρακτηριστικών

Για την αναπαράσταση των υποψήφιων λύσεων (υποψήφια υποσύνολα) στον αλγόριθμο τοπικής αναζήτησης Simulated Annealing επιλέχθηκαν δύο τρόποι κωδικοποίησης, η δυαδική (binary) κωδικοποίηση και η συνεχής (real) κωδικοποίηση.

Με βάση την δυαδική κωδικοποίηση ένα τυχαίο υποσύνολο των χαρακτηριστικών που αντιστοιχούν στο $block_i$ αναπαρίσταται ως μια ακολουθία q δυαδικών ψηφίων. Κάθε δυαδικό ψηφίο αναπαριστά ένα χαρακτηριστικό, με τις τιμές 1 και 0 να αντιστοιχούν στην επιλογή του ή όχι αντίστοιχα. Η αρχικοποίηση του πληθυσμού των blocks αφορά την δημιουργία k τυχαίων strings-blocks μήκους q, όπου q το μήκος των blocks.

Στην περίπτωση της δυαδικής κωδικοποίησης η διαδικασία επιλογής χαρακτηριστικών με ελεγχόμενη προσθήκη ή αφαίρεση χαρακτηριστικών παρουσίασε κάποια προβλήματα αυξάνοντας σε μερικές περιπτώσεις το μέγεθος των υποσυνόλων των χαρακτηριστικών. Αντίστοιχα η συνεχής κωδικοποίηση όπως περιγράφεται στην ενότητα 4.6 οδήγησε σε καλύτερα αποτελέσματα και είναι αυτή που ακολουθήθηκε.

Συνάρτηση Αξιολόγησης

Ο στόχος της επιλογής ενός υποσυνόλου χαρακτηριστικών είναι η όσο το δυνατό μεγαλύτερη απόδοση με το μικρότερο δυνατό αριθμό χαρακτηριστικών. Έτσι η συνάρτηση καταλληλότητας αποτελεί συνδυασμό 2 όρων: της ακρίβειας ταξινόμησης και του αριθμού των επιλεγμένων χαρακτηριστικών. Η ακρίβεια ταξινόμησης υπολογίζεται στον ταξινομητή SVM με χρήση της τεχνικής 10-fold cross validation. Κάθε υποσύνολο χαρακτηριστικών περιέχει ένα συγκεκριμένο αριθμό χαρακτηριστικών. Εάν δύο υποσύνολα εμφανίζουν την ίδια απόδοση τότε θα προτιμηθεί αυτό που περιέχει τα λιγότερα χαρακτηριστικά. Από τους δύο όρους της συνάρτησης αξιολόγησης αυτός που έχει μεγαλύτερη βαρύτητα είναι η ακρίβεια ταξινόμησης.

Η συνάρτηση αξιολόγησης είναι η ακόλουθη:

$$fitness = Fp \cdot Accuracy + (1 - Fp) \cdot (1 - (Nonzeros / Initial\ \dim ension))$$

όπου:

- Accuracy, η απόδοση του SVM για ένα συγκεκριμένο υποσύνολο χαρακτηριστικών.
- Fp, μια παράμετρος που καθορίζει τη βαρύτητα των δύο βασικών όρων της συνάρτησης αξιολόγησης. Μετά από πειραματικές μελέτες τέθηκε η τιμή της ίση με 0.8.
- Nonzeros, ο αριθμός των χαρακτηριστικών του υποεξέταση υποσυνόλου χαρακτηριστικών.
- Initial_dimension, ο αριθμός των χαρακτηριστικών στο ακατέργαστο σύνολο δεδομένων.

Με βάση την τιμή της παραμέτρου Fp φαίνεται καθαρά πως η απόδοση του ταξινομητή έχει το μεγαλύτερο βάρος στη συνάρτηση αξιολόγησης. Αυτό πρακτικά σημαίνει ότι για δύο υποσύνολα χαρακτηριστικών με μεγάλη διαφορά στην ακρίβεια ταξινόμησης δεν εξετάζεται στη ουσία ο αριθμός των χαρακτηριστικών τους.

4.4 Προτεινόμενος Υβριδικός Αλγόριθμος

Υβριδικός Αλγόριθμος

Αρχικοποιήσεις:

- Γενετικού πληθυσμού, P τυχαία δυαδικά χρωμοσώματα.
- Πληθυσμού k τυχαίων string-blocks.
- Γενετικών Τελεστών, αριθμού γενεών.

Στη συνέχεια επαναληπτικά μέχρι να ικανοποιηθεί η συνθήκη τερματισμού εκτελούνται τα ακόλουθα βήματα:

ΒΗΜΑ 1. Εύρεση των n ενεργών blocks (με $n \le k$) στα μέλη του γενετικού πληθυσμού της τρέχουσας γενιά t .

ΒΗΜΑ 2. Για κάθε ενεργό blocks_i επαναληπτικά:

- Εύρεση των L μελών (με $L \le P$) του γενετικού πληθυσμού της τρέχουσας γενιάς t τα οποία που έχουν ενεργό το block_i.
- Κλήση μεθόδου τοπικής αναζήτησης με είσοδο τα L μέλη του γενετικού πληθυσμού και το τρέχον ενεργό string-block_i της τρέχουσας γενιάς.

Αν για όλα τα ενεργά blocks έχει κληθεί η μέθοδος τοπικής αναζήτησης τότε θα συνεχίσουμε σειριακά με την εφαρμογή των διαδικασιών της επιλογής, διασταύρωσης και μετάλλαξης

ΒΗΜΑ 3. Υπολογισμός Fitness Score για τα κάθε χρωμόσωμα $X_i \in P$.

ΒΗΜΑ 4. Επιλογή γενετικού πληθυσμού P(t+1) από τον P(t).

ΒΗΜΑ 5. Διασταύρωση μελών γενετικού πληθυσμού P(t+1).

ΒΗΜΑ 6. Μετάλλαξη.

Τοπική μέθοδος αναζήτησης

Ξεκινώντας από ένα αρχικό υποσύνολο χαρακτηριστικών του υποεξέταση $block_i$ εκτελείται μέσω της μεθόδου προσομοιωμένης ανόπτησης (Simulated Annealing) μια επαναληπτική διαδικασία (M φορές) βελτιστοποίησης της απόδοσης του συνόλου των χαρακτηριστικών του $block_i$ και των L μελών (με $L \leq P$) του γενετικού πληθυσμού τα οποία έχουν ενεργό το $block_i$ στην τρέχουσα γενιά t .
Σε κάθε επανάληψη της τοπικής μεθόδου:

ΒΗΜΑ 1. Τρέχον υποσύνολο χαρακτηριστικών: τρέχον $block_i$.
ΒΗΜΑ 2. Υπολογισμός της καταλληλότητας (fitness) των X_i μελών του γενετικού πληθυσμού με i={1,2,.. L}.
ΒΗΜΑ 3. Τρέχουσα λύση: η μέγιστη τιμή από τις καταλληλότητες που υπολογίστηκαν στο βήμα 2 .
ΒΗΜΑ 4. Μετάβαση σε νέο υποσύνολο χαρακτηριστικών για το $block_i$.
ΒΗΜΑ 5. Επανάληψη του βήματος 2.
ΒΗΜΑ 6. Νέα λύση: η μέγιστη τιμή από τις καταλληλότητες των μελών του γενετικού πληθυσμού.
ΒΗΜΑ 7. Αποδοχή ή όχι του νέου υποσύνολου χαρακτηριστικών με βάση τη διαδικασία αποδοχής νέων λύσεων όπως ορίζεται στη μέθοδο Simulated Annealing.

Για κάθε X_i με i={1,2,.. L} μέλος του γενετικού πληθυσμού υπολογίζεται η συνάρτηση καταλληλότητας (fitness score) με βάση ένα σύνολο δεδομένων που δημιουργήθηκε: από το υποσύνολο χαρακτηριστικών του $block_i$ και τα χαρακτηριστικά των υπολοίπων ενεργών blocks του X_i. Τα χαρακτηριστικά των υπολοίπων blocks που μετέχουν στη διαδικασία αξιολόγησης παραμένουν σταθερά σε όλες τις επαναλήψεις της τοπικής μεθόδου για το τρέχον block.
Για κάθε $block_i$ που βελτιστοποιήθηκε στη φάση της τοπικής αναζήτησης αποθηκεύεται το καλύτερο στιγμιότυπό του, από τα M διαφορετικά υποσύνολα χαρακτηριστικών που αξιολογήθηκαν. Το καλύτερο υποσύνολο χαρακτηριστικών είναι αυτό που ανέδειξε το καλύτερο μέλος του γενετικού πληθυσμού σύμφωνα με τη συνάρτηση καταλληλότητας.

4.5 Προσεγγίσεις υλοποίησης

Στα πλαίσια της εργασίας δοκιμάστηκαν 2 μεθοδολογίες/στρατηγικές υλοποίησης της υβριδικής μεθόδου.

