Πανεπιστήμιο Πειραιώς – Τμήμα Πληροφορικής
Πρόγραμμα Μεταπτυχιακών Σπουδών
«Προηγμένα Συστήματα Πληροφορικής»

Μεταπτυχιακή Διατριβή

Τίτλος Διατριβής	**Ανάκτηση εικόνων με βάση το περιεχόμενο με χρήση τεχνικών Ημι-Επιτηρούμενης Μηχανικής Μάθησης.**
Ονοματεπώνυμο Φοιτητή	**Βασίλειος Κατωπόδης**
Πατρώνυμο	**Σωκράτης**
Αριθμός Μητρώου	**ΜΠΣΠ/ 09033**
Επιβλέπων	**Γεώργιος Α. Τσιχριντζής, Καθηγητής**

Ημερομηνία Παράδοσης **Απρίλιος 2013**

Τριμελής Εξεταστική Επιτροπή

(υπογραφή)	(υπογραφή)	(υπογραφή)
Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα	Όνομα Επώνυμο Βαθμίδα

Περίληψη

Σκοπός της συγκεκριμένης εργασίας είναι η ανάπτυξη ενός συστήματος, το οποίο με αυτοματοποιημένες διαδικασίες, θα μπορεί να ανακτά εικόνες με βάση τη θεματική ενότητα στην οποία ανήκουν. Το εργαλείο ουσιαστικά εξάγει γνώση από το σύνολο των χαρακτηριστικών των εικόνων και των αξιολογήσεων των χρηστών, την οποία χρησιμοποιεί για την απόδοση τιμών στις εικόνες, με αυτόματο και αποδοτικό τρόπο.

Το πρόγραμμα, για την εκπαίδευση του αλγόριθμου μάθησης, βασίζεται στις αρχές της μηχανικής μάθησης και πιο συγκεκριμένα στις αρχές της ημι-επιτηρούμενης μάθησης. Κατά την εκπαίδευση του αλγόριθμου μάθησης χρησιμοποιεί το σύνολο των εικόνων, βαθμολογημένων και μη.

Από αυτό το σύνολο το εργαλείο εξάγει τα χαμηλού επιπέδου χαρακτηριστικά των εικόνων, τα οποία αποτελούν μέρος του συνόλου εκπαίδευσης. Πιο συγκεκριμένα εξάγει τους καθολικούς οπτικούς περιγραφείς MPEG-7 για κάθε εικόνα. Τα χαρακτηριστικά που εξάγονται, είναι αυτά που σχετίζονται με το χρώμα και την υφή των εικόνων.

Το εργαλείο χρησιμοποιεί σαν αλγόριθμο μάθησης τον Transductive Support Vector Machine (TSVM). Ο αλγόριθμος μάθησης αποτελεί επέκταση του SVM αλγόριθμου και ακολουθεί τις αρχές της μεταγωγικής μάθησης. Σε αντίθεση με τους αλγόριθμους επαγωγικής μάθησης, ο αλγόριθμος δεν εξάγει μια γενικευμένη συνάρτηση απόφασης, αλλά υπολογίζει τις τιμές για τα παραδείγματα του συνόλου δοκιμής.

Στην ανάπτυξη του εργαλείου χρησιμοποιήθηκε η υλοποίηση του SVM, το SVM Light, το οποίο χρησιμοποιεί μια μορφή τοπικής μάθησης για την λύση των προβλημάτων βελτιστοποίησης.

Abstract

The main objective of this paper is the development of a software system which has the ability to retrieve images based on their content, in an automated way. The software tool extracts knowledge from the images' features and the users' ratings. This knowledge is used to rank the images in an automated and efficient manner.
The learning algorithm of the software tool is based on the principles of the machine learning and more specifically on the principles of the semi-supervised learning. The system makes use of both labeled and unlabeled data for the training purposes.
The software tool extracts the low-level description characterizations of the set of images that consists a part of the training set. Particularly extract MPEG-7 global visual descriptions from associated visual content of images. The extracted description characterizations measure the color and the texture of the given images.
The software tool makes use of Transductive Support Vector Machine (TSVM) as the learning algorithm. The TSVMs extend SVMs and they follow the principles of transduction learning.
While regular SVMs try to induce a general decision function for a learning task, this algorithm takes into account a particular test set. The software tool makes use of SVM light, an implementation of SVM, which proceeds by solving a sequence of optimization problems using a form of local search.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή μου κ. Τσιχριντζή Γεώργιο, καθηγητή του Πανεπιστημίου Πειραιώς, για το ενδιαφέρον θέμα μεταπτυχιακής διατριβής που μου πρόσφερε και την εμπιστοσύνη που μου έδειξε. Ακόμα θα ήθελα να ευχαριστήσω τον κ. Σωτηρόπουλο Διονύσιο, μεταδιδακτορικό ερευνητή του Πανεπιστημίου Πειραιώς, για την καθοδήγησή του, τις πολύτιμές συμβουλές του και το χρόνο που αφιέρωσε κατά την εκπόνηση αυτής της μεταπτυχιακής διατριβής.

Πίνακας Περιεχομένων

Εισαγωγή

Στη σημερινή εποχή, λόγω της ραγδαίας ανάπτυξης της τεχνολογίας, κατακλυζόμαστε καθημερινά, από έναν, ολοένα και αυξανόμενο, όγκο δεδομένων. Δεν είναι τυχαίο ότι η σημερινή εποχή χαρακτηρίζεται και ως εποχή της πληροφορίας. Στην καθημερινότητα του ο άνθρωπος, πλέον, θα πρέπει να είναι σε θέση, να διαχειρίζεται αυτόν τον τεράστιο όγκο δεδομένων, και να τον φιλτράρει ώστε να κρατάει ότι είναι σημαντικό για αυτόν.

Ένα από τα πιο σημαντικά στοιχήματα της επιστήμης, και της επιστήμης της πληροφορικής ειδικότερα, είναι να μπορέσει να αξιοποιήσει αυτή την πληθώρα δεδομένων. Για την πρόοδο της επιστήμης είναι πολύ σημαντικό να μπορέσουμε να επεξεργαστούμε τα δεδομένα που έχουμε στη διάθεση μας, και τα οποία προέρχονται από ένα φάσμα πηγών. Αυτό το φάσμα είναι αρκετά ευρύ και περιλαμβάνει όχι μόνο διάφορους επιστημονικούς κλάδους, όπως βιολογία και φυσική, αλλά και πολλές πτυχές της καθημερινότητας, όπως πληροφορίες που προέρχονται από τις χρηματιστηριακές αγορές.

Η μηχανική μάθηση είναι ο κλάδος της τεχνητής νοημοσύνης, ο οποίος ασχολείται με τη δημιουργία και τη μελέτη των συστημάτων, τα οποία μπορούν να αποκτήσουν γνώση μέσω της επεξεργασίας δεδομένων. Η ανάλυση δεδομένων μπορεί να οδηγήσει στην ανακάλυψη της ύπαρξης πολύπλοκων μοτίβων στη ροή των δεδομένων, γνώση ιδιαίτερα σημαντική για την επεξεργασία τους. Απώτερος σκοπός της μηχανικής μάθησης είναι η πραγματοποίηση ανθρώπινων εργασιών από αυτοματοποιημένα συστήματα, χωρίς την ανάγκη ύπαρξης ή επίβλεψης από τον άνθρωπο. Ένα παράδειγμα είναι ένα σύστημα μηχανικής μάθησης, το οποίο μετά από το στάδιο της εκπαίδευσης, είναι ικανό να κατατάσσει έγγραφα με βάση το θέμα του περιεχόμενου τους. Τα αυτοματοποιημένα συστήματα όχι μόνο μπορούν να εκτελέσουν τις εργασίες χωρίς την ύπαρξη του ανθρώπινου παράγοντα, αλλά έχουν την ικανότητα να τις εκτελούν και με πολύ μεγαλύτερη ταχύτητα, και σε πολλές περιπτώσεις με μεγαλύτερη αποτελεσματικότητα, από ότι ένας άνθρωπος.

Ένα από τα επικρατέστερα είδη μηχανικής μάθησης, είναι η εκπαίδευση των αυτοματοποιημένων συστημάτων μέσω παραδειγμάτων, δηλαδή δεδομένων με τιμές. Υπάρχουν δύο διαφορετικές τάσεις για αυτό το είδος της μηχανικής μάθησης, η επιτηρούμενη και η μη επιτηρούμενη μάθηση. Η κύρια διαφορά τους έγκειται στο γεγονός ότι στην επιτηρούμενη μάθηση έχουμε παραδείγματα, τα οποία όλα έχουν τιμή, ενώ στην μη επιτηρούμενη δεν έχουμε γνώση για την τιμή κανενός δεδομένου. Στην πρώτη περίπτωση αξιολογούμε τη λύση, στην οποία καταλήγει το σύστημα, μέσω της διαφοράς των εκτιμήσεων του συστήματος για τις τιμές των δεδομένων από τις πραγματικές τιμές. Στη δεύτερη περίπτωση δεν υπάρχει αντίστοιχη αξιολόγηση για τα αποτελέσματα του συστήματος, αφού τα παραδείγματα που επεξεργάζεται δεν έχουν τιμές.

Τα τελευταία χρόνια μια νέα τάση έχει επικρατήσει στην έρευνα της επιστήμης της μηχανικής μάθησης, η ημι-επιτηρούμενη μάθηση. Σαν σύλληψη βρίσκεται ανάμεσα στη φιλοσοφία της επιτηρούμενης και της μη επιτηρούμενης μάθησης. Τα συστήματα ημι-επιτηρούμενης μάθησης, κατά τη διάρκεια της εκπαίδευσης, κάνουν χρήση και των δύο ειδών δεδομένων. Σε πολλές περιπτώσεις, έχουμε στη διάθεση μας δεδομένα χωρίς τιμή, τα οποία μπορούμε να χρησιμοποιήσουμε για την επίτευξη μεγαλύτερης ακρίβειας στην εκπαίδευση του συστήματος. Συνήθως το σύνολο εκπαίδευσης αποτελείται από μικρό αριθμό δεδομένων με τιμή, και από μεγάλο αριθμό δεδομένων χωρίς τιμή. Εκτός από το θεωρητικό ενδιαφέρον που έχει η ημι-επιτηρούμενη μάθηση σαν ελπιδοφόρα νέα τάση στην μηχανική μάθηση, έχει και καθαρά πρακτικό ενδιαφέρον. Τα δεδομένα με τιμή είναι συνήθως δύσκολο και σπάνιο να βρεθούν, και η απόκτηση τους είναι στις περισσότερες των περιπτώσεων, ακριβής διαδικασία. Η ημι-επιτηρούμενη μάθηση, αξιοποιώντας δεδομένα χωρίς τιμή, αποτελεί μια εξαιρετική λύση σε πολλές εφαρμογές.

Η μηχανική μάθηση μπορεί να χωριστεί σε δύο κατηγορίες, το μοντέλο **επαγωγικής μάθησης** (**inductive learning**) και το μοντέλο **μεταγωγικής μάθησης** (**transductive learning**). Σκοπός του επαγωγικού μοντέλου είναι ο υπολογισμός, με βάση τα δεδομένα εκπαίδευσης, μιας γενικευμένης συνάρτησης, η οποία να μπορεί να υπολογίζει την τιμή όλων των παραδειγμάτων από το πεδίο των δεδομένων. Αντίθετα η μεταγωγική μάθηση δέχεται σαν είσοδο δυο σύνολα δεδομένων, το σύνολο εκπαίδευσης και το σύνολο δοκιμής, και σκοπός του είναι η εύρεση των τιμών του συνόλου δοκιμής. Η μεταγωγική μάθηση βασίζεται στην αρχή του Vapnik, σύμφωνα με την οποία, κατά την επίλυση ενός προβλήματος, πρέπει να αποφεύγεται η λύση ενός πιο γενικού προβλήματος σαν ενδιάμεσο βήμα της λύσης του. Οι αλγόριθμοι που υλοποιούν αυτήν την κατεύθυνση, δεν προσπαθούν να καταλήξουν σε μια γενική θεωρία για το πεδίο των δεδομένων, αλλά σκοπός τους είναι, χρησιμοποιώντας την γνώση που απέκτησαν από τα δεδομένα εκπαίδευσης, να δώσουν τιμές στα σημεία του συνόλου δοκιμής.

Η παρούσα εργασία θα κάνει χρήση του αλγόριθμου **Transductive Support Vector Machine** (**TSVM**). Το **TSVM** αποτελεί επέκταση του αλγόριθμου **Support Vector Machine** (**SVM**), και ανήκει στην κατηγορία αλγόριθμων ημι-επιτηρούμενης μάθησης και βασίζεται στις αρχές της μεταγωγικής μάθησης. Ο αλγόριθμος SVM αποτελεί αλγόριθμο επιτηρούμενης μάθησης και κάνει χρήση μόνο των παραδειγμάτων, δηλαδή δεδομένων που έχουν τιμή. Το TSVM, σε αντίθεση με το SVM, χρησιμοποιεί τις πληροφορίες που του παρέχουν τα δεδομένα, χωρίς τιμή. Σαν σύνολο εκπαίδευσης δέχεται το σύνολο των δεδομένων, δηλαδή των παραδειγμάτων και των δεδομένων χωρίς τιμή.

Σκοπός της εργασίας είναι η δημιουργία ενός συστήματος αυτοματοποιημένης ανάκτησης εικόνων με βάση το περιεχόμενο τους. Η ανάκτηση των εικόνων πραγματοποιείται με την αξιοποίηση τεχνικών ημι-επιτηρούμενης μηχανικής μάθησης. Το σύστημα αρχικά και αφού του προμηθεύσουμε τις εικόνες προς επεξεργασία, θα κάνει εξαγωγή των χαρακτηριστικών τους, τα οποία θα αποθηκεύει σε μορφή διανύσματος. Στη συνέχεια θα δίνει τη δυνατότητα σε χρήστες να βαθμολογούν τις εικόνες αυτές, με γνώμονα τη θεματική ενότητα που έχουν επιλέξει. Οι εικόνες, στην πρώτη φάση της βαθμολόγησης, εμφανίζονται με τυχαία σειρά στο χρήστη. Τα διανύσματα χαρακτηριστικών και οι αντίστοιχες βαθμολογίες των χρηστών, θα αποτελούν τα παραδείγματα με τιμή του συνόλου εκπαίδευσης. Αυτά τα παραδείγματα σε συνδυασμό με τα διανύσματα χαρακτηριστικών των εικόνων που δεν έχουν βαθμολογηθεί, θα χρησιμοποιούνται για την εκπαίδευση του συστήματος μας.

Στη συνέχεια υπολογίζονται και εξάγονται τα αποτελέσματα του αλγόριθμου μηχανικής μάθησης για το σύνολο των εικόνων, που δεν έχει βαθμολογηθεί από τους χρήστες και ανήκουν στο επιλεγμένο θέμα. Τα αποτελέσματα αυτά αναπαριστώνται με τη μορφή γραφικής παράστασης. Ακόμα το σύστημα θα εμφανίζει στο χρήστη το ποσοστό εικόνων, που έχει ανακτήσει σωστά ο αλγόριθμος από την αναζητούμενη θεματική ενότητα. Από τη δεύτερη φάση βαθμολόγησης των εικόνων και μετά, οι εικόνες θα εμφανίζονται στο χρήστη με βάση τη βαθμολογία που έχουν λάβει από τον αλγόριθμο, από την προηγούμενη βαθμολόγηση. Με αυτόν τον τρόπο θα εμφανίζονται στον χρήστη οι εικόνες αυτές, που είναι πιο πιθανό να ανήκουν στο επιλεγμένο θέμα. Στο τέλος της κάθε βαθμολόγησης, θα υπάρχει η δυνατότητα της επιπλέον εκπαίδευσης του συστήματος, προσθέτοντας τα νέα δεδομένα και στη συνέχεια, της προβολής των νέων αποτελεσμάτων, καθώς και του νέου ποσοστού επιτυχίας ανάκτησης των εικόνων από την αναζητούμενη κλάση.

Μηχανική Μάθηση

Η μηχανική μάθηση είναι ένα πεδίο της τεχνητής νοημοσύνης, η οποία αφορά αλγορίθμους που επιτρέπουν σε αυτοματοποιημένα συστήματα να αποκτούν εμπειρική «γνώση». Οι υπολογιστές μπορούν να αποκτήσουν γνώση μέσα από παραδείγματα, εξάγοντας χαρακτηριστικά. Σημαντικός σκοπός της μηχανικής μάθησης είναι να μπορεί να μαθαίνει να αναγνωρίζει πολύπλοκα μοτίβα και να κάνει έξυπνες επιλογές βασιζόμενες στα δεδομένα. Το πρόβλημα έγκειται στο γεγονός ότι οι πιθανές συμπεριφορές σε όλα τα πιθανά δεδομένα, είναι τόσες πολλές που είναι αδύνατο να καλυφθούν από ένα σύνολο δεδομένων παραδειγμάτων. Για αυτό το λόγο θα πρέπει το αυτοματοποιημένο σύστημα να μπορεί να βγάζει γενικά συμπεράσματα από τα παραδείγματα που του δίνουμε, ώστε να μπορεί να έχει χρήσιμα αποτελέσματα σε νέες περιπτώσεις.