1η στρατηγική υλοποίησης της υβριδικής μεθόδου

Σύμφωνα με αυτήν την στρατηγική κάθε μέλος του πληθυσμού του γενετικού αλγορίθμου έχει τη δική του ομάδα από blocks. Έστω για παράδειγμα το μέλος του γενετικού πληθυσμού X_i. Αν έχουμε χωρίσει τις εικόνες σε k blocks, τότε στο X_i. αντιστοιχεί μια ομάδα από k blocks . Η κλήση της μεθόδου τοπικής αναζήτησης από το X_i έχει ως αποτέλεσμα την βελτιστοποίηση της αντίστοιχης ομάδας του των blocks. Κατά την εφαρμογή των τελεστών της επιλογής και διασταύρωσης στα μέλη του γενετικού πληθυσμού πραγματοποιείται και ανταλλαγή των αντίστοιχων blocks ενός γονέα στο παιδί του, γίνεται δηλαδή μεταφορά γενετικής πληροφορίας στο χαμηλότερο επίπεδο, αυτό των περιεχομένων των blocks.

2η στρατηγική υλοποίησης της υβριδικής μεθόδου

Για κάθε μέλος του πληθυσμού του γενετικού αλγορίθμου υπάρχει μια κοινή ομάδα από blocks. Αν για παράδειγμα έχουμε χωρίσει τις εικόνες σε k blocks, τότε έχει αρχικοποιηθεί μια ομάδα από k blocks. Κάθε μέλος του πληθυσμού του γενετικού αλγορίθμου χρησιμοποιεί αυτήν την ομάδα της οποίας τα μέλη σε κάθε εκτέλεση της μεθόδου τοπικής αναζήτησης βελτιστοποιούνται σύμφωνα με το κριτήριο αξιολόγησης. Σε κάθε επόμενη γενιά του αλγορίθμου λαμβάνουν μέρος τα καλύτερα υποσύνολα χαρακτηριστικών των blocks έτσι όπως προέκυψαν από προηγούμενες γενιές.

Στα πλαίσια της υλοποίησης ακολουθήθηκε η 2η στρατηγική γιατί παρουσίαζε καλύτερα αποτελέσματα με μικρότερο υπολογιστικό κόστος.

4.6 Συνεχής κωδικοποίηση (real coding)

Ο μηχανισμός επιλογής χαρακτηριστικών στη φάση της τοπικής αναζήτησης βασίστηκε σε μια στοχαστική διαδικασία σύμφωνα με την οποία κάθε χαρακτηριστικό μπορεί να επιλεγεί ή όχι στη διαδικασία της αξιολόγησης με βάση κάποια πιθανότητα p.

Στην περίπτωση της συνεχούς κωδικοποίησης τα υποσύνολα χαρακτηριστικών αναπαρίσταται ως μια ακολουθία (string) q πραγματικών αριθμών στο διάστημα [0,1]. Οι αριθμοί αυτοί ορίζουν την πιθανότητα με την οποία το χαρακτηριστικό της αντίστοιχης θέσης του string θα επιλεγεί ή όχι. Με βάση αυτήν την προσέγγιση κατά τη φάση της τοπικής αναζήτησης η συνάρτηση μετάβασης σε νέα υποσύνολα χαρακτηριστικών επιλέγοντας ένα χαρακτηριστικό αυξάνει ή μειώνει ανάλογα και την τιμή της αντίστοιχης πιθανότητας του χαρακτηριστικού.

Σε αυτήν την περίπτωση κωδικοποίησης η διαδικασία επιλογής των χαρακτηριστικών για την δημιουργία υποψηφίων λύσεων του προβλήματος (υποσύνολα χαρακτηριστικών) με ελεγχόμενη την προσθήκη ή αφαίρεση χαρακτηριστικών είχε ως αποτέλεσμα τον περιορισμό του μεγέθους των υποσυνόλων οδηγώντας τον αλγόριθμο στην εύρεση βέλτιστων λύσεων με μικρό αριθμό χαρακτηριστικών. Η επιλογή κάθε χαρακτηριστικού παραμένει σε μεγαλύτερο βαθμό ανεπηρέαστη από την αρχική αρχικοποίηση δίνοντας την ευκαιρία σε περισσότερα χαρακτηριστικά να αποτελέσουν μέρος μιας υποψήφιας λύσης και να αξιολογηθούν ως προς τη χρησιμότητά τους.

ΚΕΦΑΛΑΙΟ 5. ΠΕΙΡΑΜΑΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ

Οι στόχοι της πειραματικής αξιολόγησης είναι να εξεταστεί η απόδοση της προτεινόμενης μεθόδου επιλογής χαρακτηριστικών σε εικόνες προσώπων υψηλής διάστασης και επιπλέον να συγκριθεί η απόδοσή της με άλλες μεθόδους επιλογής χαρακτηριστικών. Για τους πειραματισμούς χρησιμοποιήθηκαν 4 ευρέως χρησιμοποιούμενες βάσεις εικόνων προσώπων (face image databases) οι όποιες περιγράφονται παρακάτω. Τα πειράματα υλοποιήθηκαν μέσω του περιβάλλοντος MATLAB και της ανοικτής πλατφόρμας αλγορίθμων μηχανικής μάθησης WEKA (Waikato Environment for Knowledge Analysis). Για την ταξινόμηση χρησιμοποιήθηκε η βιβλιοθήκη LIBSVM.

5.1 Πειραματική μεθοδολογία

Για κάθε βάση δεδομένων προσώπων δημιουργήθηκε και ένα πειραματικό σύνολο δεδομένων. Αναλυτικότερα σε κάθε σύνολο δεδομένων, κάθε εικόνα προσώπου διάστασης *pxq* pixels αναπαρίσταται από ένα διάνυσμα x_i διάστασης N, όπου N= *pxq* . Τα διανύσματα αποθηκεύονται σε έναν πίνακα *nxN* όπου n το πλήθος των εικόνων του συνόλου των παραδειγμάτων.

Στη συνέχεια οι εικόνες των προσώπων και στα 4 σύνολα δεδομένων διαχωρίστηκαν σε k μη επικαλυπτόμενες περιοχές blocks με k={16, 25 ,36, 64, 121} μήκους q= N/k όπου N το πλήθος των pixels της κάθε εικόνα.

Για κάθε διαφορετική τιμή του k έγιναν 10 πειραματικές επαναλήψεις σε κάθε σύνολο δεδομένων με σκοπό την εξασφάλιση όσο το δυνατόν μεγαλύτερης αξιοπιστίας στην απόδοση της προτεινόμενης μεθοδολογίας. Τελικά αποτελέσματα θεωρούνται οι μέσοι όροι του αριθμού των χαρακτηριστικών των επιλεγμένων υποσυνόλων και των αντίστοιχων αποδόσεων τους και για τις 10 επαναλήψεις. Στα πειράματα χρησιμοποιήθηκε στον SVM ο γραμμικός πυρήνας (linear kernel).

Το μέγεθος του γενετικού πληθυσμού εξαρτάται από τον αριθμό των blocks. Για 16, 25, 36, 64, 121 blocks το μέγεθος του πληθυσμού είναι 20, 30, 40, 60, 120 μέλη αντίστοιχα. Η πιθανότητα διασταύρωσης τέθηκε μετά από πειραματισμούς ίση με 0.8 ενώ η πιθανότητα μετάλλαξης τέθηκε ίση με 0.09. Η αρχικοποίηση του γενετικού πληθυσμού έγινε με τυχαία κατανομή των 0 και 1. Ο αριθμός των επαναλήψεων στο Γενετικό αλγόριθμο τέθηκε ίσος με 20, ενώ ο αριθμός των επαναλήψεων στην τοπική μέθοδο Simulated Annealing είναι ανάλογος του μεγέθους των blocks.

Τέλος, για όλα τα σύνολα δεδομένων συγκρίναμε τα αποτελέσματα της δικής μας μεθοδολογίας με τα αποτελέσματα των μεθόδων επιλογής χαρακτηρισμών του WEKA: Relief-F, Gain Ratio και FilteredSubsetEval. Η μέθοδος Relief-F στηρίζεται στον αλγόριθμο RELIEF-F που αναφέρθηκε στο κεφάλαιο 2. Η μέθοδος Gain Ratio χρησιμοποιεί το κριτήριο gain ratio που αποτελεί τροποποίηση του μεγέθους του κέρδους της πληροφορίας (information gain).
Οι μέθοδοι Relief-F και Gain Ratio ανήκουν στην κατηγορία φίλτρων που κάνουν μοναδιαία αποτίμηση χαρακτηριστικού, δηλαδή αξιολογούν κάθε χαρακτηριστικό ξεχωριστά με βάση τη συσχέτιση του με την κατηγορία.
Η κλάση αποτίμησης υποσυνόλων χαρακτηριστικών FilteredSubsetEval αποτελεί συνδυασμό:

- Φίλτρου τυχαίας δειγματοληψίας υποσυνόλων χαρακτηριστικών στο σύνολο δεδομένων.