2.1 Ορισμός

Η μηχανική μάθηση είναι ένας από τους πιο σημαντικούς τομείς έρευνας της Τεχνητής Νοημοσύνης. Στόχος της είναι η δημιουργία συστημάτων που να μπορούν συνεχώς να

βελτιώνουν την απόδοση τους σε ένα συγκεκριμένο έργο που επιτελούν, χρησιμοποιώντας την εμπειρία που αποκομίζουν κατά την εκτέλεση της εργασίας.

Με τον όρο **εξόρυξη γνώσης** (**data mining**) αναφερόμαστε στην χρήση μεθόδων μηχανικής μάθησης σε μεγάλο όγκο δεδομένων. Αρχικά το πρόβλημα που αντιμετώπιζαν οι ερευνητές ήταν η έλλειψη δεδομένων εκπαίδευσης. Πλέον αυτό το πρόβλημα δεν υφίσταται και το κυρίως πρόβλημα έχει μετατεθεί στη διαχείριση του μεγάλου όγκου δεδομένων εκπαίδευσης από τους αλγόριθμους της μηχανικής μάθησης.

Η εξόρυξη γνώσης πρόβλεψης αναφέρεται σε δύο μεθόδους, την **ταξινόμηση** (**classification**) και την **παλινδρόμηση** (**regression**). Και στις δύο τεχνικές ο στόχος είναι η πρόβλεψη μιας τιμής (label) από τις γνωστές τιμές άλλων μεταβλητών.

Η μηχανική μάθηση προσφέρει πολλά πλεονεκτήματα, που την καθιστούν αρκετά πιο δημοφιλή. Μερικά από αυτά τα πλεονεκτήματα είναι ο ολοένα και περισσότερο αυξανόμενος όγκος δεδομένων προς επεξεργασία, οι περιορισμοί που υπάρχουν σε πολλά πεδία όσον αφορά τις δυνατότητες της ανθρώπινης ανάλυσης και φυσικά το σαφώς χαμηλότερο κόστος που προσφέρει η εκμάθηση ενός αυτοματοποιημένου συστήματος σε σχέση με την εκπαίδευση ενός συνόλου ειδικών.

Η μηχανική μάθηση χρησιμοποιείται σε πολλά και διαφορετικά πεδία. Μερικά από αυτά είναι η ταξινόμηση, η ταυτοποίηση και η αναγνώριση προτύπου.

2.2 Μη Επιτηρούμενη και Επιτηρούμενη Μάθηση

Υπάρχουν δύο θεμελιωδώς διαφορετικοί τύποι μηχανικής μάθησης: η **Μη Επιτηρούμενη Μάθηση** (**Unsupervised Learning**) και η **Επιτηρούμενη Μάθηση** (**Supervised Learning**).

2.2.1 Μη Επιτηρούμενη Μάθηση

Η μη επιτηρούμενη μάθηση ή μάθηση χωρίς επίβλεψη, αναφέρεται στο πρόβλημα της εύρεσης της δομής σε ένα σύνολο δεδομένων, χωρίς να γνωρίζουμε τις επιθυμητές εξόδους. Σε πιο τυπική μορφή θεωρούμε ότι στην μη επιτηρούμενη μάθηση έχουμε ένα σύνολο $X = (x_1, x_2, .., x_n)$ από n, παραδείγματα ή σημεία, όπου $x_i \in X$ για όλα τα $i \in [n] = \{1, ..., n\}$. Πιο βολική θεώρηση είναι να θεωρήσουμε την πίνακα με διαστάσεις (n x d):

$$X = (x_i^T)_{i \in [n]}^T,$$

όπου περιέχει όλα τα δεδομένα σαν σειρές.

Τυπικά σε αυτές τις περιπτώσεις υποθέτουμε ότι αυτά τα σημεία είναι ανεξάρτητα κατανεμημένα από μια κοινή διανομή στον χώρο X. Σκοπός μας είναι η εύρεση μιας δομής στα δεδομένα X, η οποία θα μπορεί να μας είναι χρήσιμη. Συχνά θεωρείται ότι το θεμελιώδες πρόβλημα της μη επιτηρούμενης μάθησης είναι ο υπολογισμός της πυκνότητας που παρήγαγε το X. Εφόσον για το σύνολο δεδομένων που δίνονται στον υπολογιστή σαν παράδειγμα δεν έχουμε τις σωστές εξόδους, δεν υπάρχει κριτήριο λάθους για την πιθανή λύση.

2.2.2 Επιτηρούμενη Μάθηση

Η επιτηρούμενη μάθηση ή μάθηση με επίβλεψη, αναφέρεται στο πρόβλημα εύρεσης μιας συνάρτησης από ένα σύνολο δεδομένων εκπαίδευσης, για τα οποία γνωρίζουμε τις εξόδους τους. Τα δεδομένα εκπαίδευσης είναι ένα σύνολο παραδειγμάτων. Στην επιτηρούμενη μάθηση κάθε παράδειγμα είναι ένα ζεύγος το οποίο αποτελείται από το αντικείμενο εισόδου, τυπικά ένα

διάνυσμα, και από την επιθυμητή τιμή εξόδου. Έστω τα ζεύγη εισόδου (x_i, y_i), τότε τα $y_i \in Y$ καλούνται οι ετικέτες ή τιμές των παραδειγμάτων xi. Αν οι τιμές είναι αριθμοί τότε το $y = (y_i)^T{}_{i\in[N]}$, είναι το διάνυσμα των τιμών. Όπως και στην μη επιτηρούμενη μάθηση, έχουμε ως δεδομένο ότι τα ζεύγη εισόδου ανεξάρτητα κατανεμημένα από μια διανομή στον χώρο X x Y. Η εργασία της επιτηρούμενης μάθησης μπορεί να αξιολογηθεί άμεσα, αφού η χαρτογράφηση μπορεί να εκτιμηθεί μέσω της απόδοσης στις προβλέψεις της στα παραδείγματα δοκιμής. Όταν $y = \Re$ ή $y = \Re^d$, δηλαδή όταν τα οι τιμές ανήκουν στον χώρο των πραγματικών αριθμών και είναι συνεχόμενα, τότε η εργασία της επιτηρούμενης μάθησης λέγεται regression. Στην επιτηρούμενη μάθηση μπορούμε να έχουμε άμεση αξιολόγηση για τα σημεία εξόδου, αφού τα γνωρίζουμε από τα παραδείγματα.

Υπάρχουν δύο οικογένειες αλγορίθμων για την επιτηρούμενη μάθηση: οι Γεννητικοί αλγόριθμοι και οι διακρίνοντες αλγόριθμοι. Οι γεννητικοί αλγόριθμοι προσπαθούν να μοντελοποιήσουν την πυκνότητα των κλάσεων $p(x \mid y)$ με βάση κάποια διαδικασία μη επιτηρούμενης μάθησης. Εφαρμόζοντας το θεώρημα του Bayes μπορούμε να εξάγουμε την πυκνότητα πρόβλεψης, η οποία είναι:

$$p(y \mid x) = \frac{p(x \mid y) * p(y)}{\int_y p(x \mid y) * p(y) dy}$$

Το $p(x \mid y) * p(y) = p(x, y)$ αποτελεί την από κοινού πυκνότητα τω δεδομένων, από την οποία υπολογίζονται τα ζευγάρια (x_i, y_i).

Αντίθετα οι μέθοδοι διάκρισης δεν προσπαθούν να υπολογίσουν πως η τιμή x_i έχει παραχθεί, αλλά επικεντρώνονται στον υπολογισμό της τιμής $p(y \mid x)$. Ο αλγόριθμος support vector machine (SVM), ο οποίος αποτελεί τη βάση του transductive support vector machine (TSVM), αλγόριθμο ημι-επιτηρούμενης μάθησης και με τον οποίο θα ασχοληθούμε στην εργασία, αποτελεί μέθοδο διάκρισης. Οι αλγόριθμοι αυτοί μάλιστα, δεν προσπαθούν να υπολογίσουν την τιμή του $p(y \mid x)$, αλλά προσπαθούν να υπολογίσουν αν οι τιμές που παίρνει είναι μεγαλύτερες ή μικρότερες από 0,5. Το γεγονός εάν τα μοντέλα διάκρισης είναι πιο κοντά στη φιλοσοφία της επιτηρούμενης μάθησης και για αυτό πιο αποτελεσματικά στην πράξη, είναι ακόμα υπό αμφισβήτηση.

2.3 Ημι-Επιτηρούμενη Μάθηση

Η **Ημι–Επιτηρούμενη Μάθηση** (**Semi-supervised learning – SSL**), βρίσκεται σαν φιλοσοφία ανάμεσα στην επιτηρούμενη και την μη επιτηρούμενη μάθηση. Η ημι–επιτηρούμενη μάθηση γίνεται χρήση για εκπαίδευση και των δύο τύπων δεδομένων, και αυτών που έχουν τιμή (**labeled data**) και αυτών που δεν έχουν (**unlabeled data**). Συνήθως στην ημι-επιτηρούμενη μάθηση ο όγκος των δεδομένων με τιμή είναι σημαντικά μικρότερος από τον αντίστοιχο των δεδομένων που δεν έχουν.

Έτσι συνδυάζονται οι διαφορετικές φιλοσοφίες της μη επιτηρούμενης και της επιτηρούμενης μάθησης, όπου η πρώτη έχει δεν περιέχει δεδομένα με τιμή και η δεύτερη δεν περιέχει δεδομένα χωρίς τιμή. Σύμφωνα με τις έρευνες ότι ο συνδυασμός των δύο τύπων δεδομένων συνήθως έχει καλύτερα αποτελέσματα στη μηχανική μάθηση. Η επιλογή των δεδομένων που έχουν τιμή για εκπαίδευση του συστήματος έχει μεγάλη σημασία για το αποτέλεσμα της μάθησης.

Η εύρεση δεδομένων με τιμή είναι δύσκολη και χρονοβόρα διαδικασία, και επίσης έχει υψηλό κόστος, αφού την εύρεση τους αναλαμβάνουν συνήθως άτομα με πείρα. Αντίθετα τα απλά δεδομένα είναι πιο εύκολο να βρεθούν. Το αρνητικό με τα δεδομένα χωρίς τιμή είναι ότι έχουν περιορισμένους τρόπους χρήσης. Για αυτό το λόγο η ημι-επιτηρούμενη μάθηση χρησιμοποιεί πάντα μεγάλο αριθμό τέτοιων δεδομένων, με σκοπό να φτιάξει καλύτερους ταξινομητές. Εξαιτίας του γεγονότος ότι η ημι-επιτηρούμενη μάθηση απαιτεί λιγότερη ανάμιξη του ανθρώπινου παράγοντα και έχει μεγαλύτερη ακρίβεια στα αποτελέσματα, έχει προσελκύσει μεγάλο ενδιαφέρον τόσο στην πράξη όσο και στην θεωρία. Επίσης έχει μεγάλο ενδιαφέρον τόσο ως μοντέλο μηχανική μάθησης όσο και ως μοντέλο για ανθρώπινη μάθηση.

2.3.1 Ιστορία της ημι-επιτηρούμενης μάθησης

Η ιδέα της χρησιμοποίησης δεδομένων που δεν έχουν τιμή για τον σκοπό της ταξινόμησης, χρησιμοποιήθηκε πρώτη φορά στον αλγόριθμο **Αυτό-Εκπαίδευσης** (**Self-Training**). Ο αλγόριθμος αυτός χρησιμοποιεί επαναληπτικά μια επιτηρούμενη μέθοδο. Στην αρχή η εκπαίδευση γίνεται μόνο στα δεδομένα με τιμή. Σε κάθε επανάληψη χαρακτηρίζει ένα μέρος των δεδομένων, που είναι ακόμα χωρίς τιμή, σύμφωνα με την παρούσα συνάρτηση απόφασης. Τότε η επιτηρούμενη μέθοδος επανεκπαιδεύεται χρησιμοποιώντας και τα νέα δεδομένα με τιμή, σαν δεδομένα εκπαίδευσης.

Ο αλγόριθμος αυτό-εκπαίδευσης έχει σημαντικό μειονέκτημα ότι η απόδοση του αλγόριθμου εξαρτάται σε μεγάλο βαθμό από την επιτηρούμενη μέθοδο που χρησιμοποιεί. Με λάθος επιλογή επιτηρούμενης μεθόδου μπορεί να μην έχουμε κανένα κέρδος από την χρησιμοποίηση δεδομένων χωρίς τιμή. Υπάρχουν και περιπτώσεις που δεν μπορούμε να γνωρίζουμε ακριβώς την αποτελεσματικότητα του αλγόριθμου αυτό-εκπαίδευσης.

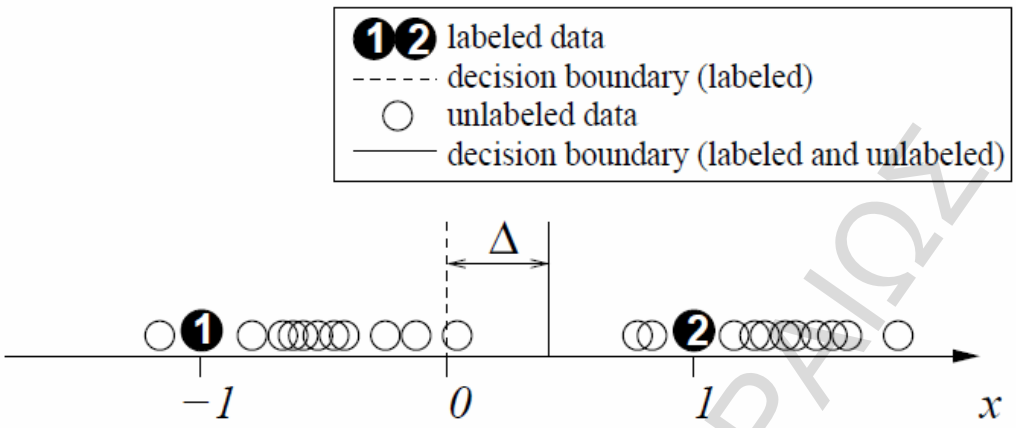
Ακόμα πιο κοντά στην ιδέα της ημι-επιτηρούμενης μάθησης είναι η μεταγωγική μέθοδος από τον Vapnik. Σε αντίθεση με την επαγωγική, δεν συνάγει μια συνάρτηση απόφασης, αλλά υπολογίζει τις ετικέτες μόνο των δεδομένων δοκιμής.

Σημαντικές εξελίξεις έγιναν στη δεκαετία του 1970, όπου ανέκυψε το πρόβλημα εύρεσης του κανόνα γραμμικής διάκρισης Fisher (Fisher linear discriminant) σε δεδομένα χωρίς τιμή. Αφορούσε στην περίπτωση που η πυκνότητα κάθε κλάσης είναι Gaussian. Η πιθανότητα του μοντέλου αυξάνεται με τη χρήση δεδομένων με και χωρίς τιμή, και τη χρήση ενός επαναληπτικού αλγόριθμου expectation-maximization (EM). Μετέπειτα η ύπαρξη μια συνιστώσας για κάθε κλάση επεκτάθηκε σε πολλαπλές συνιστώσες για κάθε κλάση.

Το ενδιαφέρον για την ημι-επιτηρούμενη μάθηση αυξήθηκε στη δεκαετία του 1990, κυρίως λόγω των εφαρμογών σε προβλήματα φυσιολογικής γλώσσας και ταξινόμησης κειμένου. Ο πρώτος που χρησιμοποίησε επίσημα τον όρο “ημι-επιτηρούμενης μάθησης“ για ταξινόμηση με χρήση δεδομένων με και χωρίς ετικέτα, ήταν ο Merz το 1992. Ο όρος έχει χρησιμοποιηθεί ξανά για διαφορετικό σκοπό όμως.

2.3.2 Δεδομένα χωρίς τιμή.

Η ημι-επιτηρούμενη μάθηση βασίζεται στην αξιοποίηση των δεδομένων χωρίς τιμή για την επίτευξη μεγαλύτερης ακρίβειας στις προβλέψεις, σε σχέση με την επιτηρούμενη μάθηση. Αυτό επιτυγχάνεται με την προϋπόθεση πάντα ότι ο διαμοιρασμός των παραδειγμάτων, τα οποία τα δεδομένα χωρίς τιμή θα βοηθήσουν να αποσαφηνίσουμε, να είναι σχετικά με το πρόβλημα ταξινόμησης.



Εικόνα 2.1

Μετατόπιση του ορίου απόφασης με την χρησιμοποίηση των δεδομένων χωρίς τιμή.

Σε μαθηματική βάση, μπορούμε να πούμε ότι η γνώση του p(x), που κερδίζουμε μέσω των δεδομένων χωρίς τιμή, θα πρέπει να φέρει πληροφορία, η οποία θα είναι χρήσιμη στο αποτέλεσμα του p(x|y). Αν δεν συμβαίνει αυτό, τότε δεν θα υπάρχει διαφορά στην ακρίβεια σε σχέση με την επιτηρούμενη μάθηση, η οποία δεν χρησιμοποιεί καθόλου δεδομένα με ετικέτα. Υπάρχει ακόμα το ενδεχόμενο η χρήση δεδομένων χωρίς τιμή να έχει τα αντίθετα από τα επιθυμητά αποτελέσματα, και να μειώσουν την ακρίβεια πρόβλεψης.