- Φίλτρου αποτίμησης της αξίας ενός υποσυνόλου χαρακτηριστικών με βάση την μεγαλύτερη συνάφεια των χαρακτηριστικών με την κατηγορία (προβλεπτική ικανότητα) και το μεγαλύτερο βαθμό ανομοιότητας μεταξύ τους. Η μέθοδος απορρίπτει υποσύνολα χαρακτηριστικών που έχουν μεγάλη συσχέτιση μεταξύ τους [26].
- Μεθόδου αναζήτησης με βάση τον αλγόριθμο αναρρίχησης λόφων (hill climbing).

5.2 Πειραματικά σύνολα εικόνων

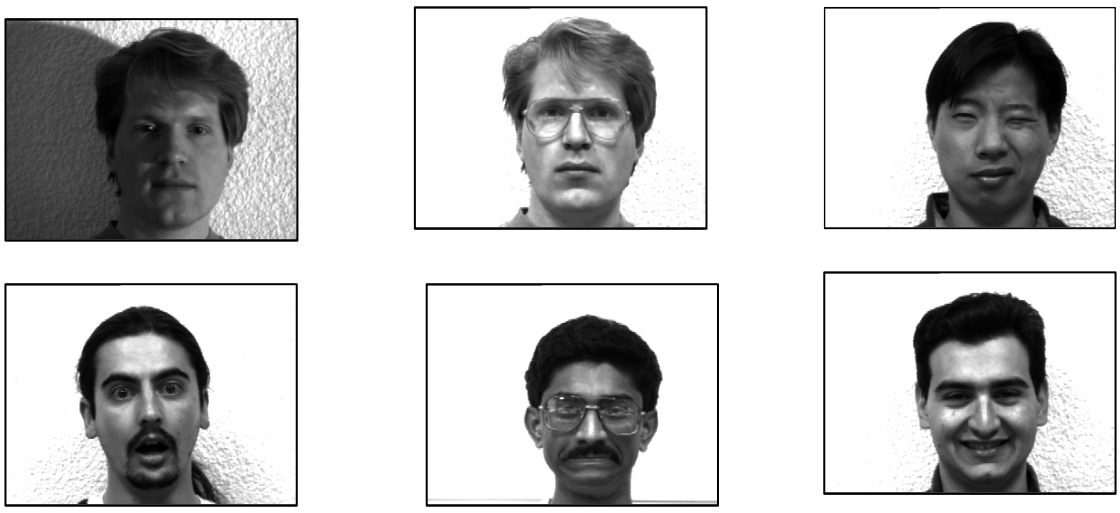
Η Βάση εικόνων προσώπου ORL

Η βάση ORL [28] είναι διαθέσιμη από το AT & T Laboratories Cambridge και αποτελείται συνολικά από 400 ασπρόμαυρες εικόνες διαστάσεων 92x112 με 256 επίπεδα γκρι ανά pixel. Οι εικόνες ανήκουν σε 40 διαφορετικά άτομα με 10 εικόνες ανά άτομο. Τα πρόσωπα έχουν φωτογραφηθεί σε όρθια στάση και σε πρόσθιες όψεις με διάφορες παραλλαγές σε: συνθήκες φωτισμού, εκφράσεις προσώπου, ανοικτά κλειστά μάτια, χαμόγελο ή όχι, γυαλιά ή όχι. Η εικόνα 5.1 δείχνει εικόνες από τη βάση ORL.

Εικόνα 5.1 Εικόνες προσώπων από τη βάση ORL.

Η Βάση εικόνων προσώπου Yale Face Database

Η Βάση Yale Face είναι διαθέσιμη από το Yale University Department of Computer Science [29] και αποτελείται από 165 grayscale εικόνες διαστάσεων 320 x 243. Οι εικόνες ανήκουν σε 15 άτομα με 11 εικόνες ανά άτομο, μια για κάθε διαφορετική έκφραση ή παραλλαγή: center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised and wink. Για υπολογιστικές ανάγκες μειώθηκε η διάσταση των εικόνων σε 84x110. Η εικόνα 5.2 δείχνει κάποια στιγμιότυπα από τη βάση Yale.



Εικόνα 5.2 Εικόνες προσώπων από τη βάση Yale.

Η Βάση εικόνων προσώπου FERET Database

Η βάση δημιουργήθηκε από το «Defense Advanced Research Products Agency (DARPA)» [30] για την ανάπτυξη αυτοματοποιημένων μεθόδων αναγνώρισης προσώπων με σκοπό τη βελτίωση θεμάτων ασφαλείας, την ανάπτυξη συστημάτων ευφυΐας και τήρησης του νόμου. Περιέχει 14.051 grayscale εικόνες που απεικονίζουν πρόσωπα με εμπρόσθιες, αριστερές και δεξιές κατόψεις με διάφορες παραλλαγές σε: συνθήκες φωτισμού, εκφράσεις προσώπου, ανοικτά κλειστά μάτια, χαμόγελο ή όχι,

γυαλιά ή όχι. Επίσης φωτογραφίες του ίδιου προσώπου τραβηχτήκαν με διαφορά μεγαλύτερη του ενός έτους. Από τη βάση FERET δημιουργήθηκε ένα σύνολο από 440 grayscale εικόνες διαστάσεων 256x384. Οι εικόνες ανήκουν σε 40 άτομα με 11 εικόνες ανά άτομο. Για υπολογιστικές ανάγκες μειώθηκε η διάσταση των εικόνων σε 120x80. Η εικόνα 5.3 δείχνει κάποια στιγμιότυπα από τη βάση FERET.

Εικόνα 5.3 Εικόνες προσώπων από τη βάση FERET.

Η Βάση εικόνων προσώπου Face Recognition Data, University of Essex, UK

Η Βάση εικόνων προσώπου Face Recognition Data, University of Essex είναι διαθέσιμη από το University of Essex [31]. Η βάση περιέχει 7900 εικόνες 24bit RGB διαστάσεων 180x200. Η ηλικία των ατόμων που συμμετέχουν είναι κυρίως από 18-20 ετών με διάφορες παραλλαγές σε γυαλιά και γενειάδα ενώ ο φωτισμός των εικόνων είναι τεχνητός με μίγμα βολφραμίου και φθοριζόντων ουσιών. Από την συγκεκριμένη βάση δημιουργήθηκε ένα σύνολο από 400 εικόνες οι οποίες μετατράπηκαν σε κλίμακα του γκρι (grayscale μορφή). Οι εικόνες ανήκουν σε 20 άτομα με 20 εικόνες ανά άτομο. Για υπολογιστικές ανάγκες οι μειώθηκε η διάσταση των εικόνων σε 90x100. Η εικόνα 5.4 δείχνει κάποια στιγμιότυπα από τη βάση Essex.

Εικόνα 5.4 Εικόνες προσώπων από τη βάση Essex.

5.3 Πειραματικά αποτελέσματα και συγκρίσεις

Αρχικά στον Πίνακα 5.1 δίνεται μια συνοπτική περιγραφή για τα διαθέσιμα σύνολα δεδομένων καθώς και η αντίστοιχη ακρίβεια ταξινόμησης του ταξινομητή SVM με χρήση όλων των χαρακτηριστικών.

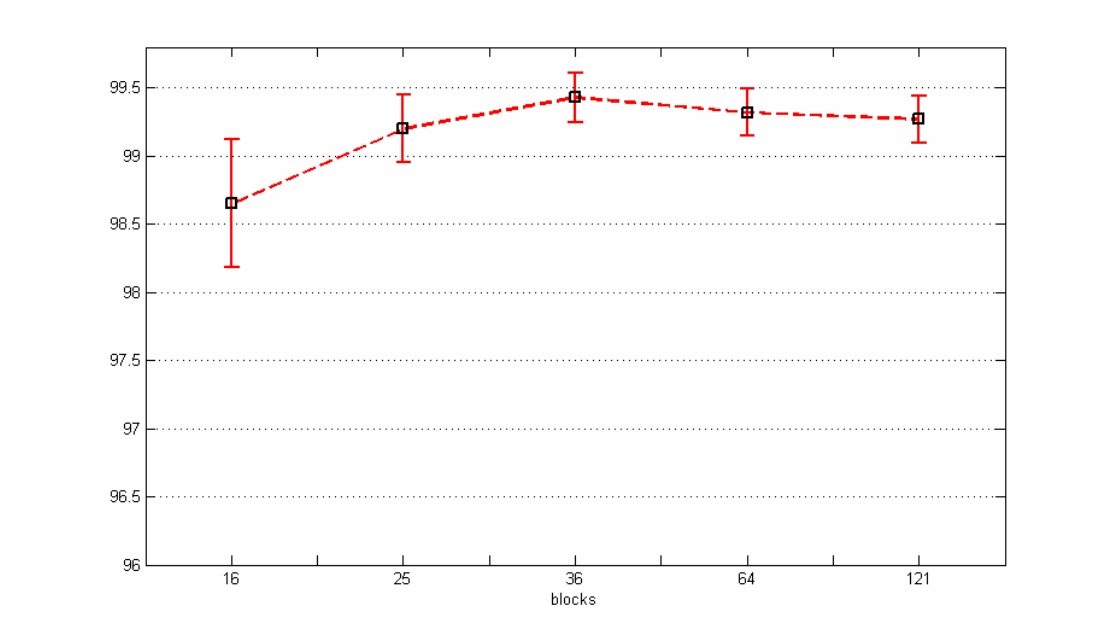
Πίνακας 5.1 Συνοπτική περιγραφή των διαθέσιμων συνόλων και της αρχικής απόδοσης του ταξινομητή SVM

Σύνολο δεδομένων	Αριθμός παραδειγμάτων	Αριθμός χαρακτηριστικών	Αριθμός κατηγοριών	Απόδοση SVM
Esssex faces	400	9000	20	93,25
ORL faces	400	10304	40	97,75
Yale Faces	165	9240	11	87,27
FERET faces	440	9600	40	94,5

Για όλα τα σύνολα δεδομένων και για κάθε μέγεθος block (16,25,36,121) υπολογίστηκαν για τη συνάρτηση καταλληλότητας (fitness function), την απόδοση

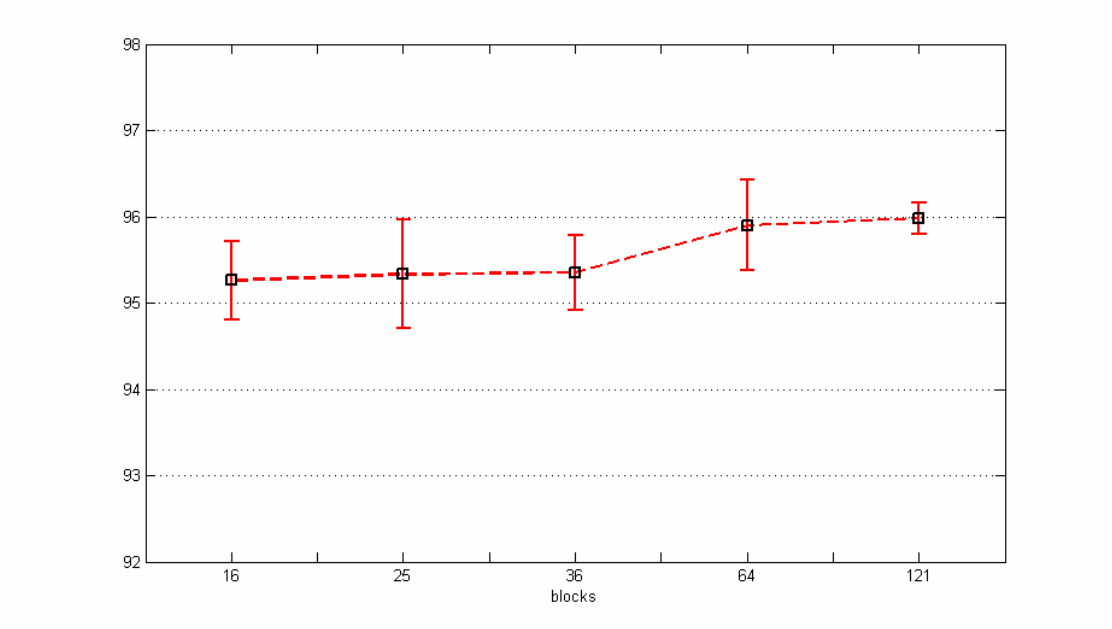
του ταξινομητή SVM και τον αριθμό των επιλεγμένων χαρακτηριστικών: η μέση τιμή (mean), η τυπική απόκλιση (standard deviation), η μέγιστη τιμή (max) και η ελάχιστη τιμή (min) αντίστοιχα.

Ξεκινώντας με τους πειραματισμούς στο σύνολο δεδομένων της βάσης ORL παρατηρούμε στον Πίνακα 5.2 ότι από την εφαρμογή της προτεινομένης μεθόδου προέκυψε σε κάθε διαφορετική περίπτωση αριθμού blocks μια σημαντική μείωση της διάστασης με παράλληλη αύξηση της απόδοσης του ταξινομητή από την αρχική. Η απόδοση του ταξινομητή στο αρχικό σύνολο δεδομένων με 10304 χαρακτηριστικών είναι 97,75% ενώ η προτεινόμενη μέθοδος πέτυχε μεγαλύτερη απόδοση με λιγότερα από 150 χαρακτηριστικά. Όπως παρατηρούμε στο Σχήμα 5.1 η ομάδα των 121 blocks μαζί με την ομάδα των 61 blocks παρουσιάζουν τη μικρότερη διακύμανση στις τιμές της συνάρτησης καταλληλότητας (fitness) με τυπική απόκλιση ± 0.17. Την μεγαλύτερη διακύμανση στις τιμές της συνάρτησης καταλληλότητας παρουσιάζει η ομάδα των 16 blocks. Σχετικά με τον αριθμό των χαρακτηριστικών η μικρότερη διακύμανση τιμών παρουσιάζεται για την ομάδα των 121 blocks.



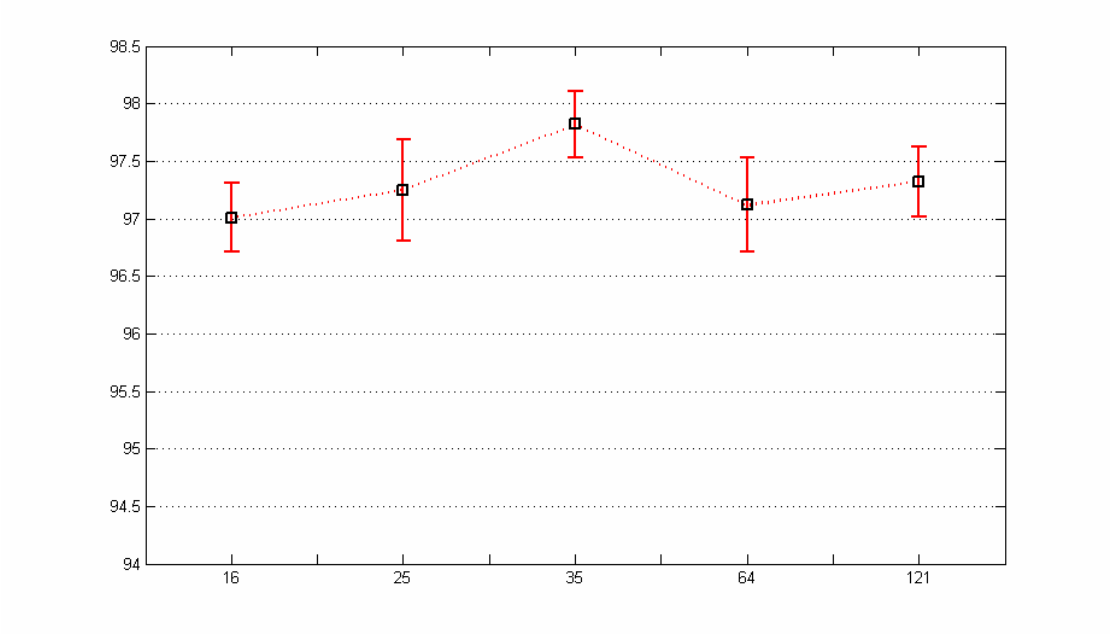
Σχήμα 5.1 Γράφημα μέσων όρων (means) και τυπικών αποκλίσεων (standard deviations) της απόδοσης (fitness) για όλα τα blocks του συνόλου ORL. Διακρίνονται τα άκρα του διαστήματος εμπιστοσύνης (confidence interval) για κάθε block.

Συνεχίζοντας στο σύνολο δεδομένων της βάσης Yale όπως παρατηρούμε στον Πίνακα 5.3 βελτιώθηκε η απόδοση του ταξινομητή η οποία ήταν 87,27% και με τα 9240 χαρακτηριστικά. Για όλες τις διαφορετικές περιπτώσεις αριθμού blocks η προτεινόμενη μέθοδος πετυχαίνει απόδοση μεγαλύτερη από 94,5%. με λιγότερα από 201 χαρακτηριστικά. Στο Σχήμα 5.2 παρατηρούμε ότι η ομάδα των 121 blocks παρουσιάζει τη μικρότερη διακύμανση στις τιμές της συνάρτησης καταλληλότητας με τυπική απόκλιση ± 0.18 .