Με τη χρήση δεδομένων χωρίς τιμή μπορεί να καταφέρνουμε να μειώνουμε την ανθρώπινη ανάμιξη, αλλά δεν είναι σίγουρο ότι μειώνουμε και την εργασία που απαιτείται από τον ειδικό. Αυτό συμβαίνει γιατί απαιτείται μεγαλύτερη προσπάθεια για τον σχεδιασμό των μοντέλων, των λειτουργιών και γενικά του συστήματος ήμι-επιτηρούμενης μάθησης που θέλουμε να υλοποιήσουμε. Αυτός ο σχεδιασμός αποτελεί σημαντικής σημασίας για το αποτέλεσμα της μηχανικής μάθησης.

Ακόμα τα απλά δεδομένα δεν είναι πάντα σίγουρο ότι θα έχουν τα αποτελέσματα που αναμένουμε. Χρησιμοποιώντας ότι το βέλτιστο μοντέλο υπόθεσης για το πρόβλημα μας μπορεί να έχει σαν αποτέλεσμα την μείωση της αποδοτικότητας του ταξινομητή. Παραδείγματος χάρη υπάρχουν μέθοδοι ημι-επιτηρούμενης μάθησης που θεωρούν ότι το όριο απόφασης πρέπει να αποφεύγουν περιοχές με υψηλό p(x). Έτσι χρησιμοποιώντας τέτοιες μεθόδους σε πρόβλημα με δύο Gaussian κατανομές, που υπερκαλύπτουν η μία την άλλη, το όριο απόφασης θα περνά από την πυκνότερη περιοχή, και το αποτέλεσμα θα είναι να έχουμε πολύ χαμηλή απόδοση. Αντίθετα χρησιμοποιώντας την μια άλλη ημι-επιτηρούμενη μέθοδο, την ΕΜ, θα είχαμε λύσει πολύ πιο εύκολα το πρόβλημα.

Όπως είδαμε και παραπάνω η εύρεση της καταλληλότερης μεθόδου δεν είναι πάντα τόσο εύκολη διαδικασία και αποτελεί ένα από τα σημαντικότερα ζητήματα της ημι-επιτηρούμενης μεθόδου. Εξαιτίας του γεγονότος ότι τα δεδομένα με ετικέτα είναι δυσεύρετα, οι μέθοδοι που χρησιμοποιούμε κάνουν ισχυρές εικασίες μοντέλων. Ιδανικά θα πρέπει να επιλέγουμε την μέθοδο, της οποίας οι εικασίες να ταιριάζουν με τη δομή του προβλήματος. Δυστυχώς τις περισσότερες φορές κάτι τέτοιο θεωρείται δύσκολο στην πράξη. Υπάρχουν όμως κάποιοι γενικοί κανόνες για την επιλογή της κατάλληλης μεθόδου, όπως στην περίπτωση που οι κλάσεις σχηματίζουν συστάδες δεδομένων και τότε θα ήταν σωστό να χρησιμοποιήσουμε την μέθοδο ΕΜ.

2.3.3 Παραδείγματα

Η μηχανική μάθηση βασίζεται στα πρότυπα της μάθησης των ανθρώπων. Φαίνεται ακόμα, ότι ο ανθρώπινος τρόπος σκέψης σε πολλές περιπτώσεις έχει τις ίδιες αρχές με αυτές της ημι-επιτηρούμενης μάθησης.

Παράδειγμα, όπου οι άνθρωποι εφαρμόζουν την ημι-επιτηρούμενη μάθηση, είναι ο τρόπος που τα βρέφη, συσχετίζουν τα αντικείμενα με τις λέξεις. Όταν θέλουν να συσχετίσουν ένα αντικείμενο με μια λέξη, και έχουν ακούσει μια λέξη πολλές φορές, η σύνδεση θα είναι πιο δυνατή. Αυτές οι περιπτώσεις αποτελούν τα unlabeled δεδομένα στην ημι-επιτηρούμενη μάθηση, αφού ακούει την λέξη χωρίς να δει το αντικείμενο. Αν το βρέφος δεν έχει ακούσει τη συγκεκριμένη λέξη προηγουμένως, τότε η σύνδεση είναι πιο αδύνατη.

Ένα άλλο παράδειγμα είναι η δυνατότητα των ανθρώπων να αναγνωρίσουν ένα πρόσωπο, έχοντας το δει αρχικά μόνο την προβολή του από δύο γωνίες.

Ο άνθρωπος έχει τη δυνατότητα να καταλάβει, βλέποντας μια σειρά από υπόλοιπες προβολές του προσώπου, να οι προβολές αυτές ανήκουν στο ίδιο πρόσωπο. Η σειρά αυτών των προβολών αποτελούν τα unlabeled δεδομένα στον τρόπο μάθησης του ανθρώπου.

2.4 Επεξεργασία δεδομένων

Κατά τη διάρκεια της συλλογής των δεδομένων για το σύστημα μας, θα πρέπει να δοθεί ιδιαίτερη σημασία στην ορθότητα των στιγμιότυπων του συνόλου εκπαίδευσης. Αυτό που πρέπει να προσεχθεί είναι η ύπαρξη σφαλμάτων στις τιμές των χαρακτηριστικών στα στιγμιότυπα. Το φαινόμενο αυτό ονομάζεται θόρυβος (noise).

Ο θόρυβος μπορεί να οφείλεται στο γεγονός ότι τα δεδομένα τα συλλέξαμε από πειραματικές μετρήσεις. Είναι συχνό φαινόμενο γενικά όταν επεμβαίνει ανθρώπινος παράγοντας. Όταν έχουμε εκτεταμένο θόρυβο στα δεδομένα μας, μπορεί ο αλγόριθμος μάθησης να μην επιλέξει την βέλτιστη λύση για το σύστημα.

Ένα άλλο φαινόμενο που μπορεί να αποπροσανατολίσει το σύστημα κατά τη διάρκεια της εκπαίδευσης, είναι η απουσία τιμών από τα στιγμιότυπα. Σε αυτήν την περίπτωση λείπουν από κάποια στιγμιότυπα μερικά χαρακτηριστικά. Για να αντιμετωπίσουμε το πρόβλημα της απουσίας τιμών θα πρέπει να εκτελέσουμε μερικές από τις παρακάτω ενέργειες :

- Να διαγράψουμε τα δεδομένα που δεν έχουν όλα τα χαρακτηριστικά.
- Να αντικαταστήσουμε τα χαρακτηριστικά που λείπουν με την μέση τιμή του αντίστοιχου χαρακτηριστικού.
- Να αντικαταστήσουμε τα χαρακτηριστικά που λείπουν με την μέση τιμή του αντίστοιχου χαρακτηριστικού των στιγμιότυπων της ίδιας κλάσης.
- Τέλος μπορούμε να αντικαταστήσουμε τα χαρακτηριστικά που λείπουν με τους διάφορους πιθανούς συνδυασμών. Θα μπορούσαμε να επιλέξουμε τους συνδυασμούς που προέρχονται από τα στιγμιότυπα της ίδιας κλάσης μόνο.

Σε πολλές περιπτώσεις για να έχουμε καλύτερη απόδοση στο σύστημα μάθησης είναι προτιμότερο να κάνουμε διακριτοποίηση των συνεχών μεταβλητών. Η διακριτοποίηση γίνεται με τους παρακάτω τρόπους:

- Με κατάτμηση του διαστήματος τιμών μιας μεταβλητής σε διαστήματα ίσου μεγέθους.
- Με κατάτμηση του διαστήματος τιμών μιας μεταβλητής σε διαστήματα, τα οποία να περιέχουν ίσο αριθμό περιπτώσεων.
- Με την μέθοδο MDL, σύμφωνα με την οποία επιλέγουμε συνέχεια τα σημεία περικοπής του διαστήματος τιμών μιας μεταβλητής, τα οποία ελαχιστοποιούν την εντροπία. Η διαδικασία σταματάει όταν ικανοποιηθεί ένα συγκεκριμένο κριτήριο, το οποίο βασίζεται στο κριτήριο ελαχίστου μήκους περιγραφής.

2.5 Αξιώματα Ημι-Επιτηρούμενης Μάθησης

Η ημι-επιτηρούμενη μάθηση χρησιμοποιεί τα δεδομένα χωρίς τιμή, για να τροποποιήσει την υπόθεση που απέκτησε το σύστημα από τα δεδομένα με τιμή. Στις μεθόδους, που αναπαριστούν τις υποθέσεις με πιθανότητες, τα απλά δεδομένα δίνονται από τον τύπο p(x) και η υπόθεση από τον τύπο p(y|x). Είναι εύκολο να διαπιστώσουμε αν η πιθανότητα p(x) των δεδομένων χωρίς τιμή, επηρεάζει την πιθανότητα p(y|x).

Για να έχει νόημα η χρήση της ημι-επιτηρούμενης μάθησης θα πρέπει να ισχύουν συγκεκριμένα αξιώματα. Ακόμα και η επιτηρούμενη μάθηση βασίζεται σε αξιώματα. Ένα από τα πιο γνωστά αξιώματα είναι το **Αξίωμα Ομαλότητας της Επιτηρούμενης Μάθησης** (**Smoothness Assumption of Supervised Learning**) , του οποίου ο ορισμός είναι ο εξής:

Αν δύο σημεία x1, x2 είναι κοντά, τότε και οι αντίστοιχες τους έξοδοι y1, y2 θα πρέπει επίσης να είναι κοντά.

Είναι εμφανές ότι χωρίς τέτοιες βασικές υποθέσεις θα ήταν αδύνατο να βγάλουμε γενικά συμπεράσματα από ένα πεπερασμένο σύνολο δεδομένων εκπαίδευσης σε ένα μεγάλο αριθμό άγνωστων δεδομένων δοκιμής.

2.5.1 Το Αξίωμα Ομαλότητας της Ημι-Επιτηρούμενης Μάθησης

Το προηγούμενο αξίωμα ομαλότητας για την επιτηρούμενη μάθηση γενικεύεται για την ημι-επιτηρούμενη μάθηση. Εκτός από το αξίωμα ότι οι έξοδοι θα πρέπει να μεταβάλλονται ομαλά, ανάλογα με τη θέση της εισόδου, πρέπει να λαμβάνουμε υπόψη και την πυκνότητα των εισόδων. Η λογική του αξιώματος είναι ότι η συνάρτηση εξόδου είναι πιο ομαλή σε σημεία με υψηλή πυκνότητα, παρά σε σημεία με χαμηλή.

Αξίωμα ομαλότητας της ημι-επιτηρούμενης μάθησης: Αν δύο σημεία x1, x2 σε περιοχή υψηλής πυκνότητας είναι κοντά, τότε επίσης κοντά θα είναι και τα αντίστοιχα σημεία εξόδου y1, y2.

Μεταβατικά αυτό το αξίωμα δηλώνει ακόμα ότι αν δύο σημεία είναι ενωμένα από ένα μονοπάτι υψηλής πυκνότητας, δηλαδή αν ανήκουν στην ίδια συστάδα, τότε τα σημεία εξόδου τους είναι πιθανό να είναι κοντά. Αντιστρόφως, αν δύο σημεία χωρίζονται από μια περιοχή χαμηλής πυκνότητας, τότε τα σημεία εξόδου τους δεν είναι ανάγκη να είναι κοντά.

2.5.2 Το Αξίωμα των Συστάδων

Ένα από τα πρώτα αξιώματα της ημι-επιτηρούμενης μάθησης είναι το αξίωμα των συστάδων. Το αξίωμα αυτό βασίζεται στην παραδοχή ότι τα σημεία της κάθε κλάσης τείνουν να σχηματίζουν μια συστάδα. Τότε τα σημεία χωρίς τιμή μπορούν να βοηθήσουν στην εύρεση των ορίων κάθε συστάδας με μεγαλύτερη ακρίβεια. Μπορούμε έτσι να τρέξουμε έναν αλγόριθμο συσταδοποίησης και να χρησιμοποιήσουμε τα δεδομένα με ετικέτα για να ορίσουμε σε κάθε συστάδα μια κλάση. Ο ορισμός του αξιώματος είναι ο εξής:

Το αξίωμα των συστάδων: Αν τα σημεία ανήκουν στην ίδια συστάδα, είναι πιθανό να ανήκουν και στην ίδια κλάση.

Το αξίωμα αυτό, αν λάβουμε υπόψη τον ορισμό καθαυτό των κλάσεων, θεωρείται λογικό, αφού σε μια περιοχή υπάρχουν συνεχόμενα πολλά αντικείμενα, το πιο πιθανό είναι να μην ανήκουν σε διαφορετικές κλάσεις. Το αξίωμα των συστάδων δεν ισχύει αντίστροφα, δηλαδή δεν δηλώνει ότι κάθε κλάση απεικονίζεται σαν μια συμπαγής συστάδα, αλλά ότι συνήθως δεν παρατηρείται μία συστάδα να αποτελείται από αντικείμενα διαφορετικών κλάσεων.

Το αξίωμα συστάδων μπορεί να θεωρηθεί σαν μια ειδική περίπτωση του αξιώματος ομαλότητας της ημι-επιτηρούμενης μάθησης, δεδομένου του γεγονότος ότι οι συστάδες μπορούν να θεωρηθούν σαν σύνολο σημείων, που μπορούν να ενωθούν από κοντές καμπύλες, οι οποίες βρίσκονται μόνο σε περιοχές υψηλής πυκνότητας. Το αξίωμα μπορεί να οριστεί και με διαφορετικό ισοδύναμο τρόπο:

Διαχωρισμός μικρής πυκνότητας: Το όριο απόφασης πρέπει να βρίσκεται σε περιοχή χαμηλής πυκνότητας.

Αν το όριο απόφασης βρίσκεται σε περιοχή υψηλής πυκνότητας τότε θα χωρίσει μια συστάδα σε δύο διαφορετικές κλάσεις. Ακόμα για πολλά αντικείμενα διαφορετικών κλάσεων στην ίδια συστάδα θα χρειαστεί το όριο απόφασης να κόψει την συστάδα, δηλαδή να περάσει μέσω υψηλής πυκνότητας περιοχή. Παρόλο που οι δύο ορισμοί είναι ισοδύναμοι, παράγουν διαφορετικούς αλγόριθμους.

2.5.3 Το Αξίωμα Manifold

Ο ορισμός του αξιώματος:

Το αξίωμα manifold: Τα δεδομένα πολλαπλών διαστάσεων προβάλλονται σε manifold λίγων διαστάσεων.

Ένα συχνό πρόβλημα των στατιστικών μεθόδων και των αλγόριθμων μάθησης είναι η κατάρα της διστακτικότητας. Βασίζεται στο γεγονός ότι η ένταση αυξάνεται εκθετικά με την αύξηση των διαστάσεων, και ότι αυξάνονται εκθετικά ο αριθμός των παραδειγμάτων που χρειάζεται για στατιστικές εργασίες. Αν όμως τα δεδομένα μπορούν να προβληθούν σε manifold με λίγες

διαστάσεις, τότε ο αλγόριθμος μάθησης μπορεί να λειτουργήσει στο διάστημα της αντίστοιχης διάστασης, αποφεύγοντας έτσι την κατάρα της διστακτικότητας.

Η χρησιμοποίηση manifold μπορεί να θεωρηθεί σαν υλοποίηση του αξιώματος ομαλότητας της ημι-επιτηρούμενης μάθησης. Τέτοιοι αλγόριθμοι χρησιμοποιούν το manifold για τον υπολογισμό γεωδαιτικών αποστάσεων. Αν θεωρήσουμε το manifold σαν προσέγγιση των περιοχών με υψηλή πυκνότητα, τότε το αξίωμα ομαλότητας της ημι-επιτηρούμενης μάθησης ουσιαστικά περιορίζεται στο κανονικό αξίωμα ομαλότητας της επιτηρούμενης μάθησης εφαρμοσμένο πάνω στο manifold.

2.6 Μεταγωγική Μέθοδος

Σύμφωνα με τη φιλοσοφία που πρόβαλε ο Vapnik, τα προβλήματα που έχουν πολλές διαστάσεις θα πρέπει να ακολουθούν την αρχή του Vapnik. Σύμφωνα με αυτήν την αρχή, όταν προσπαθούμε ένα επιλύσουμε ένα πρόβλημα, δεν θα πρέπει να επιλύουμε ένα μεγαλύτερο σαν ενδιάμεσο βήμα.

Ως παράδειγμα μπορούμε να πάρουμε την περίπτωση της επιτηρούμενης μάθησης, όπου σκοπός μας είναι η πρόβλεψη των τιμών y ενός συνόλου αντικειμένων x. Τα γενικευμένα μοντέλα υπολογίζουν την πυκνότητα του x, σαν ένα ενδιάμεσο βήμα, ενώ οι μέθοδοι διάκρισης υπολογίζουν κατευθείαν τις ετικέτες.

Στις περιπτώσεις που απαιτείται να προβλέψουμε τις ετικέτες ενός συγκεκριμένου συνόλου δοκιμής, η **Μεταγωγική Μέθοδος (Transductive Method)** μπορεί να θεωρηθεί πιο άμεση από την **Επαγωγική Μέθοδο (Inductive Method)**. Ενώ η επαγωγική μέθοδος καταλήγει σε μια συνάρτηση για την εύρεση τιμών σε όλο το πεδίο X, και σύμφωνα με την οποία υπολογίζει μετέπειτα τις τιμές των συγκεκριμένων σημείων που μας ενδιαφέρουν, η μεταγωγική μέθοδος υπολογίζει απευθείας τις τιμές μόνο για το σύνολο δοκιμής που μας ενδιαφέρει.

Οι μεταγωγικοί αλγόριθμοι μπορούν να έχουν καλύτερα αποτελέσματα από τους επαγωγικούς αλγόριθμους που έχουν εκπαιδευτεί στο ίδιο σύνολο δεδομένων. Η διαφορά στην απόδοση των δύο αλγορίθμων μπορεί να βασίζεται στο γεγονός ότι η μεταγωγική μέθοδος ακολουθεί την αρχή του Vapnik πιο πιστά από την επαγωγική. Ένας ακόμα λόγος είναι ότι ο μεταγωγικός αλγόριθμος χρησιμοποιεί προς όφελος του τα δεδομένα, για τα οποία δεν έχουμε τιμή, με τρόπο παρόμοιο των αλγόριθμων ημι-επιτηρούμενης μάθησης.

2.7 Ταξινόμηση

Στο εργαλείο που αναπτύξαμε θα έχουμε ένα σύνολο εικόνων στις οποίες θα αποδίδονται διακριτές τιμές. Αυτές οι τιμές αποτελούν τις βαθμολογίες των εικόνων. Η κάθε εικόνα θα βαθμολογείται με βάση τη σχέση του περιεχόμενου της με ένα συγκεκριμένο θέμα, το οποίο εμείς επιλέγουμε κάθε φορά. Οι βαθμολογίες αυτές αποτελούν τις τιμές των παραδειγμάτων μας. Επειδή οι βαθμολογίες έχουν διακριτές τιμές, οι εικόνες μας θα παίρνουν τιμές από ένα πεπερασμένο σύνολο. Αυτού του είδους η μάθηση, με την οποία θα ασχοληθούμε στο πρόγραμμα μας, λέγεται ταξινόμηση.

Η ταξινόμηση θεωρείται εργασία κυρίως της επιτηρούμενης μάθησης. Για να εκπαιδεύσεις έναν ταξινομητή, χρειάζεται ένα σύνολο εκπαίδευσης που τους αντιστοιχούν τιμές. Τα δεδομένα με τιμές όμως είναι ακριβά και δύσκολο να βρεθούν, γιατί συνήθως χρειάζονται να τα βαθμολογούν άνθρωποι, οι οποίοι πρέπει να έχουν την απαιτούμενη πείρα. Μερικά παραδείγματα συνόλων δεδομένων με τιμή είναι:

- Φωνητική αναγνώριση. Η ακριβής μεταγραφή ομιλίας σε φωνητικό επίπεδο είναι υπερβολικά αργή διαδικασία και απαιτεί εξειδικευμένο προσωπικό. Η διαδικασία μπορεί να έχει 400 φορές μεγαλύτερη διάρκεια από τον χρόνο της ομιλίας καθεαυτό. Ακόμα το πρόβλημα είναι ακόμα μεγαλύτερο όταν έχουμε να κάνουμε με ξένες γλώσσες ή με τοπικούς διαλέκτους, όπου η εύρεση ειδικών είναι πιο δύσκολη διαδικασία.
- Κατηγοριοποίηση κειμένου. Σε αυτήν την κατηγορία εμπίπτουν πολλές περιπτώσεις, όπως η κατηγοριοποίηση των μηνυμάτων του χρήστη, προτεινόμενα άρθρα από το ιντερνέτ και 'ξεκαθάρισμα' των ηλεκτρονικών μηνυμάτων. Είναι μια διαδικασία που ουσιαστικά είναι αδύνατο να γίνει από τον μέσο χρήστη καθημερινά, ο οποίος μπορεί να έχει πλήθος μηνυμάτων και άρθρων.
- Επιτήρηση μέσω βίντεο. Η αναγνώριση ανθρώπων μέσα από κάμερες ασφαλείας και φωτογραφίες απαιτεί μεγάλη χρονική διάρκεια και πλήθος προσωπικού.
- Πρόβλεψη δομής των πρωτεϊνών. Αυτή η διαδικασία μπορεί να διαρκέσει και μήνες και να απαιτεί εργαστηριακή δουλειά υψηλού κόστους από ειδικούς επιστήμονες για την εξακρίβωση της 3D δομής μόνο μίας πρωτεΐνης.
- Ταξινόμηση εικόνων και βίντεο. Με την ταξινόμηση εικόνων είναι το θέμα της εργασίας. Η αναγνώριση των θεμάτων του περιεχομένου των εικόνων και βίντεο είναι μια διαδικασία χρονοβόρα και απαιτεί την ενασχόληση πολλών ατόμων για τον χαρακτηρισμό του multimedia περιεχομένου.

2.7.1 Ορισμός

Στην ταξινόμηση αντιμετωπίζουμε ένα πρόβλημα με N κλάσεις: C1, C2, …, Cn. Επίσης κάθε στιγμιότυπο του προβλήματος έχει m χαρακτηριστικά. Ακόμα έχουμε ένα σύνολο εκπαίδευσης, το οποίο είναι ένα σύνολο στιγμιότυπων του προβλήματος και για τα οποία γνωρίζουμε εξαρχής σε ποια κλάση ανήκουν.

Το ζητούμενο της ταξινόμησης είναι η δημιουργία ενός μοντέλου για την ταξινόμηση των νέων άγνωστων στιγμιότυπων. Όταν αναφερόμαστε στην ταξινόμηση εννοούμε την αντιστοίχηση ενός στιγμιότυπου σε μια από τις προκαθορισμένες κλάσεις.

Για να έχουμε επιτυχημένη ταξινόμηση θα πρέπει:

- Να υπάρχει σαφής καθορισμός των κλάσεων του προβλήματος, οι οποίες πρέπει να είναι καθορισμένες και να μην μεταβάλλονται κατά τη διάρκεια της ταξινόμησης.
- Τα στιγμιότυπα που θα ταξινομήσουμε πρέπει να είναι αντιπροσωπευτικά του προβλήματος.

Μια υπόθεση h θεωρείται πως υπερταιριάζει με το στιγμιότυπα του συνόλου εκπαίδευσης, αν υπάρχει μια άλλη υπόθεση h', η οποία να έχει μεγαλύτερο σφάλμα από την h για το σύνολο εκπαίδευσης, αλλά ταυτόχρονα να έχει μικρότερο σφάλμα για το σύνολο των στιγμιότυπων. Δηλαδή η h' να αποτελεί καλύτερη προσέγγιση του πραγματικού μοντέλου από την h.

Ένα σύστημα έχει κλάση που να υπερταιριάζει με τα στιγμιότυπα του συνόλου εκπαίδευσης, όταν το μοντέλο έχει μεγάλο αριθμό παραμέτρων. Το γεγονός αυτό καθιστά πιο δύσκολη την

εργασία του συστήματος στην κατασκευή πολύπλοκων μοντέλων. Ακόμα ένας λόγος είναι να μην έχει γίνει κατάλληλη επιλογή των χαρακτηριστικών των στιγμιότυπων. Τέλος, σημαντικό ρόλο στην εμφάνιση του φαινομένου αυτού, παίζει και ο θόρυβος που μπορεί να περιέχεται στα δεδομένα.

Μεταγωγική Μάθηση

3.1 Εισαγωγή

Με τον όρο μεταγωγική μέθοδο χαρακτηρίζουμε, στην λογική, το συλλογισμό όπου από συγκεκριμένες υποθέσεις καταλήγουμε σε συγκεκριμένες υποθέσεις. Στην μηχανική μάθηση αυτό μεταφράζεται σαν συλλογισμός από τις υποθέσεις εκπαίδευσης σε αυτές της δοκιμής. Αντίθετα η επαγωγική μέθοδος είναι ο συλλογισμός από τις ειδικές υποθέσεις καταλήγουμε στις γενικές.

Υπάρχουν πολλές περιπτώσεις όπου εφαρμόζοντας την επαγωγική μέθοδο, δεν έχουμε τα αναμενόμενα αποτελέσματα, που θα είχαμε αν εφαρμόζαμε την μεταγωγική. Πιο συγκεκριμένα, χρησιμοποιώντας την επαγωγική μέθοδο να μην μπορούμε να καταλήξουμε σε γενικές προβλέψεις, δηλαδή να εξάγουμε έναν γενικό κανόνα από τις ειδικές υποθέσεις που έχουμε στη διάθεση μας, και να καταλήγουμε σε προβλέψεις χρησιμοποιώντας την μεταγωγική μέθοδο. Το μειονέκτημά της μεταγωγικής μεθόδου είναι ότι σε πολλές περιπτώσεις παρατηρούμε να έχουμε προβλέψεις ασυνεπείς. Αυτό συμβαίνει εξαιτίας του γεγονότος ότι η μεταγωγική μέθοδος βασίζεται περισσότερο στις αρχικές προβλέψεις.

Στην μηχανική μάθηση, όπως έχουμε αναφέρει, την μεταγωγική μέθοδο την εισήγαγε ο Vladimir Vapnik στη δεκαετία του 1990. Η πεποίθηση του ήταν ότι η μεταγωγική μέθοδος είναι προτιμότερη από την επαγωγική, για το γεγονός ότι η επαγωγική απαιτεί τη λύση ενός πιο γενικού προβλήματος, για να μπορέσει να λύσει στη συνέχεια ένα πρόβλημα συγκεκριμένου τύπου. Πίστευε δηλαδή, ότι στην προσπάθεια λύσης ενός συγκεκριμένου προβλήματος, πρέπει να αποφεύγεται η λύση ενός πιο γενικού προβλήματος, σαν ενδιάμεσο βήμα της λύσης του, και ότι πρέπει κάποιος να επικεντρώνεται στην εύρεση των αποτελεσμάτων που χρειάζεται για το συγκεκριμένο πρόβλημα.

Εκτός της αποφυγής του πιο περίπλοκου ενδιάμεσου βήματος, η μεταγωγική μέθοδος μπορεί να εξάγει τα επιθυμητά αποτελέσματα σε περιπτώσεις όπου δεν μπορεί η μεταγωγική. Ένα τέτοιο παράδειγμα είναι η περίπτωση της δυαδικής ταξινόμησης. Στην περίπτωση αυτή τα παραδείγματα εισόδου τείνουν να συγκεντρώνονται σε δύο συστάδες. Στη δυαδική ταξινόμηση χρησιμοποιώντας ένα μεγάλο σύνολο παραδειγμάτων εισόδου, μπορούμε να βρούμε τις δύο αυτές συστάδες, και έτσι να έχουμε μια πιο σαφή εικόνα για τα δεδομένα με τιμή. Αυτές οι προβλέψεις δεν θα μπορούσαν να εξαχθούν αν χρησιμοποιούσαμε ένα επαγωγικό μοντέλο, το οποίο θα προσπαθούσε να εξάγει μια συνάρτηση, βασισμένο μόνο στις περιπτώσεις των δεδομένων δοκιμής.

Τέλος η μεταγωγική μέθοδος, σε αντίθεση με την επαγωγική, έχει την ικανότητα να υπολογίζει αποτελέσματα με προσέγγιση. Υπάρχουν περιπτώσεις όπου ο ακριβής υπολογισμός των αποτελεσμάτων είναι μη επιθυμητός, όπως στις περιπτώσεις που απαιτείται μεγάλος υπολογιστική δύναμη ή στις περιπτώσεις που απαιτείται μεγάλος χρόνος υπολογισμού. Σε αυτές τις περιπτώσεις μπορούμε με την μεταγωγική μέθοδο να υπολογίσουμε προσεγγιστικά τα αποτελέσματα, στον βαθμό που μας ικανοποιούν για τους υπολογισμούς μας.

3.2 Μεταγωγική Μάθηση

3.2.1 Κεντρική Ιδέα της Μεταγωγικής Μάθησης

Η βασική ιδέα της **Μεταγωγικής Μάθησης** (**Transductive Learning**) είναι ότι σε αντίθεση με την **Επαγωγική Μάθηση** (**Inductive Learning**), προβλέψεις γίνονται μόνο για τον ορισμένο αριθμό των δεδομένων εκπαίδευσης. Ο αλγόριθμος μάθησης με αυτόν τον τρόπο μπορεί να χρησιμοποιήσει τις πληροφορίες που του παρέχουν τα σημεία δοκιμής (data points) για την επίλυση του προβλήματος της ημι-επιτηρούμενης μάθησης. Οι **Transductive Support Vector Machines** (**TSVM**) υλοποιούν την ιδέα του Transductive learning, περιλαμβάνοντας στους υπολογισμούς του ορίου απόφασης το σύνολο των δεδομένων, δηλαδή τα δεδομένα που έχουν τιμή, καθώς και αυτά που δεν έχουν. Αντίθετα το απλό **Transductive Learning Machine** (**SVM**), λαμβάνει υπόψη του κατά τον υπολογισμό του ορίου απόφασης μόνο τα δεδομένα για τα οποία έχουμε τιμή.

Χαρακτηριστικό παράδειγμα όπου μπορεί να βρει εφαρμογή το inductive learning είναι η κατάταξη ενός αριθμού εγγράφων σε κατάλληλα και σε ακατάλληλα, σε σχέση με ένα θέμα που έχουμε επιλέξει. Αρχικά ο χρήστης αξιολογεί έναν αριθμό εγγράφων, και στη συνέχεια αυτό το δείγμα θα χρησιμοποιηθεί σαν σύνολο εκπαίδευσης για το πρόβλημα κατάταξης κειμένου με δυο τιμές. Ο στόχος είναι η εύρεση ενός κανόνα, ο οποίος θα ταξινομεί σωστά τα έγγραφα με βάση τη συνάφεια τους ως προς το θέμα που έχουμε θέσει.

Αντίστοιχα θέματα μπορούν να αντιμετωπιστούν και σαν προβλήματα επιτηρούμενης μάθησης. Όμως υπάρχουν δύο σημαντικές διαφορές που διαφοροποιούν το παράδειγμα μας από τα κλασικά προβλήματα του inductive learning.

Σε αντίθεση με το inductive learning, ο αλγόριθμος εκπαίδευσης δεν χρειάζεται αναγκαία να μάθει έναν γενικό κανόνα, που θα ισχύει για όλα τα σημεία του χώρου. Αντιθέτως το μόνο που χρειάζεται είναι να μπορεί να προβλέψει με ακρίβεια τις τιμές για έναν συγκεκριμένο αριθμό παραδειγμάτων. Ακόμα τα παραδείγματα της δοκιμής είναι γνωστά εκ των προτέρων, γεγονός που επιτρέπει να μπορούν να παρατηρηθούν από τον αλγόριθμο μάθησης κατά τη διάρκεια της εκπαίδευσης. Αυτό επιτρέπει στον αλγόριθμο μάθησης να εκμεταλλευθεί όποια πληροφορία προσφέρει η τοποθεσία των παραδειγμάτων δοκιμής.

Για αυτούς τους λόγους το Transductive learning θεωρείται ειδική περίπτωση ημι-επιτηρούμενης μάθησης, αφού επιτρέπει τον αλγόριθμο εκμάθησης να χρησιμοποιήσει τα παραδείγματα χωρίς τιμή από το σύνολο δοκιμής.

3.2.2 Ορισμός

Θεωρώντας όλα τα έγγραφα που έχουμε στην βάση δεδομένων μας σαν ένα σύνολο παραδειγμάτων, τότε μπορούμε να τα αναπαραστήσουμε ως:

$$S = \{1,2,...,n\}$$

Όπου **n** είναι το σύνολο των παραδειγμάτων. Καθένα από τα n παραδείγματα αποτελεί ένα έγγραφο, και παριστάνεται σαν ένα διάνυσμα. Το κάθε διάνυσμα που ανήκει στο σύνολο των παραδειγμάτων έχει d διαστάσεις, που αναπαριστούν τις ιδιότητες των εγγράφων. Το σύνολο αυτών των διανυσμάτων το ονομάζουμε **X** και έχει την εξής μορφή:

$$X = (x_1, x_2, ..., x_n)$$

Τα labels αντίστοιχα ανήκουν στο σύνολο **Y** και παράγονται ανεξάρτητα σύμφωνα με μια κατανομή **P**, και έχουν την μορφή:

$$Y = (y_1, y_2, ..., y_n)$$

Στο παράδειγμα μας οι τιμές των παραδειγμάτων παίρνουν δύο μόνο τιμές: -1 και +1. Σαν σύνολο εκπαίδευσης μπορούμε να επιλέξουμε ένα υποσύνολο παραδειγμάτων και τα υπόλοιπα παραδείγματα να τα χρησιμοποιήσουμε σαν σύνολο δοκιμής, έτσι ώστε να έχουμε

$$S_{test} = S \setminus S_{train}$$

Σε αντίθεση με τους inductive αλγόριθμους εκμάθησης, οι οποίοι κατά τα διάρκεια εκπαίδευσης έχουν πρόσβαση μόνο στα δεδομένα που έχουν δεδομένα, οι Transductive αλγόριθμοι χρησιμοποιούν επιπλέον και τα δεδομένα χωρίς τιμή. Έτσι ένας inductive αλγόριθμος χρησιμοποιεί τα σύνολα $Xtrain$, $Ytrain$ και $Ytest$ για να παράγει προβλέψεις για τα labels των παραδειγμάτων δοκιμής. Τις προβλέψεις τις συμβολίζουμε με:

$$Y^*_{test} = (y^*_1, y^*_2, ..., y^*_u),$$

Όπου **u** ο αριθμός των παραδειγμάτων του συνόλου δοκιμής.