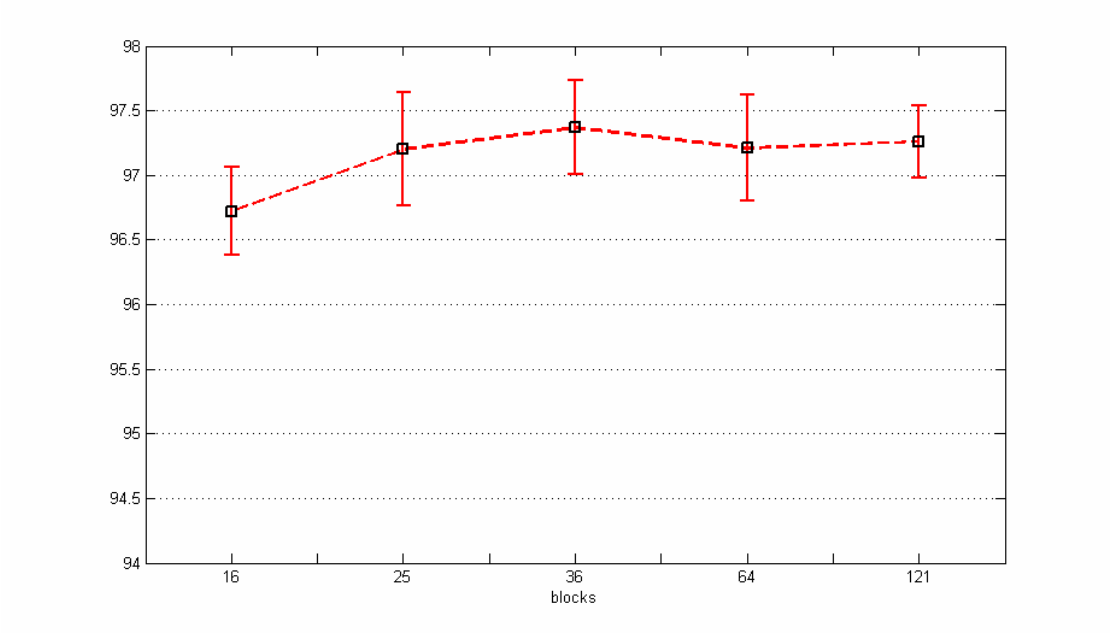
Σχήμα 5.2 Γράφημα μέσων όρων (means) και τυπικών αποκλίσεων (standard deviations) της απόδοσης (fitness) για όλα τα blocks του συνόλου Yale. Διακρίνονται τα άκρα του διαστήματος εμπιστοσύνης (confidence interval) για κάθε block.

Στο σύνολο δεδομένων Face Recognition Data University of Essex η απόδοση της προτεινόμενης μεθόδου είναι αρκετά υψηλή αφού με πολύ λιγότερα χαρακτηριστικά πετυχαίνει μεγαλύτερη ακρίβεια. Και με τα 9000 χαρακτηριστικά ο ταξινομητής αρχικά έχει απόδοση 93,25%, ενώ για όλες τις διαφορετικές περιπτώσεις αριθμού blocks όπως παρουσιάζονται στον Πίνακα 5.4 η προτεινόμενη μέθοδος πετυχαίνει με λιγότερα από 210 χαρακτηριστικά ακρίβεια ταξινόμησης μεγαλύτερη από 96%.



Σχήμα 5.3 Γράφημα μέσων όρων (means) και τυπικών αποκλίσεων (standard deviations) της απόδοσης (fitness) για όλα τα blocks του συνόλου Essex. Διακρίνονται τα άκρα του διαστήματος εμπιστοσύνης (confidence interval) για κάθε block.

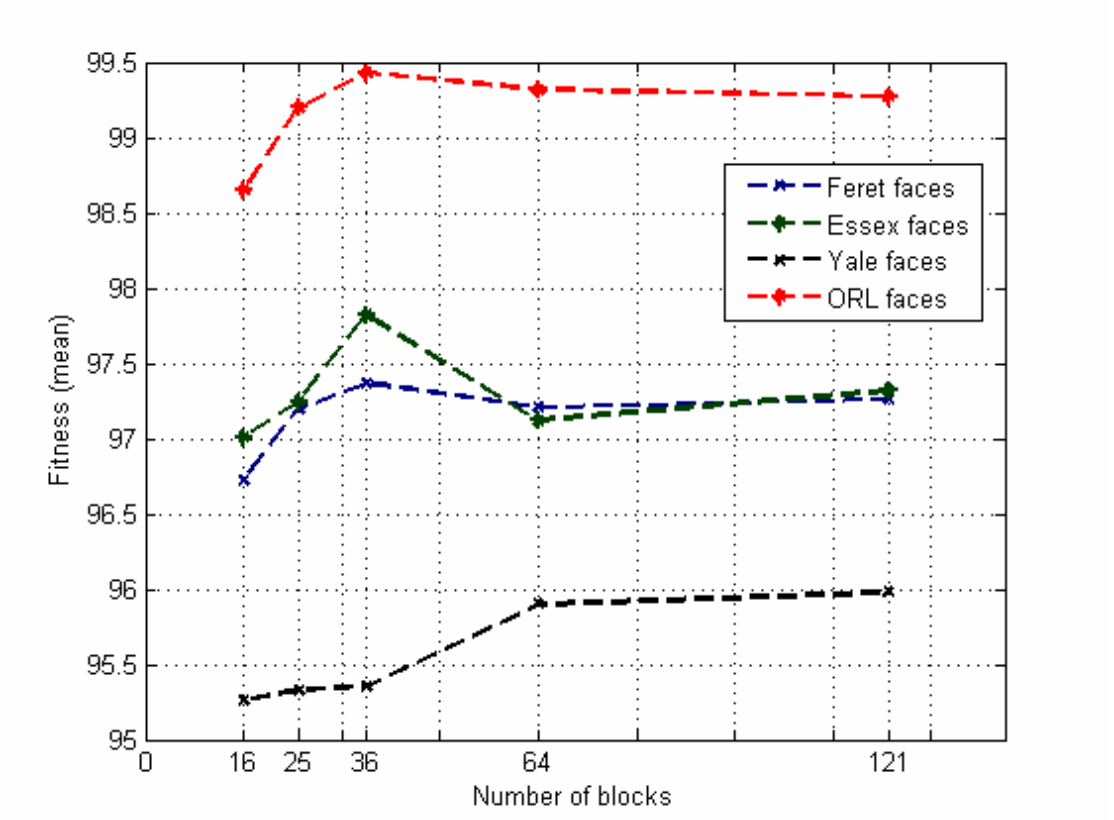
Τέλος στο σύνολο δεδομένων της βάσης FERET η προτεινόμενη μέθοδος και εδώ παρουσιάζει καλά αποτελέσματα αφού όπως παρατηρούμε στον Πίνακα 5.5 για όλες τις διαφορετικές περιπτώσεις αριθμού blocks πετυχαίνει με λιγότερα από 215 χαρακτηριστικά απόδοση μεγαλύτερη από 96,5%. Σημειώνεται ότι ο ταξινομητής αρχικά και με τα 9600 χαρακτηριστικά έχει απόδοση 94,5%.



Σχήμα 5.4 Γράφημα μέσων όρων (means) και τυπικών αποκλίσεων (standard deviations) της απόδοσης (fitness) για όλα τα blocks του συνόλου FERET. Διακρίνονται τα άκρα του διαστήματος εμπιστοσύνης (confidence interval) για κάθε block.

Στα Σχήματα 5.1, 5.2, 5.3 και 5.4 παρουσιάστηκαν για κάθε σύνολο δεδομένων με την βοήθεια του γραφήματος error bar το διάστημα εμπιστοσύνης (95% confidence interval) για κάθε ομάδα blocks. Τα γραφήματα για κάθε σύνολο δεδομένων δείχνουν ότι σε αυτό το διάστημα θα βρίσκεται με πιθανότητα 0,95 η μέση τιμή της απόδοσης (fitness) για κάθε ομάδα blocks σε κάθε περίπτωση. Το διάστημα αυτό για την περίπτωση των 121 blocks είναι μικρότερο, αφού η τυπική απόκλιση είναι μικρότερη σε αυτή την περίπτωση σε σχέση με τις υπόλοιπες ομάδες blocks.

Παρατηρώντας στο Σχήμα 5.5 τα αποτελέσματα για όλα τα σύνολα δεδομένα μπορούμε να πούμε ότι σε γενικές γραμμές τα καλύτερα αποτελέσματα παρουσιάζονται για αριθμό blocks μεγαλύτερο του 25. Η ερμηνεία που μπορεί να δοθεί σε αυτό το γεγονός είναι ότι καθώς έχουμε διαχωρισμό των εικόνων σε περισσότερα τμήματα διευκολύνεται ο αλγόριθμος ώστε να συνδυάσει και να ανακαλύψει χαρακτηριστικά με πιο αποδοτικό τρόπο.