Σκοπός του αλγόριθμου είναι να ελαττώσει το κλάσμα των λανθασμένων προβλέψεων για τα labels αυτά. Έτσι έχουμε τον τύπο:

$$Err_{test}(Y^*_{test}) = \frac{1}{u} \sum_{iEStet} \delta_0 \, /1(y^*_i, y_i)$$

Όπου το y^*_i αποτελεί την πρόβλεψη για το παράδειγμα x_i , το οποίο ανήκει στο σύνολο δοκιμής. Στην περίπτωση που η πρόβλεψη του αλγόριθμου εκμάθησης είναι σωστή, τότε θα έχουμε $\delta_{0/1}$ (a, b) ίσο με το μηδέν. Αν είναι λάθος η πρόβλεψη τότε θα είναι ίσο με το ένα.

Από τον ορισμό του Transductive learning μπορεί κάποιος να υποθέσει ότι αποτελεί παραλλαγή του inductive learning, αφού μπορούμε να εξάγουμε έναν κανόνα κατάταξης, μια συνάρτηση, από τα παραδείγματα εκπαίδευσης και να τον εφαρμόσουμε στα παραδείγματα δοκιμής για να κάνουμε τους υπολογισμούς. Η εγγενής διαφορά όμως αυτών των δύο αλγορίθμων είναι ότι το Transductive λαμβάνει υπόψη του τις πληροφορίες που του παρέχουν τα δεδομένα δοκιμής, ενώ το inductive learning όχι.

Στο transductive learning έχουμε το πλεονέκτημα ότι ο αριθμός των τιμών των προβλέψεων για τα παραδείγματα μας είναι πεπερασμένος, αφού πεπερασμένος είναι και ο αριθμός των παραδειγμάτων μας. Έτσι το υποθετικό διάστημα Η ενός transductive learner είναι αναγκαστικά πεπερασμένο. Μπορούμε να οικοδομήσουμε το **H** σε μια ένθετη κατασκευή:

$$H1 \subset H2 \subset ... \subset H = \{-1, +1\}^n$$

Η κατασκευή θα πρέπει να έχει τέτοια μορφή ώστε αν στο σύνολο **S**, ο αλγόριθμος προβλέψει τα σωστά labels ή οι προβλέψεις περιέχουν λίγα λάθη, τότε το εύρος των τιμών θα περιέχεται

σε ένα υποθετικό διάστημα μικρής πληθικότητας. Αυτή η δόμηση του υποθετικού διαστήματος Η, μπορεί να επιτευχθεί χρησιμοποιώντας γενικευμένα όρια σφάλματος από τη θεωρία στατιστικής εκμάθησης. Για την περίπτωση που αναζητούμε μια υπόθεση με μικρό σφάλμα εκπαίδευσης είναι πιθανό να βρούμε άνω όριο για το σφάλμα δοκιμής. Το σφάλμα εκπαίδευσης:

$$Err_{test}(Y^*_{train}) = \frac{1}{l}\sum_{iEStrain}\delta_{0/1}(y^*_i, y_i),$$

όπου **l** είναι ο αριθμός των παραδειγμάτων του συνόλου εκπαίδευσης. Με πιθανότητα **1-n,μ** έχουμε το όριο για το σφάλμα δοκιμής:

$$Err(Y^*_{test}) \le Err_{train}(Y^*_{train}) + \Omega(l,u,|Hi|,n),$$

όπου το διάστημα **Ω** εξαρτάται από τον αριθμό των παραδειγμάτων του συνόλου εκπαίδευσης, τον αριθμό των παραδειγμάτων του συνόλου δοκιμής και από την πληθικότητα |Hi|. Όσο μικρότερη είναι η |Hi|, τόσο μικρότερη είναι η παρέκκλιση ανάμεσα στα σφάλματα της εκπαίδευσης και της δοκιμής.

Διασφαλίζοντας ανώτατο όριο μπορούμε να είμαστε σίγουροι ότι θα υπάρχει ακριβής πρόβλεψη για τα labels των παραδειγμάτων δοκιμής. Έτσι σε αντίθεση με το inductive learning, το transductive learning έχει τη δυνατότητα να χρησιμοποιήσει προηγούμενη γνώση που μπορούμε να έχουμε σχετικά με τη σχέση της γεωμετρίας του συνόλου $\mathbf{X = (x_1, x_2, \ldots, x_n)}$ και την αντίστοιχη της κατανομής $\mathbf{P(y_1, y_2, \ldots, y_n)}$. Στις περιπτώσεις όπου ισχύει υπάρχει μια τέτοια γνώση, έχουμε τη δυνατότητα να ελαττώσουμε τον αριθμό των παραδειγμάτων εκπαίδευσης που είναι αναγκαία για να έχουμε την επιθυμητή ακρίβεια προβλέψεων.

3.2.3 Μεταγωγικοί Αλγόριθμοι

Θεωρούμε **Μεταγωγικούς Αλγόριθμους (Transductive Algorithms)** τους αλγόριθμους εκμάθησης αυτούς, οι οποίοι ενεργούν μέσα στα πλαίσια ενός μεταγωγικού μοντέλου. Δηλαδή είναι οι αλγόριθμοι αυτοί που λαμβάνουν ένα σύνολο εκπαίδευσης με τιμές και ένα σύνολο δοκιμών και έχουν σαν σκοπό την εύρεση των τιμών για το σύνολο δοκιμών.

Όπως γίνεται αντιληπτό ο ορισμός αυτός είναι πολύ γενικός, καθώς μπορεί να ισχύει σε αλγόριθμους επιτηρούμενης μάθησης, όσο και σε αλγόριθμους ημι-επιτηρούμενης μάθησης. Μπορούμε να χρησιμοποιήσουμε αλγόριθμο επιτηρούμενης μάθησης στο σύνολο εκπαίδευσης και να εξάγουμε μια υπόθεση, με την οποία θα δώσουμε τιμές στα απλά δεδομένα. Ακόμα μπορούμε να χρησιμοποιήσουμε αλγόριθμο ημι-επιτηρούμενης μάθησης στα δεδομένα εκπαίδευσης και στα δεδομένα δοκιμής, να εξάγουμε πάλι μια συνάρτηση και να δώσουμε τιμές στα δεδομένα δοκιμής. Και οι δύο αλγόριθμοι παράγουν μια γενική υπόθεση, σύμφωνα με την οποία υπολογίζουν τις τιμές των δεδομένων δοκιμής. Το ενδιαφέρον το δικό μας όμως, εστιάζεται στο γεγονός ότι οι αλγόριθμοι μάθησης να είναι γνήσια transductive, δηλαδή να μπορεί παράγει υπόθεση η οποία θα μπορεί να υπολογίσει τα τιμών των δεδομένων δοκιμής και μόνο.

Οι μεταγωγικοί αλγόριθμοι μπορούν να χωριστούν σε δύο γενικές κατηγορίες: στους αλγόριθμους που αναθέτουν διακριτές τιμές σε απλά σημεία, και σε εκείνους που αναθέτουν συνεχείς τιμές. Η πρώτη κατηγορία αλγορίθμων τείνουν να προέρχονται από έναν αλγόριθμο συσταδοποίησης, προσθέτοντας μερική επιτήρηση. Αυτοί οι αλγόριθμοι με τη σειρά τους χωρίζονται σε δύο κατηγορίες: σε αυτούς που συγκεντρώνουν σε συστάδες με τη μέθοδο του διαχωρισμού, και σε αυτούς που συγκεντρώνουν σε συστάδες με τη μέθοδο της συσσώρευσης.

Η δεύτερη κατηγορία αλγορίθμων συνήθως προέρχονται από αλγόριθμους μάθησης manifold, προσθέτοντας μερική επιτήρηση.

3.3 Μεταγωγική και Επαγωγική Μάθηση

3.3.1 Διαφορές Μεταγωγικής και Επαγωγικής Μάθησης

Η μεταγωγική μάθηση έχει ως εκ φύσεως να εκτελέσει πιο εύκολο έργο από την επαγωγική μάθηση. Σκοπός της είναι η εκμετάλλευση της επιπλέον πληροφορίας που περιέχεται στα δεδομένα χωρίς τιμή και να την προσθέσει στην πληροφορία που έχουμε ήδη από τα δεδομένα με τιμή του συνόλου εκπαίδευσης.

Στην επαγωγική μάθηση βρίσκουμε τη συνάρτηση αυτή, η οποία θα μπορεί να κάνει προβλέψεις για όλο το πεδίο τιμών. Αντίθετα την μεταγωγική μάθηση, την απασχολεί μόνο η πρόβλεψη των τιμών της συνάρτησης για τα σημεία δοκιμής που μας ενδιαφέρουν. Το πρόβλημα αυτό αποτελεί ευκολότερο στόχο, αφού η λύση ενός επαγωγικού προβλήματος περιέχει και τη λύση ενός μεταγωγικού, αξιολογώντας τη συνάρτηση για τα δοθέντα δεδομένα δοκιμής, ενώ δεν συμβαίνει το αντίστροφο.

Η μεταγωγική μάθηση λειτουργεί καλύτερα γιατί το σύνολο δοκιμής μπορεί να δώσει παραγοντοποίηση της κλάσης της συνάρτησης. Στην περίπτωση που έχουμε δύο συναρτήσεις ισοδύναμες, συναρτήσεις δηλαδή που δεν μπορούν να διαχωριστούν βασιζόμενες σε παραδείγματα του συνόλου εκπαίδευσης ή δοκιμής, τότε είναι αρκετό να χρησιμοποιήσουμε μόνο μια συνάρτηση αντιπροσωπευτική της κάθε ισοδύναμης κλάσης και καμία άλλη. Η κλάση συναρτήσεων είναι πεπερασμένη και για αυτό μπορούμε να έχουμε ένα γενικευμένο όριο λάθους.

Στην ημι-επιτηρούμενη μάθηση κάθε σημείο χωρίς τιμή μας δίνει πληροφορίες στην κατανομή P(x). Αν το σημείο θα είναι χρήσιμο εξαρτάται σε μεγάλο βαθμό από την κατανομή. Αν παραδείγματος χάρη η κατανομή ικανοποιεί το αξίωμα της ημι-επιτηρούμενης ομαλότητας, τότε ακόμα και ένα μόνο σημείο μπορεί να μας δώσει πληροφορίες. Ένας τρόπος, που ένα σημείο μπορεί να επηρεάσει το αποτέλεσμα της ημι-επιτηρούμενης μάθησης, είναι ότι αλλάζει την πυκνότητα των σημείων, στην περιοχή που βρίσκεται, και για αυτό επηρεάζει την απόφαση μας που θα θέσουμε ομαλότητα στον χώρο. Με άμεση συνέπεια να επηρεάζει τις προβλέψεις μας για τα αποτελέσματα των σημείων δοκιμής.

Στην μεταγωγική μάθηση, όσα περισσότερα σημεία δοκιμής έχουμε στη διάθεση μας, τόσο περισσότερο πλησιάζουμε στην επαγωγική μάθηση, γιατί θα έχουμε να προβλέψουμε τα αποτελέσματα για ένα σύνολο σημείων το οποίο σταδιακά θα καλύπτει όλο τον χώρο.

Η επαγωγική μάθηση είναι χρήσιμη για δύο διαφορετικούς λόγους. Ο πρώτος λόγος είναι ότι τα όρια είναι πιο σφιχτά από τα αντίστοιχα της επαγωγικής μάθησης. Ο δεύτερος είναι μετρώντας το μέγεθος των ισοδύναμων κλάσεων είναι μια ευκαιρία για να αλλάξουμε τη σειρά στην κατασκευή των κλάσεων των συναρτήσεων, ενέργεια που είναι κοντά στους στόχους της ημι-επιτηρούμενης μάθησης.

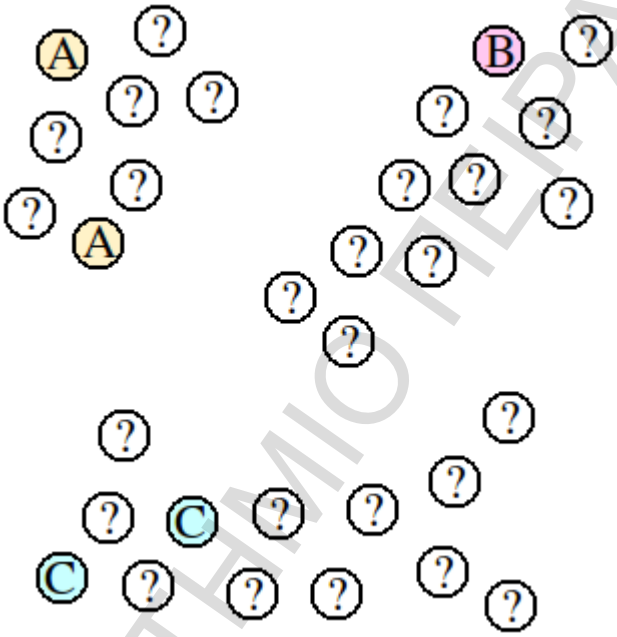
Οι μεταγωγικοί αλγόριθμοι μπορούν ακόμα να επιλεχθούν για υπολογιστικούς λόγους. Για παράδειγμα η Bayesan committee machine, που τα αποτελέσματα της είναι επέκταση ενός συνόλου βασικών συναρτήσεων. Για λόγους υπολογισμού, μόνο συναρτήσεις που επικεντρώνονται σε σημεία του συνόλου δοκιμής λαμβάνονται υπόψη.

Οι graph-based αλγόριθμοι μπορούν να ερμηνευθούν και σαν ημι-επιτηρούμενοι και σαν μεταγωγικοί αλγόριθμοι. Οι graph-based αλγόριθμοι είναι μεταγωγικοί γιατί δεν μπορούμε να κάνουμε πρόβλεψη για ένα σημείο το οποίο δεν ανήκει στο σύνολο δοκιμής. Αν συμπεριλάβουμε ένα τέτοιο σημείο στο γράφημα μπορεί να έχει και αρνητικές συνέπειες, αφού μπορεί να μας παρέχει πληροφορίες παραπλανητικές όσον αφορά την κατανομή P(x). Στην

μεταγωγική μάθηση τα σημεία δοκιμής θα πρέπει να προέρχονται από την κατανομή P(x), ή τουλάχιστόν από μια κατανομή σχετική με τη P(x).

3.3.2 Πλεονεκτήματα Μεταγωγικής Μεθόδου

Τα πλεονεκτήματα της μεταγωγικής μεθόδου σε σχέση με την επαγωγική, μπορούμε να τα παρατηρήσουμε καλύτερα στην περίπτωση, όπου μας δίνεται ένα σύνολο σημείων και πρέπει να προβλέψουμε τις τιμές όλων των σημείων. Τα σημεία που έχουν τιμές ανήκουν σε τρεις κλάσεις.



Εικόνα 3.1
Σημεία που ανήκουν σε κλάσεις

Αν θελήσουμε να λύσουμε το πρόβλημα με την επαγωγική μέθοδο, τότε θα πρέπει να εκπαιδεύσουμε ένα αλγόριθμο επιτηρούμενης μάθησης για να προβλέψουμε τιμές. Συνήθως όμως, σε περιπτώσεις όπως αυτό το πρόβλημα, τα δεδομένα με τιμή είναι πολύ λιγότερα από αυτά που δεν έχουν, και έτσι ο αλγόριθμος επιτηρούμενης μάθησης θα έχει πολύ λίγα σημεία να χρησιμοποιήσει για να παράγει το μοντέλο πρόβλεψης. Ειδικά σε περιπτώσεις που τα δεδομένα με τιμή είναι αρκετά περιορισμένα, είναι σχεδόν αδύνατο ο αλγόριθμος να καταφέρει να καταλήξει σε μοντέλο, το οποίο να αντιλαμβάνεται τη δομή αυτών των δεδομένων. Για παράδειγμα, αν χρησιμοποιηθεί ο αλγόριθμος του κοντινότερου-γείτονα, τότε σημεία που βρίσκονται κοντά σε σημεία άλλης κλάσης, θα ταξινομηθούν σαν τέτοια, ενώ μπορεί να ανήκουν εμφανώς σε μια συστάδα, η οποία απαρτίζεται από σημεία άλλης κλάσης.

Αν εφαρμόσουμε μεταγωγική μέθοδο θα έχουμε το πλεονέκτημα ότι θα λαμβάνουμε υπόψη μας το σύνολο των δεδομένων, δηλαδή και των δύο ειδών σημείων. Σε αυτήν την περίπτωση, μεταγωγικοί αλγόριθμοι θα μπορούν να δίνουν τιμές στα σημεία, σύμφωνα με τη συστάδα στην οποία ανήκουν.

Το βασικό πλεονέκτημα, λοιπόν, της μεταγωγικής μεθόδου, είναι ότι μπορεί να κάνει καλύτερες προβλέψεις, έχοντας στη διάθεση της λιγότερα δεδομένα με τιμή. Ένα μειονέκτημά της είναι όμως ότι εξ ορισμού δεν μπορεί να δημιουργήσει ένα γενικό μοντέλο πρόβλεψης. Αν στο σύνολο δεδομένων προσθέσουμε ένα νέο σημείο, τότε ο μεταγωγικός αλγόριθμος θα πρέπει να εκτελεστεί από την αρχή, για συμπεριλάβει στα δεδομένα εισόδου και το νέο σημείο. Αυτή η διαδικασία είναι φυσικά μη αποδοτική στις περιπτώσεις όπου τα δεδομένα εισόδου έρχονται στη διάθεσή μας σταδιακά ή σειριακά. Ακόμα ο υπολογισμός για νέα σημεία, μπορεί να έχει σαν αποτέλεσμα να αλλάξουν οι υπολογισμοί για τις τιμές των αρχικών σημείων. Αντίθετα με τις επαγωγικές μεθόδους μπορούμε να υπολογίσουμε τις τιμές νέων σημείων άμεσα, χρησιμοποιώντας το ήδη υπολογισμένο μοντέλο προβλέψεων.