Σχήμα 5.5 Σύγκριση του μέσου όρου της συνάρτησης καταλληλότητας (fitness) για όλα τα μεγέθη των blocks {16,25,36,64,121 } στα διαθέσιμα σύνολα δεδομένων.

Πίνακας 5.2 Αξιολόγηση συνόλου δεδομένων της βάσης ORL

Blocks	Fitness				Features				SVM accuracy			
	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min
16 (28x33)	98.65	± 0.47	99.36	98.02	142	± 38	205	93	98.66	± 0.53	99.5	98
25 (22x18)	99.2	± 0.25	99.53	98.82	150	± 36	197	101	99.37	± 0.27	99.75	99
36 (18x15)	99.43	± 0.18	99.67	99.16	148	± 38	201	98	99.65	± 0.21	99.9	99.23
64 (14x11)	99.32	± 0.17	99.52	99.08	150	± 37	185	101	99.65	± 0.24	99.7	99.52
121(10x8)	99.27	± 0.17	99.4	99.99	148	± 33	191	101	99.45	± 0.22	99.7	99.1

Πίνακας 5.3 Αξιολόγηση συνόλου δεδομένων της βάσης Yale

Blocks	Fitness				Features				SVM accuracy			
	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min
16 (21x27)	95.26	± 0.45	95.67	94.33	174	± 25	210	122	94.55	± 0.62	95.15	93.4
25 (16x22)	95.33	± 0.63	96.35	94.64	201	± 34	288	170	94.96	± 0.79	95.9	93.9
36 (14x18)	95.35	± 0.44	96.08	94.76	192	± 10	212	175	94.99	± 0.54	95.61	94
64 (10x13)	95.9	± 0.52	96.4	95.22	172	± 21	193	145	95.34	± 0.66	96	94.45
121(7x10)	95.98	± 0.18	96.23	95.8	171	± 17	187	146	95.43	± 0.24	95.8	95.2

Πίνακας 5.4 Αξιολόγηση συνόλου δεδομένων της βάσης Face Recognition Data University of Essex

Blocks	Fitness				Features				SVM accuracy			
	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min
16 (22x25)	97.01	± 0.3	97.51	96.42	168	± 37	290	168	96.81	± 0.4	97.5	96
25 (18x20)	97.25	± 0.44	97.98	96.42	205	± 47	290	156	97.13	± 0.59	98	96
36 (15x16)	97.82	± 0.29	98.25	97.40	198	± 26	232	144	97.82	± 0.34	98.2	97.3
64 (11x12)	97.12	± 0.41	97.4	96.38	199	± 11	211	183	96.9	± 0.53	97.3	96
121(8x9)	97.32	± 0.3	97.69	96.85	180	± 25	206	145	97.13	± 0.33	97.5	96.6

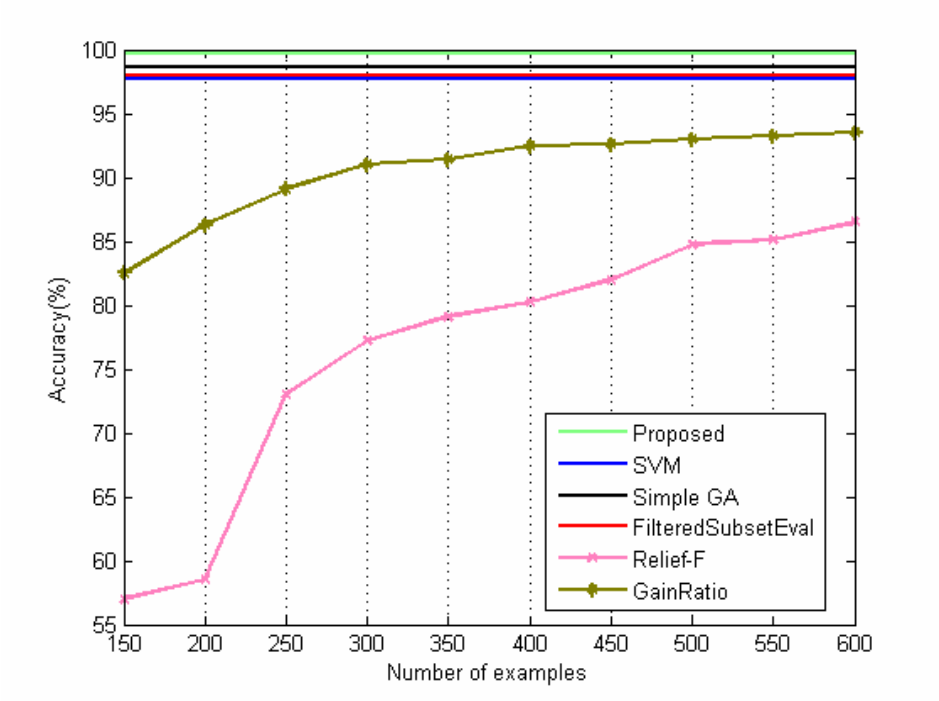
Πίνακας 5.5 Αξιολόγηση συνόλου δεδομένων της βάσης FERET

Blocks	Fitness				Features				SVM accuracy			
	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min	Mean	± *Std*	Max	Min
16 (30x20)	96.72	± 0.34	97.45	96.39	215	± 23	249	190	96.46	± 0.44	97.45	96
25 (24x16)	97.2	± 0.44	97.77	96.4	198	± 29	245	150	97.01	± 0.60	98	95.9
36 (20x13)	97.37	± 0.36	97.92	96.68	214	± 20	252	186	97.24	± 0.43	98	96.5
64 (15x10)	97.21	± 0.41	97.9	96.76	195	± 6	201	189	97	± 0.24	97.9	96.45
121(10x7)	97.26	± 0.28	97.53	96.83	181	± 4	185	176	97	± 0.35	97.4	96.5

5.4 Σύγκριση τεχνικών επιλογής χαρακτηριστικών

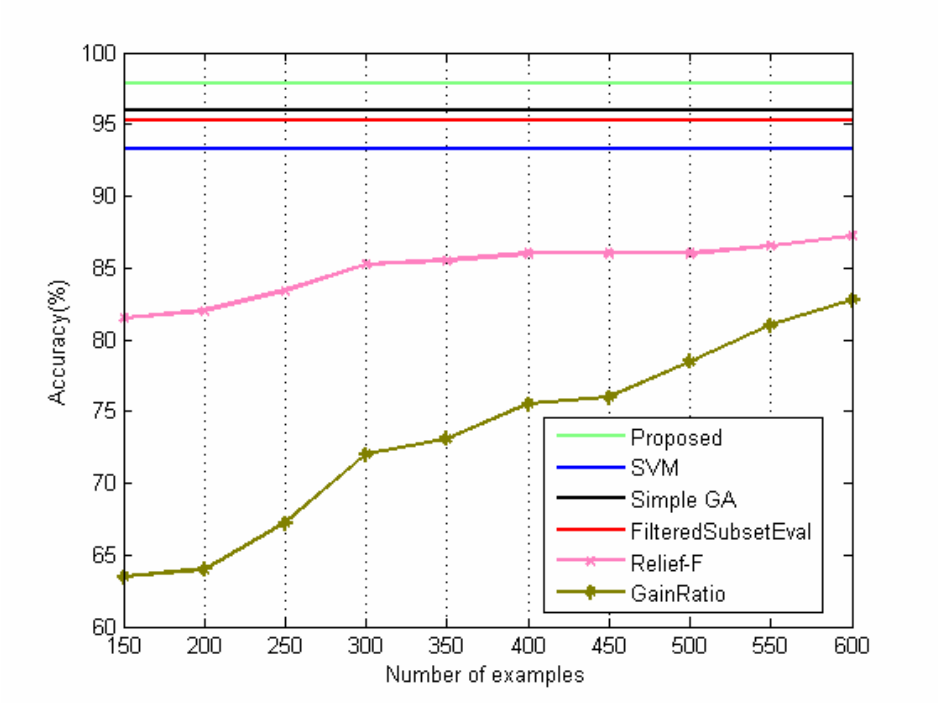
Σε αυτήν την ενότητα συγκρίνεται η απόδοση της προτεινόμενης μεθόδου με τις αποδόσεις των μεθόδων Gain Ratio, Relief-f και FilteredSubsetEval, την απόδοση ενός απλού Γενετικού αλγορίθμου που υλοποιήθηκε στα πλαίσια της εργασίας καθώς την απόδοση του ταξινομητή SVM στα αρχικά σύνολα δεδομένων με όλα τα χαρακτηριστικά.