3.4 Μεταγωγική και Ημι-Επιτηρούμενη Μάθηση

Στην κοινότητα της μηχανικής μάθησης υπάρχει μια σύγχηση, όσον αφορά τους ορισμούς και τις διαφορές της Μεταγωγικής μάθησης και της Ημι-Επιτηρούμενης μάθησης. Πολλές φορές μάλιστα, οι ορισμοί της μεταγωγικής και ημι-επιτηρούμενης μάθησης ταυτίζονται. Το κοινό τους σημείο είναι το γεγονός ότι και οι δύο χρησιμοποιούν τα παραδείγματα που δεν έχουν τιμή, για την εξαγωγή αποτελεσμάτων. Οι μεταγωγικοί μέθοδοι είναι πάντα και μέθοδοι ημι-επιτηρούμενης μάθησης. Χρησιμοποιούν πληροφορία, η οποία περιέχεται στα δεδομένα δοκιμής.

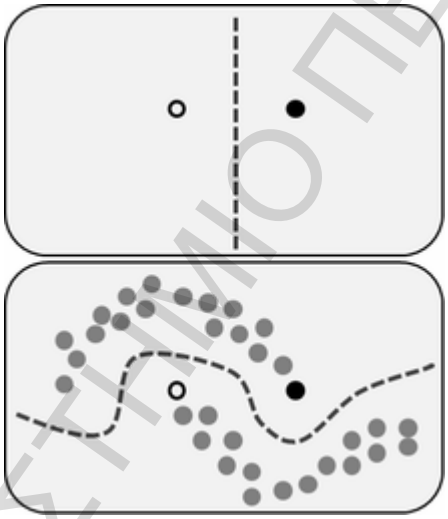
Τα αρχικά μοντέλα της Επιτηρούμενης και Ημι-Επιτηρούμενης Μάθησης είναι από τη φύση τους επαγωγικά. Στην επιτηρούμενη μάθηση το σύνολο εκπαίδευσης περιέχει μόνο δεδομένα, στα οποία έχουν ανατεθεί τιμές, οπότε μας ενδιαφέρει μόνο η απόδοση του αλγόριθμου στα δεδομένα δοκιμής. Στην ημι-επιτηρούμενη μάθηση το σύνολο εκπαίδευσης περιέχει και δεδομένα χωρίς τιμή. Από τα δεδομένα χωρίς τιμή μπορούμε να βγάλουμε επιπλέον πληροφορίες για το σύνολο δεδομένων μας. Για αυτόν τον λόγο προτιμάται η χρήση ημι-επιτηρούμενης μάθησης, όταν θέλουμε να εφαρμόσουμε μεταγωγική μοντέλο.

Τα μοντέλα ημι-επιτηρούμενης μάθησης, ανάλογα τον στόχο τους, χωρίζονται σε δύο κατηγορίες. Η πρώτη κατηγορία έχει σαν στόχο την πρόβλεψη των τιμών σε δεδομένα δοκιμής. Η δεύτερη κατηγορία έχει σαν στόχο την πρόβλεψη των τιμών των απλών δεδομένων στο σύνολο εκπαίδευσης. Την πρώτη κατηγορία την ονομάζουμε Επαγωγική ημι-επιτηρούμενη μάθηση, ενώ τη δεύτερη Μεταγωγική ημι-επιτηρούμενη μάθηση.

Transductive Support Vector Machine

4.1 Εισαγωγή

Η βασική ιδέα του **Transductive Support Vector Machine** (**TSVM**) έγκειται στο γεγονός ότι χρησιμοποιεί τις πληροφορίες που παρέχουν τα δεδομένα χωρίς τιμές, για τον υπολογισμό του ορίου απόφασης. Σκοπός του αλγόριθμου είναι να βρει το όριο απόφασης που να έχει την μέγιστη απόσταση ανάμεσα στα πιο κοντινά σημεία των κλάσεων, που χωρίζει. Το αρχικό **Support Vector Machine** (**SVM**), έκανε χρήση μόνο των δεδομένων με τιμές για τον υπολογισμό του μέγιστου περιθωρίου ανάμεσα στο όριο απόφασης και τα σημεία των κλάσεων. Με την χρήση των πληροφοριών που μας παρέχουν τα απλά δεδομένα, μπορούμε να διαπιστώσουμε ότι υπάρχουν περιπτώσεις όπου το όριο απόφασης του SVM, διέρχεται από περιοχές υψηλής πυκνότητας. Γεγονός το οποίο δεν είναι επιθυμητό, αφού σκοπός είναι το όριο απόφασης να διέρχεται από περιοχές χαμηλής πυκνότητας. Με την χρήση του TSVM βρίσκουμε νέο όριο απόφασης όπου θα διαχωρίζει και τα δύο είδη δεδομένων, όπως φαίνεται στην Εικόνα 4.1.



Εικόνα 4.1.

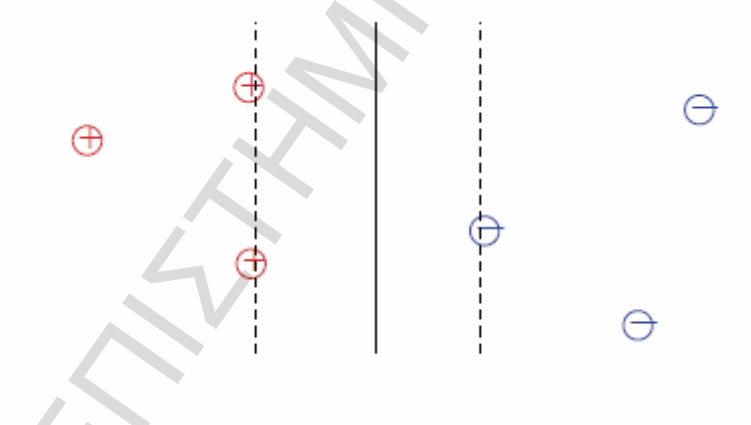
4.2 Support Vector Machine

4.2.1 Κεντρική Ιδέα

Τα Support Vector Machine (SVM) αποτελούν μοντέλα της επιτηρούμενης μάθησης, με αλγόριθμους μάθησης, οι οποίοι είναι σχεδιασμένοι για να αναλύουν δεδομένα και να αναγνωρίζουν πρότυπα, και χρησιμοποιούνται κυρίως για ταξινόμηση και για παλινδρόμηση. Το βασικό μοντέλο του SVM παίρνει σαν είσοδο ένα σύνολο δεδομένων και προβλέπει, για κάθε στοιχείο εισόδου σε ποια κλάση ανήκει. Δίνοντας του ένα σύνολο δεδομένων, για τα οποία έχουμε τις τιμές τους, για εκπαίδευση, ένας SVM αλγόριθμος κατασκευάζει ένα μοντέλο το οποίο αναθέτει τις τιμές στα νέα δεδομένα, δηλαδή τα αντιστοιχεί στην κλάση στην οποία ανήκουν.

Το SVM μοντέλο αποτελεί μια αναπαράσταση των παραδειγμάτων εισόδου, σαν ένα σύνολο από σημεία στον χώρο, και τα αναπαριστά με τέτοιο τρόπο, ώστε τα παραδείγματα που ανήκουν σε διαφορετικές κατηγορίες να χωρίζονται από ένα περιθώριο. Σκοπός του SVM μοντέλου είναι το κενό αυτό, να είναι όσο το δυνατόν πιο πλατύ. Τα νέα παραδείγματα απεικονίζονται στον ίδιο χώρο και το SVM προβλέπει σε ποια κατηγορία ανήκει, βασισμένο σε ποια πλευρά του κενού ανήκει.

Στην παρακάτω εικόνα απεικονίζεται το όριο απόφασης σαν ευθεία γραμμή. Με το όριο απόφασης αντιστοιχούμε τα στοιχεία εισόδου σύμφωνα με την κλάση στην οποία ανήκουν. Το περιθώριο του ορίου απόφασης απεικονίζεται με διακεκομμένες γραμμές και διέρχονται από τα σημεία της κάθε συστάδας που είναι πιο κοντά στο όριο απόφασης. Σκοπός του SVM είναι η μεγιστοποίηση αυτού του περιθωρίου.



Εικόνα 4.2.

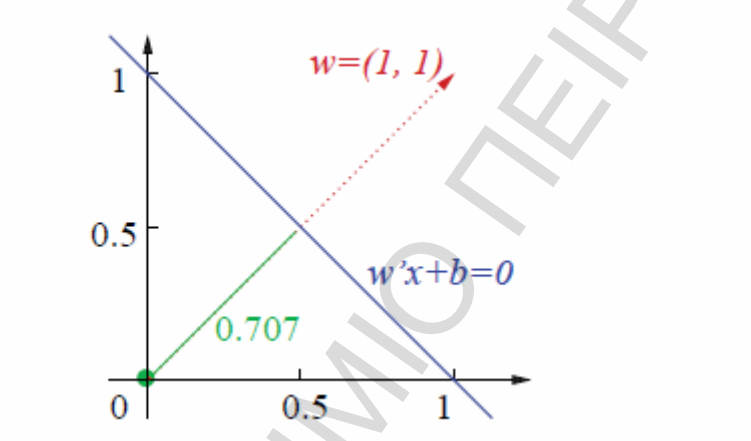
Το μοντέλο SVM ανήκει τυπικά στους γραμμικούς δυαδικούς ταξινομητές, δηλαδή να ταξινομεί τα στοιχεία εισόδου σε δύο κλάσεις. Γενικά όμως έχουμε τη δυνατότητα να εκτελούμε και μη γραμμικές ταξινομήσεις, χρησιμοποιώντας το λεγόμενο «kernel trick», σύμφωνα με το οποίο μπορούμε να απεικονίσουμε τα στοιχεία εισόδου σε διαστήματα πολλών διαστάσεων.

4.2.2 Ορισμός

Για πιο εύκολη ανάλυση του SVM υποθέτουμε ότι έχουμε δύο κλάσεις, η μία παίρνει τιμές -1 και η άλλη +1. Ακόμα υποθέτουμε ότι το όριο απόφασης είναι γραμμικό στο χώρο $\mathbf{R}^D$. Το όριο απόφασης θα δίνεται από τον τύπο:

$$\{x \mid w^T x + b = 0\},$$

όπου το w ανήκει στον χώρο $\mathbf{R}^D$ και είναι η παράμετρος που ορίζει την κλίση και την κλίμακα του ορίου απόφασης, και όπου το b ανήκει στον χώρο **R**. Το όριο απόφασης είναι πάντα κάθετο στο διάνυσμα w.



Εικόνα 4.3

Το όριο απόφασης χωρίζει τον χώρο σε δύο μέρη. Για την εύρεση του ορίου απόφασης ορίζουμε τη συνάρτηση $f(x) = w^T x + b$ και ψάχνουμε τις τιμές για $f(x) = 0$. Στο ένα μέρος του χώρου η συνάρτηση έχει τιμές μεγαλύτερες του μηδέν, ενώ για το άλλο μέρος θα έχει τιμές μικρότερες του μηδέν. Ορίζουμε ως signed distance την απόσταση ενός σημείου δεδομένων ως προς το όριο απόφασης με τον τύπο $yf(x)/\|w\|$.

Το signed distance είναι θετικό στην θετική πλευρά και αρνητικό στην αρνητική.

Υποθέτουμε ότι υπάρχει τουλάχιστον μία ευθεία γραμμή, η οποία να αποτελεί το όριο απόφασης, και η οποία να χωρίζει τα labeled data ώστε να βρίσκονται στην σωστή μεριά του ορίου απόφασης. Το περιθώριο σε αυτήν την περίπτωση, από το όριο απόφασης στο κοντινότερο σημείο δίνεται από τον τύπο:

$$\min_{i=1}^{l} y_i f(x_i)/\|w\|.$$

Σκοπός μας είναι η εύρεση ενός ορίου απόφασης που να μεγιστοποιεί το γεωμετρικό περιθώριο. Έτσι για να βρούμε το μέγιστο περιθώριο θα πρέπει να βρούμε τα μέγιστα w και b, για τα οποία θα ισχύει ο παραπάνω τύπος.

Με μετατροπές μπορούμε να καταλήξουμε στον παρακάτω πρόγραμμα, στο οποίο είναι πιο εύκολο να το βελτιστοποιήσουμε:

$$\min_{w,b} \|w\|^2 ,$$

$$y_i (w^T x_i + b) \geq 1 \text{ για } i = 1...l.$$

Στην περίπτωση που το όριο απόφασης δεν είναι γραμμικό, ο τύπος είναι ο εξής:

$$\min_{w,b,\xi} \sum_{i=1}^{l} \xi_i + \lambda \|w\|^2 ,$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i \text{ για } i = 1...l. \text{ και για } \xi_i \geq 0 ,$$

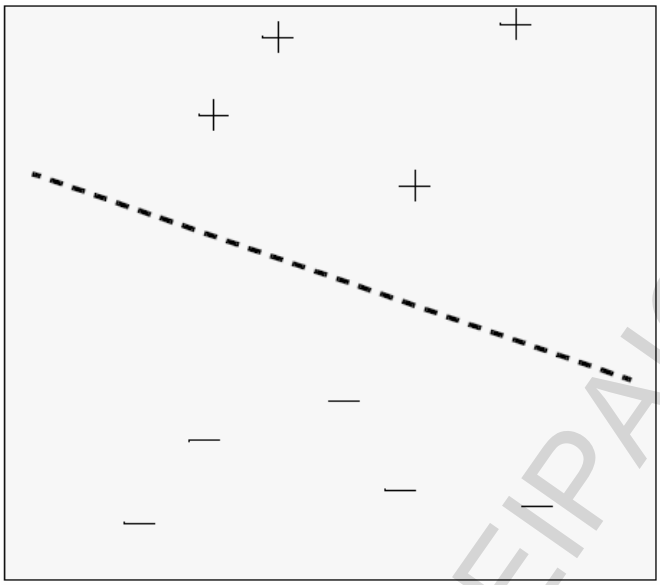
όπου ξ_i είναι οι μεταβλητές χαλαρότητας, το ποσό χαλαρότητας για κάθε παράδειγμα. Το άθροισμα τους αποτελεί το συνολικό ποσό χαλαρότητας, και σκοπός μας είναι να το ελαχιστοποιήσουμε μαζί με το τετράγωνο $||w||^2$. Το βάρος λ ισορροπεί τους δύο αυτούς στόχους. Αυτό το πρόγραμμα προσπαθεί να βρει τους την μέγιστη απόσταση του περιθωρίου, αλλά επιτρέπει σε μερικά σημεία του συνόλου εκπαίδευσης να είναι στην λάθος πλευρά από το όριο απόφασης.

4.3 Transductive Support Vector Machine (TSVM)

4.3.1 Κεντρική Ιδέα

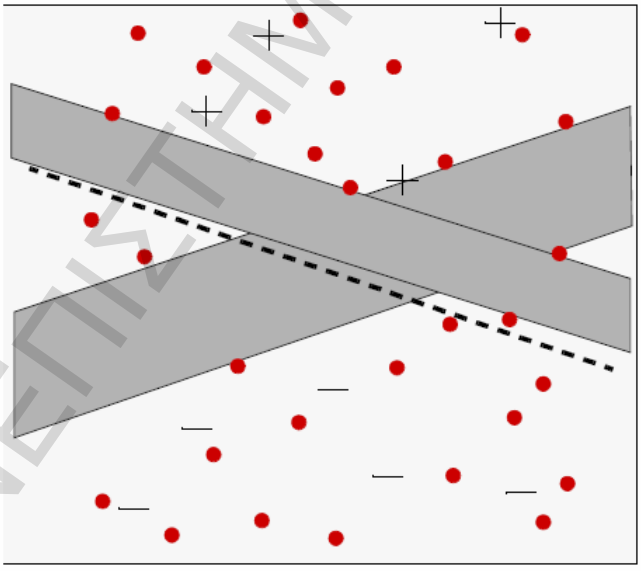
Τα Transductive support vector machines (TSVM) βασίζονται στην ιδέα της μεταγωγικής μάθησης, δηλαδή οι προβλέψεις για τις τιμές να γίνονται μόνο για τα σημεία του συνόλου δοκιμής (test points). Τα TSVM υλοποιούν την ιδέα περιλαμβάνοντας τα δεδομένα, για τα οποία δεν έχουμε τιμές, στον υπολογισμό του περιθωρίου.

Στην παρακάτω εικόνα απεικονίζεται, με διακεκομμένη γραμμή, το όριο απόφασης για το απλό SVM, όπου δεν λαμβάνονται υπόψη τα απλά δεδομένα, αλλά μόνο τα παραδείγματα με τιμές.



Εικόνα 4.4

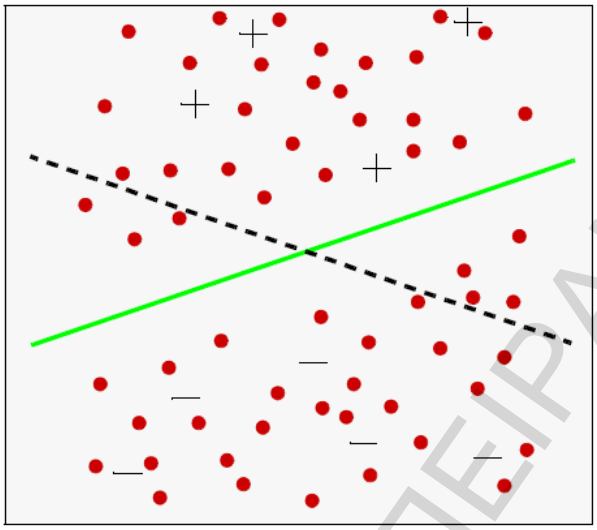
Στην παρακάτω εικόνα εμφανίζονται τα περιθώρια των που σχηματίζει το σύνολο των δεδομένων. Τα όρια των περιθωρίων διέρχονται από τα σημεία των συστάδων που είναι πιο κοντά στο όριο απόφασης.



Εικόνα 4.5

Στην παρακάτω εικόνα απεικονίζεται το όριο απόφασης του TSVM με ευθεία γραμμή, ενώ το όριο απόφασης του απλού SVM απεικονίζεται με διακεκομμένες γραμμές. Όπως παρατηρούμε, με την χρήση των απλών δεδομένων στον υπολογισμό του περιθωρίου, έχουμε περισσότερη πληροφορία για τη δομή των δεδομένων μας, και έτσι το όριο απόφασης μπορεί να είναι πιο

συγκεκριμένο. Με την προϋπόθεση πάντα ότι οι κλάσεις των δεδομένων είναι διαχωρισμένες σε επαρκή βαθμό.



Εικόνα 4.6

4.3.2 Ορισμός

Τα TSVM θεωρούν μια συγκεκριμένη σχέση ανάμεσα στο σύνολο διανυσμάτων των παραδειγμάτων $X = (x_1, x_2, \ldots, x_n)$, όπου $x_i \in R^d$, και τη διανομή των τιμών $P(y_1, y_2, \ldots, y_n)$. Σύμφωνα με την αρχή της επαγωγικής μάθησης, τα σημεία που μας ενδιαφέρουν και θα αναλύσουμε είναι πεπερασμένα σε αριθμό, οπότε και οι υποθετικές τιμές των τιμών θα είναι και αυτές πεπερασμένες σε αριθμό. Τα TSVM κατασκευάζουν μια δομή στον υποθετικό χώρο H, η οποία βασίζεται στο περιθώριο των υπερεπίπεδων:

$$y_i(w^T x_i + b) \geq 1 - \xi_i$$

όλων των σημείων του συνόλου των παραδειγμάτων, δηλαδή περιλαμβάνοντας τα διανύσματα εκπαίδευσης και δοκιμής.

Το περιθώριο ενός υπερεπιπέδου του $X = (x_1, x_2, \ldots, x_n)$ είναι η ελάχιστη απόσταση του ορίου απόφασης από το πιο κοντινό διάνυσμα στο σύνολο X.

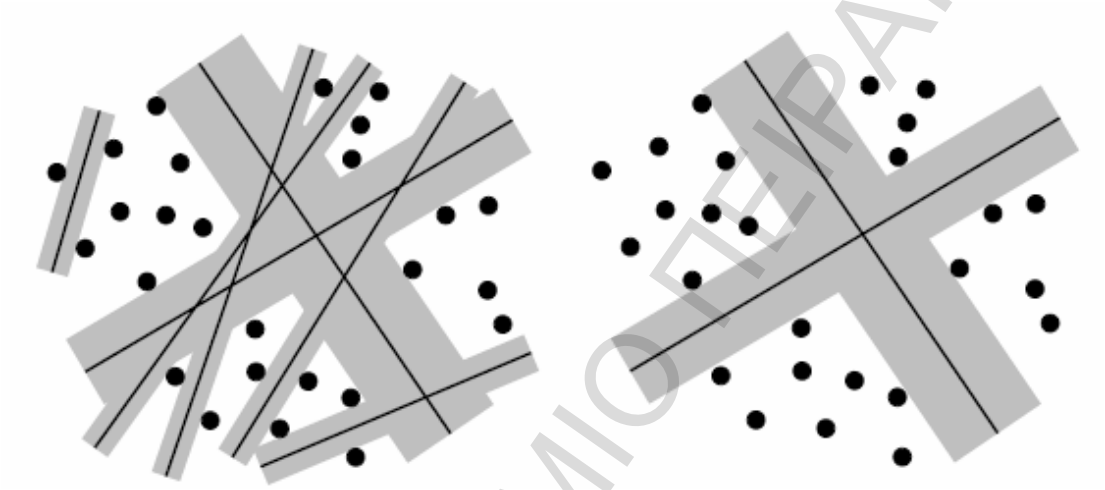
$$\min_{iE[1..n]} \left[\frac{y_i}{\|w\|} (w^* x_i + b) \right]$$

Το στοιχείο δομής H_ρ περιέχει όλες τις τιμές των σημείων του συνόλου X, τα οποία μπορούν να επιτευχθούν με τους ταξινομητές υπερεπιπέδου

$$h(x) = sign\{w^* x_i + b\},$$

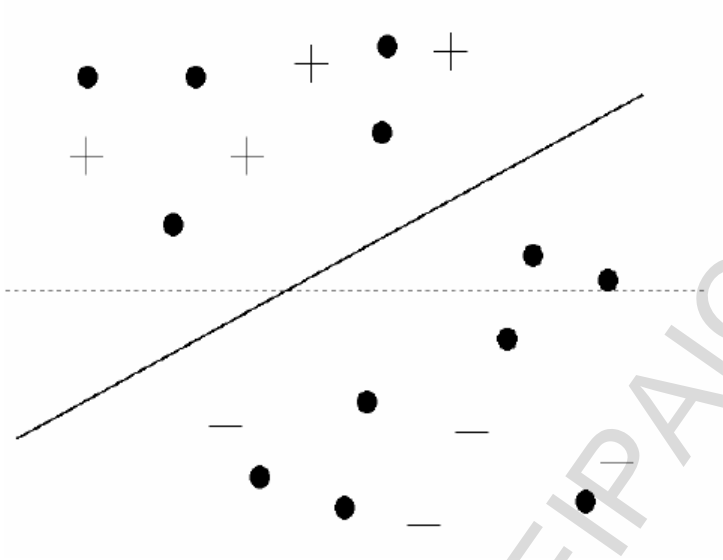
οι οποίοι έχουν περιθώριο μεγαλύτερο του ρ στο σύνολο X. Ο υπολογισμός των τιμών εξαρτάται σε μεγάλο βαθμό από το μέγεθος του περιθωρίου. Όταν μεγαλώνει το περιθώριο, τότε ο αριθμός των πιθανών συνόλων τιμών για τα σημεία του συνόλου X, μικραίνει.

Στο παρακάτω σχήμα απεικονίζεται ο ρόλος που παίζει το μέγεθος του περιθωρίου των υπερεπιπέδων, στην ανάθεση των τιμών. Τα παραδείγματα απεικονίζονται με τελείες, ενώ το περιθώριο των υπερεπιπέδων απεικονίζεται με τη γκρι περιοχή. Βλέπουμε από τα σχήματα ότι όταν έχουμε μικρά περιθώριο έχουμε μεγαλύτερο αριθμό πιθανών τιμών. Αντίθετα όταν αυξάνουμε το κατώφλι ρ του περιθωρίου, τότε ο αριθμός των πιθανών τιμών μειώνεται.



Εικόνα 4.7

Στην περίπτωση του TSVM, όπου λαμβάνουμε υπόψη και τα παραδείγματα με τιμές. Στο παρακάτω σχήμα απεικονίζουμε το αποτέλεσμα με τη χρήση παραδειγμάτων στα οποία έχουμε αναθέσει τιμές. Τα θετικά και αρνητικά παραδείγματα απεικονίζονται με το σύμβολο + και -, αντίστοιχα. Η διακεκομμένη γραμμή είναι η λύση του απλού SVM, το οποίο όπως έχουμε αναφέρει, βρίσκει το όριο διαχωρισμού με βάση το υπερεπίπεδο, το οποίο διαχωρίζει τα δεδομένα εκπαίδευσης με το μεγαλύτερο περιθώριο. Η συνεχής γραμμή δείχνει το αποτέλεσμα του TSVM. Το όριο διαχωρισμού σε αυτή τη περίπτωση βασίζεται στην ανάθεση τιμών, η οποία έχει μηδενικό σφάλμα εκπαίδευσης και το μεγαλύτερο περιθώριο, στο οποίο λαμβάνεται υπόψη τα διανύσματα της εκπαίδευσης και της δοκιμής. Το TSVM επιλέγει τη λύση, στην οποία τα label να είναι σε ευθυγράμμιση με τη δομή των συστάδων, τόσο στα παραδείγματα εκπαίδευσης, όσο και στα παραδείγματα δοκιμής.



Εικόνα 4.8

Είναι εύκολο να φανταστεί κανείς ότι άμα βασιστούμε, για τον υπολογισμό των labels, πρωτίστως στο περιθώριο που πρέπει να έχει το όριο απόφασης από τα σημεία του παραδείγματος, τότε δίνεται προτεραιότητα σε τιμές που ανήκουν σε διακριτές συστάδες, των οποίων το όριο απόφασης περνάει από υψηλής πυκνότητας περιοχές. Το μέγεθος του περιθωρίου ρ μπορεί να χρησιμοποιηθεί για τον έλεγχο της πληθικότητας του αντίστοιχου συνόλου τιμών H_ρ . Το παρακάτω θεώρημα του Vapnik παρέχει ανώτατο όριο στον αριθμό των τιμών $|H_\rho|$, που μπορούν να επιτευχθούν με υπερεπίπεδα, τα οποία έχουν περιθώριο τουλάχιστον ρ:

Για οποιοδήποτε n διανύσματα $x_1, \ldots, x_n \in R^d$, τα οποία περιέχονται σε σφαίρα με διάμετρο R, ο αριθμός $|H_\rho|$ των πιθανών δυαδικών labels $y_1, \ldots, y_n \in \{ -1, +1\}$, τα οποία προκύπτουν από τους υπερεπίπεδους ταξινομητές $h(x) = \text{sign} \{x * w + b\}$, με περιθωρίο τουλάχιστον ρ,

$$\forall_{i=1}^{n} : \frac{y_i}{\|w\|}\left[w * x_i + b\right] \geq \rho$$

Είναι περιορισμένοι από

$$\left|\mathrm{H}_{\rho}\right| \leq e^{d\left(\ln\frac{n+k}{d}+1\right)} \quad ,$$

$$d = \frac{R^2}{\rho^2} + 1$$

Ο αριθμός των δυνατών συνόλων τιμών $|H_\rho|$, δεν εξαρτάται αναγκαστικά από τον αριθμό των διαστάσεων του συνόλου των σημείων. Το TSVM ταξινομεί όλα τα σύνολα τιμών με βάση το

περιθώριο ρ στο Χ. Για την ειδική περίπτωση που απαιτείται μηδενικό λάθος εκπαίδευσης, βελτιστοποίηση του ορίου απόφασης σημαίνει την εύρεση του συνόλου τμών με το μεγαλύτερο περιθώριο σε όλο το σύνολο των διανυσμάτων.

4.3.3 Τεχνικές Βελτιστοποίησης

Το παραπάνω πρόβλημα αποτελεί το **Πρόβλημα Βελτιστοποίησης** (**Optimization Problem – OP**).

Το πρόβλημα βελτιστοποίησης έχει σαν αντικείμενο την ελαχιστοποίηση του συνόλου:

$$\vee (y^*_{u1},..., y^*_{uu}, w, b) = \frac{1}{2} * w * w,$$

το οποίο πρέπει να υπακούει στους παρακάτω περιορισμούς:

$$\forall^{l}_{i=1} : y_{li}[\vec{w} * x_{li} + b] \geq 1,$$
$$\forall^{u}_{j=1} : y^{8}_{uj}[\vec{w} * x^*_{uj} + b] \geq 1,$$
$$\forall^{u}_{j=1} : y^*_{uj} E\{-1,+1\}$$

Η λύση του προβλήματος αυτού είναι η εύρεση των τιμών αυτών των δεδομένων δοκιμής, για τις οποίες το υπερεπίπεδο, το οποίο διαχωρίζει τα δεδομένα εκπαίδευσης και δοκιμής, έχει το μέγιστο περιθώριο. Το απλό SVM υπολογίζει ένα υπερεπίπεδο με μέγιστο περιθώριο, λαμβάνοντας όμως υπόψη μόνο τα δεδομένα εκπαίδευσης και όχι τα δεδομένα δοκιμής στους υπολογισμούς του.

Ένας τρόπος για να αποφύγουμε περίπλοκες λύσεις στην πρόβλεψη των τιμών ενός συνόλου δεδομένων είναι η **Τοπική Μάθηση (Local Learning**). Η ιδέα της τοπικής μάθησης είναι ότι δοθέντος ενός σημείου δοκιμής, μια σωστή προσέγγιση είναι να επικεντρωθούμε στα σημεία εκπαίδευσης, τα οποία βρίσκονται σε κοντινή ακτίνα από αυτό το σημείο, να κατασκευάσουμε έναν τοπικό κανόνα απόφασης, και να προβλέψουμε την τιμή του σημείου σύμφωνα με αυτόν μόνο τον κανόνα. Η ιδέα της τοπικής μάθησης μπορεί να βρίσκεται και στην υλοποίηση του TSVM. Μπορούμε να χρησιμοποιήσουμε σαν απλά σημεία, σημεία από το σύνολο δοκιμής, αντί να επιλέξουμε τυχαία σημεία, και έτσι θα έχουμε το πλεονέκτημα ότι ο αλγόριθμος θα εστιάζεται στις περιοχές του χώρου όπου είναι σημαντικό να είναι ακριβής για τους υπολογισμούς μας.

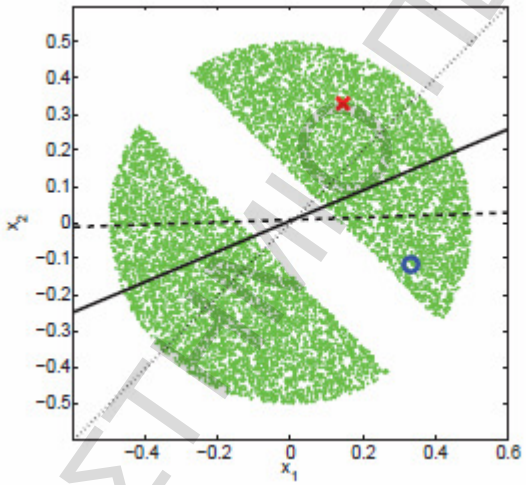
Τα μοντέλα SVM και TSVM είναι μη πιθανολογικά μοντέλα. Έτσι δεν αποτελούν μοντέλα σχεδιασμένα για να υπολογίζουν την μεταγενέστερη (a posteriori) πιθανότητα p(y|x), κατά τη διάρκεια της ταξινόμησης. Στην μηχανική μάθηση υπάρχουν πολλά άλλα μοντέλα, τα οποία υπολογίζουν την πιθανότητα p(y|x) από τα δεδομένα με τιμές του συνόλου εκπαίδευσης για την ταξινόμηση.

4.4 Μειονεκτήματα

Στο μοντέλο TSVM παίρνουμε σαν βασική υπόθεση, ότι θεωρούμε τις κλάσεις ότι είναι καλά χωρισμένες. Σε αυτή τη περίπτωση το όριο απόφασης τοποθετείται σε περιοχή με χαμηλή πυκνότητα, και έτσι δεν περνάει μέσα από περιοχή με μεγάλη πυκνότητα δεδομένων χωρίς τιμή.

Στην περίπτωση που οι κλάσεις δεν είναι καλά διαχωρισμένες, υπάρχει πρόβλημα στην απόδοση του μοντέλου TSVM. Στη συνέχεια θα παρουσιάσουμε ένα σενάριο το οποίο υποδεικνύει την αδυναμία αυτή του TSVM. Η διανομή p(x) του παραδείγματος μας είναι ενιαία και σχηματίζει έναν δίσκο. Το δίσκο τον διαπερνάει ένα κενό, το οποίο περνάει από τα σημεία όπου το y = - x. Στις περιοχές αυτές η πυκνότητα είναι ίση με το μηδέν. Στο παράδειγμα μας όμως, το όριο που διαχωρίζει τις κλάσεις είναι η διαγώνιος όπου ισχύει y = x. Οπότε έχουμε να κάνουμε με μια περίπτωση όπου οι κλάσεις δεν είναι καλά διαχωρισμένες, ακόμα και αν τα δεδομένα σχηματίζουν εμφανείς συστάδες. Είναι εμφανές ότι το όριο απόφασης δεν βρίσκεται στο περιθώριο όπου υπάρχει μικρή πυκνότητα.