Στην περίπτωση του συνόλου δεδομένων της βάσης ORL όπως παρατηρούμε στο Σχήμα 5.6, η προτεινόμενη μέθοδος παρουσιάζει την καλύτερη απόδοση 99,65% με 148 επιλεγμένα χαρακτηριστικά. Παρατηρούμε επίσης ότι οι μέθοδοι αποτίμησης μοναδιαίου χαρακτηριστικού Gain Ratio και Relief-f παρουσίασαν πολύ χαμηλή απόδοση σε κάθε περίπτωση από την προτεινόμενη μέθοδο αλλά και από τη μέθοδο αποτίμησης υποσυνόλων χαρακτηριστικών FilteredSubsetEval, η οποία έχει απόδοση 98% με 205 επιλεγμένα χαρακτηριστικά. Ο απλός Γενετικός παρουσιάζει την 3^{η} καλύτερη απόδοση μετά τον SVM, αλλά με πολύ λιγότερα χαρακτηριστικά.



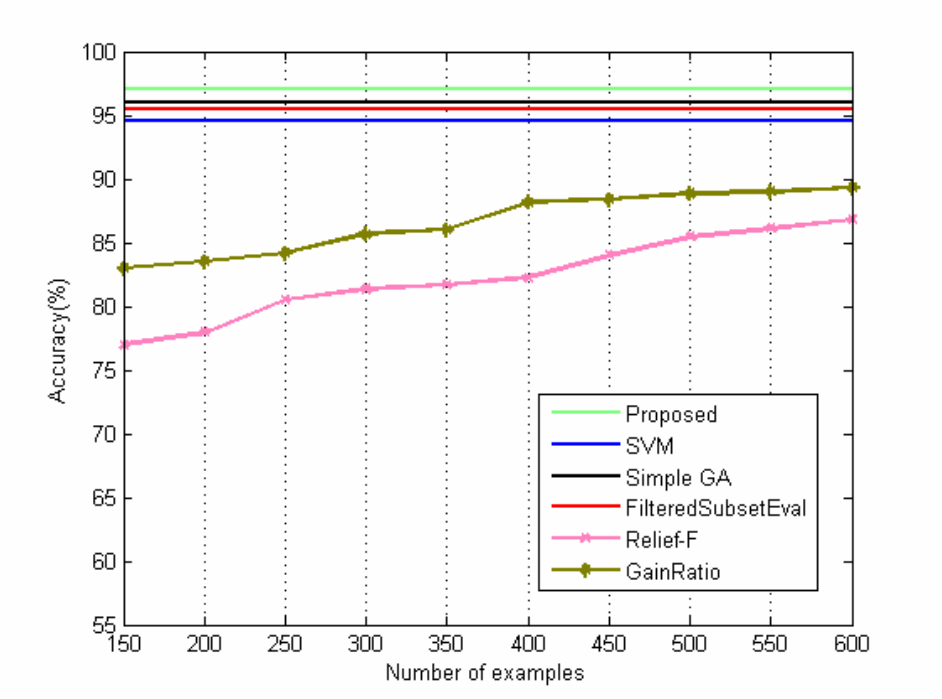
Σχήμα 5.6 Σύγκριση μεθόδων για το σύνολο για το σύνολο δεδομένων ORL.

Στο σύνολο δεδομένων Face Recognition Data University of Essex τα αποτελέσματα για την προτεινόμενη μέθοδο είναι και εδώ καλύτερα όπως φαίνεται στο Σχήμα 5.7. Συγκεκριμένα παρουσιάζει την μεγαλύτερη απόδοση 97,82% με 144 επιλεγμένα χαρακτηριστικά. Η μέθοδος που παρουσιάζει καλά αποτελέσματα είναι η FilteredSubsetEval η οποία έχει απόδοση 95,3% με 235 επιλεγμένα χαρακτηριστικά. Επίσης ο απλός γενετικός αλγόριθμος παρουσιάζει καλύτερη απόδοση από τον ταξινομητή SVM.



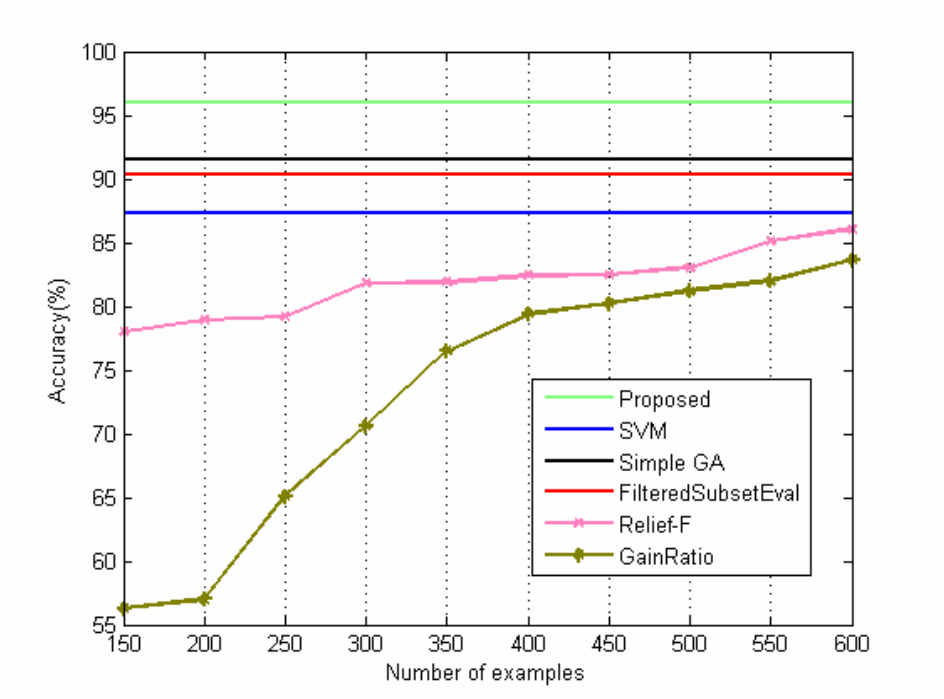
Σχήμα 5.7 Σύγκριση μεθόδων για το σύνολο δεδομένων Essex.

Συνεχίζουμε με το σύνολο δεδομένων FERET παρατηρούμε στο Σχήμα 5.8 ότι η προτεινόμενη μέθοδος παρουσιάζει την καλύτερη απόδοση 97 % με 181 επιλεγμένα χαρακτηριστικά. Ο απλός γενετικός παρουσιάζει την $2^{η}$ καλύτερη απόδοση, ενώ η μέθοδος FilteredSubsetEval παρουσιάζει καλά αποτελέσματα έχοντας απόδοση 95,5% με 205 επιλεγμένα χαρακτηριστικά.



Σχήμα 5.8 Σύγκριση μεθόδων για το σύνολο δεδομένων FERET.

Τέλος για το σύνολο δεδομένων Yale όπως παρατηρούμε στο Σχήμα 5.9 οι. μέθοδοι αποτίμησης μοναδιαίου χαρακτηριστικού GainRadio και Relief-f παρουσίασαν ξανά τη χειρότερη απόδοση, ενώ η προτεινόμενη μέθοδος εξακολουθεί να είναι καλύτερη από όλες τις μεθόδους. Ο απλός γενετικός αλγόριθμος παρουσίασε τη $2^{η}$ καλύτερη απόδοση ενώ η μέθοδος FilteredSubsetEval έχει απόδοση 90,33 % με 143 επιλεγμένα χαρακτηριστικά έναντι της προτεινόμενης μεθόδου η οποία παρουσιάζει απόδοση 95,43 με 171 επιλεγμένα χαρακτηριστικά.



Σχήμα 5.9 Σύγκριση μεθόδων για το σύνολο δεδομένων Yale.

ΚΕΦΑΛΑΙΟ 6. ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην παρούσα εργασία μελετήθηκε το πρόβλημα της επιλογής χαρακτηριστικών για το πρόβλημα της αναγνώρισης προσώπων. Στην εργασία οι εικόνες αναπαρίστανται ως διανύσματα διάστασης ίσης με το μέγεθος της εικόνας (πλήθος των pixels) και επομένως ως διανύσματα «υψηλής διάστασης». Υπάρχουσες μεθοδολογίες επιλογής χαρακτηριστικών όπως προέκυψε από τους πειραματισμούς αδυνατούν να δώσουν ικανοποιητικές λύσεις λόγω της πολυπλοκότητας του προβλήματος.

Για την αντιμετώπιση αυτών των δυσκολιών προτείναμε ένα υβριδικό σχήμα συνδυάζοντας μεθόδους καθολικής βελτιστοποίησης, έναν Γενετικό αλγόριθμο (Genetic algorithm) και τη μέθοδο της Προσομειώμένης Ανόπτησης (Simulated Annealing). Αναλυτικότερα, αρχικά οι εικόνες διαμερίστηκαν σε μικρότερες περιοχές από pixels (blocks) και έτσι ο πολυδιάστατος χώρος του προβλήματος διασπάστηκε σε υποχώρους μικρότερης διάστασης. Στη συνέχεια, σε πρώτο στάδιο οι συνδυασμοί των blocks κωδικοποιούνται ως χρωμοσώματα και ο Γενετικός αλγόριθμος επιλέγει το κατάλληλο υποσύνολο (συνδυασμό) των ενεργών blocks, ενώ στο δεύτερο στάδιο της τοπικής αναζήτησης. τα υποσύνολα χαρακτηριστικών των blocks κωδικοποιούνται ως συμβολοσειρές (strings) πραγματικών αριθμών και η μέθοδος Simulated Annealing βελτιστοποιεί κάθε block επιλέγοντας τα κατάλληλα χαρακτηριστικά.

Η αξιολόγηση του προτεινόμενου υβριδικού σχήματος διεξήχθη σε τέσσερις βάσεις εικόνων προσώπων συγκριτικά και με άλλες μεθόδους επιλογής χαρακτηριστικών. Τα αποτελέσματα είναι ιδιαίτερα ικανοποιητικά για τον υβριδικό αλγόριθμο αφού σε όλες τις περιπτώσεις παρουσιάζει την καλύτερη απόδοση με λιγότερα χαρακτηριστικά.

Η συνεχής κωδικοποίηση αντί της δυαδικής, στη φάση της τοπικής αναζήτησης, έδωσε στοιχεία ευελιξίας στην αναζήτηση των καταλληλότερων χαρακτηριστικών αφού κάθε χαρακτηριστικό επιλέγεται αν θα συμμετάσχει ή όχι στη διαδικασία αξιολόγησης βάση μιας πιθανότητας. Με αυτόν τον τρόπο η επιλογή ενός χαρακτηριστικού παραμένει σε μεγαλύτερο βαθμό ανεπηρέαστη από την αρχική αρχικοποίηση δίνοντας την ευκαιρία σε περισσότερα χαρακτηριστικά να αποτελέσουν μέρος μιας υποψήφιας λύσης και να αξιολογηθούν ως προς τη χρησιμότητά τους.

Περισσότεροι πειραματισμοί και με άλλα σύνολα δεδομένων όπως για παράδειγμα εικόνες αντικειμένων θα ενισχύσουν την αξιοπιστία της προτεινόμενης μεθόδου και θα επιβεβαιώσουν περισσότερο την αποτελεσματικότητά της στην επιλογή χαρακτηριστικών σε δεδομένα υψηλής διάστασης. Τέλος οι Γενετικοί αλγόριθμοι έχοντας έντονα ενδογενή στοιχεία παραλληλισμού δίνουν τη δυνατότητα υλοποίησης του υβριδικού σχήματος σε παράλληλες μηχανές ώστε η αναζήτηση των κατάλληλων υποσυνόλων των χαρακτηριστικών να γίνει πιο γρήγορη.

ΑΝΑΦΟΡΕΣ

[1] W. Zhao, R. Chellappa, P.J Phillips and A. Rosenfeld. “Face recognition: A literature survey”, ACM Computing Surveys, vol. 35, no. 4, pp. 399-458, Dec. 2003.

[2] Zehang Sun, George Bebis and Ronald Miller. “Object detection using feature subset selection”, Pattern Recognition 37 (2004) 2165 – 2176.

[3] P. Viola and M. Jones. “Rapid object detection using a boosted cascacde of simple features”, Proceedings on Computer Vision and Pattern Recognition, 2001.

[4] A. Gyaourova, G. Bebis, I. Pavlidis. “Infrared and visible image fusion for face recognition”, European Conference on Computer Vision, May 2004.

[5] Z. Sun, X. Yuan, G. Bebis and S. Louis. “Neural-network-based gender classi1cation using genetic eigen-feature extraction”, IEEE International Joint Conference on Neural Networks, May 2002.

[6] Z. Sun, G. Bebis, X. Yuan and S. Louis. “Genetic feature subset selection for gender classification: a comparison study”, IEEE Workshop on Applications of Computer Vision, December 2002.

[7] Z. Sun, G. Bebis and R. Miller. “Evolutionary gabor filter optimization with application to vehicle detection”, IEEE International Conference on Data Mining, November 2003, pp. 307–314.

[8] B. Bhanu and Y. Lin. “Genetic algorithm based feature selection for target detection in sar images”, Image Vision Comput. 21 (7) (2003) 591–608.

[9] N. Haering, N. da Vitoria Lobo. “Features and classi1cation methods to locate deciduous trees in images”, Comput. Vision Image Understanding 75 (1/2) (1999) 133–149.

[10] M. Kirby and L. Sirovich. “Application of the Karhunem-Loeve Procedure for the Characterization of Human Faces”, IEEE Trans. Pattern Analysis and Machine Intelligence 12 (1990) 103-108.

[11] M. Turk and A. Pentland. “Eigenfaces for Recognition, Journal of Cognitive Neurosicence”, Vol. 3, No. 1, 1991, pp. 71-86

[12] M.S. Bartlett, J.R. Movellan and T.J. Sejnowski. “Face Recognition by Independent Component Analysis”, IEEE Trans. on Neural Networks, Vol. 13, No. 6, November 2002, pp. 1450-1464

[13] K. Etemad and R. Chellappa. “Discriminant Analysis for Recognition of Human Face Images”, Journal of the Optical Society of America A, Vol. 14, No. 8, August 1997, pp. 1724-1733

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

[15] I.Kononenko. “Estimating attributes: Analysis and extension of RELIEF”, Proc. Eur. Conf. Machine Learning, 1994, pp. 171-182.

[16] 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.

[17] O. Petrocheilos. “Feature selection for classification”, MSc Thesis, Computer Science Department, University of Ioannina, Greece. June, 2009.

[18] W.Siedlecki and J.Sklansky. “A Note on Genetic Algorithms for Large-Scale Feature Selection”, Pattern Recognition Letters, vol. 10 , pp. 335-347, 1989.

[19] C. Liu, H. Wechsler. Face Recognition Using Evolutionary Pursuit, Proc. of the Fifth European Conference on Computer Vision”, ECCV'98, Vol II, 02-06 June 1998, Freiburg, Germany, pp. 596-612

[20] J.H. Holland. “Adaptation in Natural and Artificial Systems”, MIT Press, 1975

[21] Σπυρίδων Λυκοθανάσης. “Γενετικοί αλγόριθμοι και Εφαρμογές”.

[22] D.E. Goldberg. “Genetic Algorithms in search, optimization and machine learning ”,Reading, MA: Addison-Wesley, 1989.

[23] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi. “Optimization by Simulated Annealing” Science, Volume 220, Number 4598, May 1983.

[24] P. J. Phillips. “Support vector machines applied to face recognition”, In Proc.Conf. on Advances in Neural Information Processing Systems II, July 1999, pp. 803-809.

[25] S. Gutta, J. Huang, B. Takacs and H. Wechsler. "Face recognition using ensembles of networks",In Proc. 13th Int. Conf. on Pattern Recognition (ICPR 96), VOL. 4, Aug. 1996, pp. 50-54.

[26] C. Burges, "Tutorial on support vector machines for pattern recognition", Data Mining Knowledge Discovery 2 (2) (1998) 955–974.

[27] M. A. Hall (1998). "Correlation-based Feature Subset Selection for Machine Learning". Hamilton, New Zealand.

[28] AT&T "The Database of Faces", 2010. [Online]. Available at http://www.cl.cam.ac.uk/research/dtg/attarchive/facedatabase.html

[29] The Yale Face Database, 2010. [Online]. Available at http://cvc.yale.edu/projects/yalefaces/yalefaces.html.

[30] The Color FERET Database, USA, 2010. [Online]. Available at http://face.nist.gov/colorFERET/.

[31] Face Recognition Data, University of Essex, UK, 2010. [Online]. Available at http://cswww.essex.ac.uk/mv/allfaces/index.html.

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ο Σάββας Δημητριάδης γεννήθηκε στα Ιωάννινα το 1971. Το 2002 εισήχθη στο Ελληνικό Ανοικτό Πανεπιστήμιο από το οποίο αποφοίτησε το 2007. Παρακολούθησε το πρόγραμμα μεταπτυχιακών σπουδών του τμήματος Πληροφορικής του Πανεπιστημίου Ιωαννίνων από το 2008 και αποφοίτησε τον Νοέμβριο του 2010 αποκτώντας δίπλωμα με ειδίκευση στις «Τεχνολογίες-Εφαρμογές». Τα ερευνητικά του ενδιαφέροντα εστιάζονται στον τομέα της Μηχανικής Μάθησης και πιο συγκεκριμένα στα προβλήματα της Αναγνώρισης Προτύπων και της Επιλογής Χαρακτηριστικών.