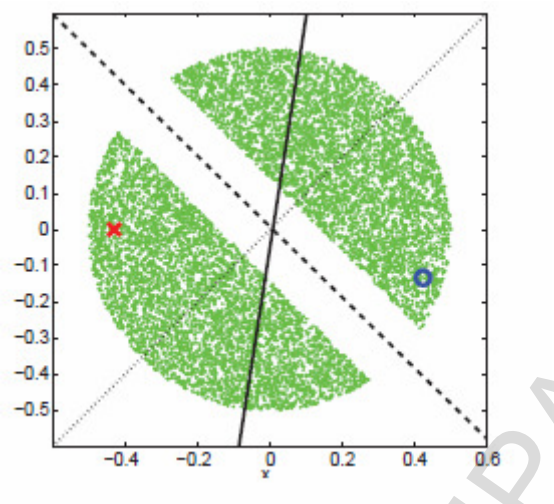
Σε αυτή την περίπτωση υπάρχουν δύο ενδεχόμενα. Στο πρώτο ενδεχόμενο, αν τα δεδομένα με τιμή εμφανίζονται στην ίδια μεριά από τα το κενό, τότε το μοντέλο TSVM στην προσπάθεια του να βρει το κενό ανάμεσα στις δύο κλάσεις, μπορεί να λάβει υπόψη του μικρά κενά που υπάρχουν ανάμεσα στα απλά δεδομένα. Το όριο απόφασης σε αυτό το ενδεχόμενο θα είναι χειρότερο από αυτό του απλού SVM, το οποίο δεν λαμβάνει καθόλου υπόψη του τα δεδομένα χωρίς τιμή, αλλά τουλάχιστον θα προσεγγίζει το πραγματικό όριο απόφασης.



Εικόνα 4.9

Τα δεδομένα χωρίς τιμή αναπαριστώνται στην εικόνα με τις πράσινες κουκίδες. Το θετικό δεδομένο με το σύμβολο “o” και το αρνητικό με το σύμβολο “x”. Το πραγματικό όριο απόφασης με τη λεπτή διακεκομμένη γραμμή, το όριο απόφασης του SVM με την ευθεία, και το όριο απόφασης του TSVM με την πιο χοντρή διακεκομμένη γραμμή. Στο σχήμα μπορούμε να δούμε ότι αν και εσφαλμένο, το όριο απόφασης του TSVM προσεγγίζει ως ένα σημείο το πραγματικό όριο απόφασης.

Το ενδεχόμενο, όπου και το αποτέλεσμα του TSVM απέχει αρκετά από το ορθό αποτέλεσμα, πραγματοποιείται όταν τα δεδομένα με τιμές δεν βρίσκονται στην ίδια μεριά του κενού. Τότε το TSVM θα προτιμήσει σαν όριο απόφασης το κενό ανάμεσα στα δεδομένα χωρίς τιμή σαν όριο απόφασης, με αποτέλεσμα να έχουμε προβλέψεις με μεγάλο σφάλμα. Ειδικά στην περίπτωση του παραδείγματός μας, ο ταξινομητής θα κάνει λάθος στις μισές προβλέψεις του.



Εικόνα 4.10

Στο σχήμα παρατηρούμε το μέγεθος του σφάλματος του TSVM στο ενδεχόμενο όπου τα παραδείγματα με τιμή βρίσκονται εκατέρωθεν του κενού στα δεδομένα χωρίς τιμή. Το όριο απόφασης του TSVM είναι ακριβώς κάθετο στο πραγματικό όριο απόφασης, με αποτέλεσμα ο ταξινομητής να έχει μεγάλο ποσοστό λάθους στις προβλέψεις του.

Περιγραφή Υλοποίησης του Συστήματος

Σκοπός της εργασίας είναι η δημιουργία ενός συστήματος, το οποίο θα μπορεί να ανακτά εικόνες από ένα σύνολο εικόνων, με βάση το περιεχόμενο τους. Για κάθε μία από τις εικόνες που έχουμε στη διάθεσή μας, θα εξάγουμε ένα σύνολο οπτικών, χαμηλού επιπέδου, χαρακτηριστικών τους. Αυτά τα σύνολα χαρακτηριστικών θα τα αξιοποιήσουμε για την εκπαίδευση ενός TSVM αλγορίθμου, ο οποίος θα έχει σκοπό την ταξινόμηση των εικόνων με βάση το κύριο θέμα τους.

5.1 Εξαγωγή Χαρακτηριστικών

Για την αναπαράσταση των χαρακτηριστικών χαμηλού επιπέδου από τις εικόνες, επιλέγουμε τους MPEG-7 οπτικούς περιγραφείς. Το standard MPEG-7, επικεντρώνεται στην περιγραφή του multimedia περιεχομένου, προσφέροντας ένα σύνολο περιγραφέων χαμηλού επιπέδου. Για τον σκοπό αυτό χρησιμοποιήσαμε την εφαρμογή VDE, η οποία ειδικεύεται στην εξαγωγή οπτικών περιγραφέων από το οπτικό περιεχόμενο εικόνων. Η εφαρμογή VDE εξάγει τα χαρακτηριστικά των εικόνων σε μορφή διανύσματος.

Για την εξαγωγή χαρακτηριστικών χαμηλού επιπέδου από τις εικόνες, επιλέξαμε να εξάγουμε περιγραφείς από το σύνολο της εικόνας. Τα χαρακτηριστικά που επιλέγουμε να εξάγουμε είναι αυτά που σχετίζονται με το χρώμα και την υφή των εικόνων. Πιο συγκεκριμένα έχουμε χρησιμοποιήσει MPEG-7 περιγραφείς χρώματος και υφής εικόνας για να πιάσουμε τα χαμηλού επιπέδου χαρακτηριστικά των εικόνων.

Οι περιγραφείς που χρησιμοποιήσαμε για την εξαγωγή των χαρακτηριστικών είναι ο **Dominant Color Layout Descriptor** (**CLD**) και ο **Edge Histogram Descriptor** (**EHD**).

Ο Color Layout Descriptor είναι ένας MPEG-7 οπτικός περιγραφέας, σχεδιασμένος για να παριστάνει την χωρική διανομή του χρώματος στο διάστημα χρώματος YCbCr. Η εικόνα, από την οποία εξάγουμε τα χαρακτηριστικά της, διαιρείται σε 8 x 8 = 64 τετράγωνα και το μέσο χρώμα του κάθε τετραγώνου υπολογίζεται σαν το αντιπροσωπευτικό χρώμα του. Ως DCT μετατροπή εκτελείται πάνω στις σειρές των αντιπροσωπευτικών χρωμάτων και επιλέγονται

μερικά χαμηλής συχνότητας συντελεστές. Ο CLD τελικά διαμορφώνεται μετά τον κβαντισμό των εναπομεινάντων συντελεστών.

Ο **Edge Histogram Descriptor** (**EHD**) συλλαμβάνει την χωρική διανομή των ακρών. Ο EHD διαιρεί την εικόνα σε 4 x 4 μικρότερες εικόνες και έτσι η διανομή των ακρών για κάθε μικρότερη εικόνα μπορεί να αναπαρασταθεί από ένα ιστόγραμμα. Για να παραχθεί ένα ιστόγραμμα, οι άκρες στις μικρότερες εικόνες κατηγοριοποιούνται σε πέντε τύπους: σε κάθετες, οριζόντιες, διαγώνιες 45 μοιρών, διαγώνιες 135 μοιρών και σε χωρίς κατεύθυνση άκρες. Για κάθε μικρότερη εικόνα που διαιρείται η αρχική εικόνα, χρειάζονται πέντε ιστογράμματα. Αφού κάθε εικόνα αποτελείται από 16 μικρότερες, τότε για κάθε εικόνα θα χρειαζόμαστε 80 ιστογράμματα.

Καθένας από τους περιγραφείς που επιλέξαμε να εξάγουμε, θα αποτελείται από ένα διάνυσμα με προκαθορισμένες διαστάσεις, εκτός της περίπτωσης του Dominant Color Descriptor, το οποίο θα αποτελείται από τις τιμές και το ποσοστό του επικρατέστερου χρώματος. Για να έχουμε μια μοναδική περιγραφή για κάθε εικόνα της συλλογής μας, τα διανύσματα των περιγραφέων συγχωνεύονται σε ένα διάνυσμα.

Η μορφή των διανυσμάτων θα είναι η εξής:

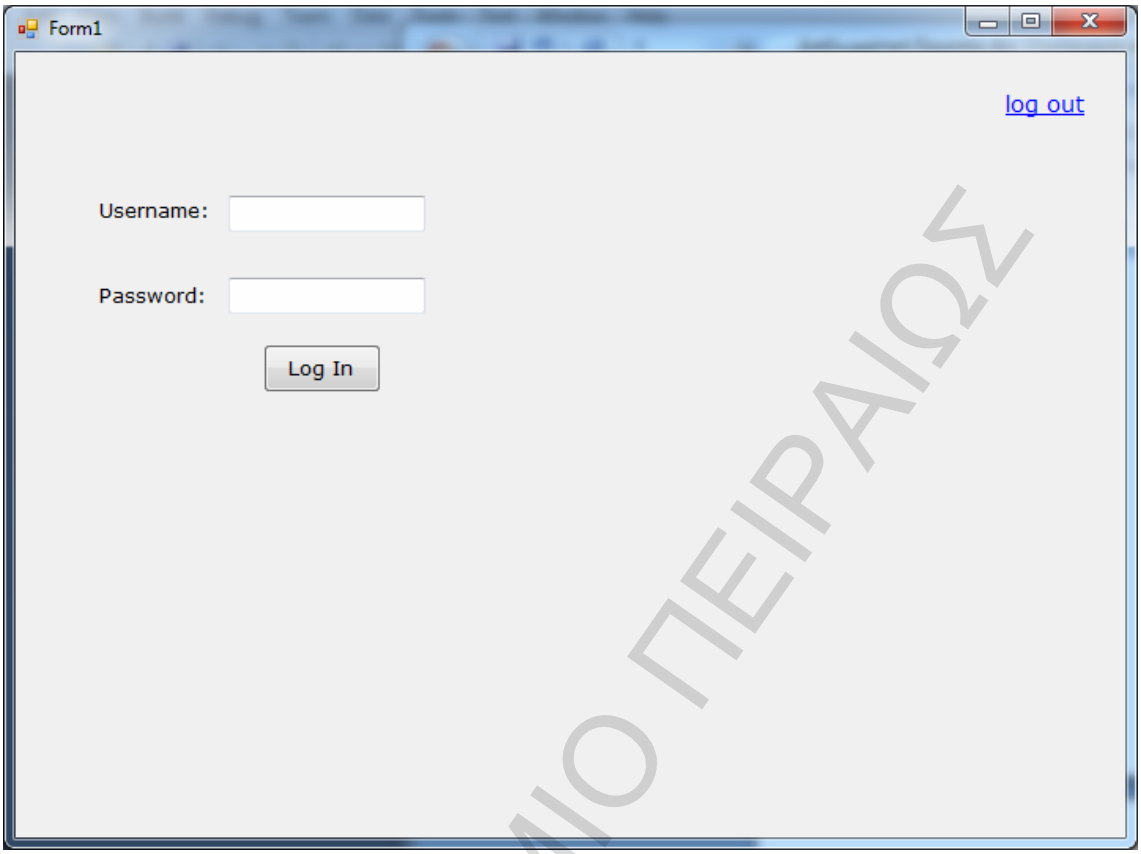
$$f_i = f(r_i) = [CLD(r_i), EHD(r_i)], \forall r_i \exists R$$

Η εξαγωγή των χαμηλού επιπέδου περιγραφέων πραγματοποιείται με τη χρήση της εφαρμογής **Visual Descriptor Extraction** (**VDE**). Αυτή η εφαρμογή χρησιμοποιείται για να εξάγει τους MPEG-7 περιγραφείς, της επιλογής μας, από τις εικόνες που έχουμε επιλέξει. Η ανάπτυξη της εφαρμογής VDE έχει βασιστεί στο eXperimentation Model του MPEG-7, χρησιμοποιώντας τους αλγόριθμους εξαγωγής του.

Για την καλύτερη οργάνωση του προγράμματος μας, επιλέξαμε να υπάρχει ένας χρήστης, ο οποίος θα έχει αυξημένες αρμοδιότητες σε σχέση με τους υπόλοιπους χρήστες, και ο οποίος θα είναι επιφορτισμένος με τον ρόλο του administrator. Με αυτόν τον τρόπο, οι χρήστες θα έχουν διακριτούς ρόλους και συγκεκριμένα δικαιώματα, όσον αφορά τις ενέργειες που μπορούν να εκτελέσουν.

5.2 Χρήση του προγράμματος

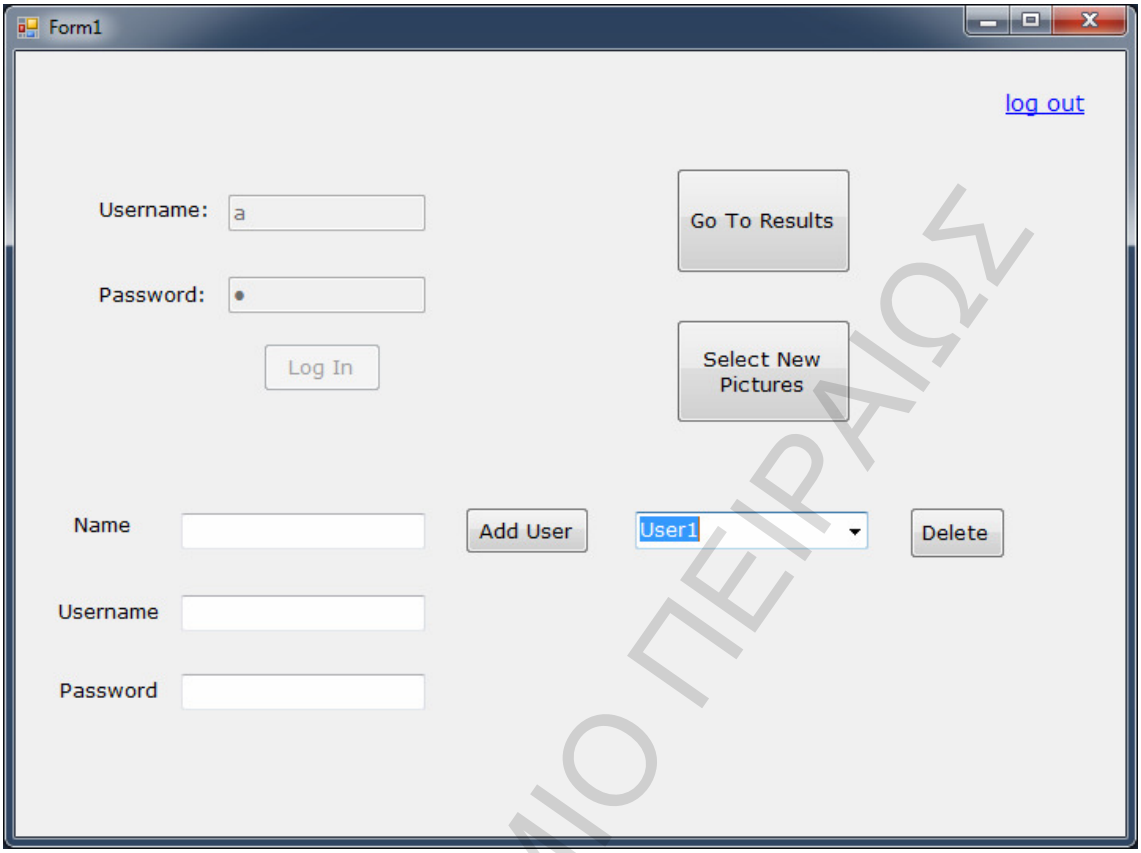
Η είσοδος των χρηστών στο πρόγραμμα θα γίνεται με την είσοδο του ονόματος του χρήστη (username) και του συνθηματικού του (password) , που θα τους έχουν δοθεί. Φυσικά το όνομα του χρήστη θα είναι πάντα μοναδικό, για να μπορούν να διακριθούν οι χρήστες. Η είσοδος στο πρόγραμμα θα έχει αυτήν την μορφή:



Εικόνα 5.1.

5.2.1 Ενέργειες του χρήστη administrator

Ο χρήστης που θα εκτελεί χρέη administrator, θα έχει τη δυνατότητα να επιλέγει τις εικόνες που θα αποτελέσουν το σύνολο δεδομένων του συστήματός μας. Ακόμα θα έχει τη δυνατότητα να διαχειρίζεται τους χρήστες στο σύστημα, δημιουργώντας νέους, δίνοντας τους το όνομα τους, το username και το αντίστοιχο password. Ακόμα θα έχει τη δυνατότητα να τους διαγράψει. Η φόρμα της σελίδας του administrator θα έχει την εξής μορφή:



Εικόνα 5.2.

Η λίστα στα δεξιά θα δείχνει τα ονόματα όλων των χρηστών. Τα πεδία κάτω αριστερά χρησιμοποιούνται για την εισαγωγή του ονόματος, του username και του password του νέου χρήστη. Όταν ο administrator προσθέσει νέο χρήστη, τότε το πρόγραμμα τον προσθέτει στη βάση δεδομένων στον πίνακα που περιέχει τους χρήστες. Ο admin θα έχει τη δυνατότητα να διαγράψει τον χρήστη και τα στοιχεία του στη συνέχεια, αν επιθυμεί.

Η επιλογή των εικόνων γίνεται από το κουμπί “Select New Pictures”, με το οποίο θα μπορούμε να περιηγούμαστε στα αρχεία του υπολογιστή και να επιλέγουμε τον φάκελο, ο οποίος θα περιέχει τις εικόνες που θέλουμε να προσθέσουμε στο σύστημα. Αφού σκοπός μας είναι να μπορεί το σύστημα να κατατάσσει ένα σύνολο εικόνων σε συγκεκριμένα κλάση, ανάλογα με το θέμα που απεικονίζουν, επιλέξαμε να οργανώσουμε τις εικόνες σε φακέλους ανάλογα με το θέμα τους. Παραδείγματος χάρη στον φάκελο με το όνομα “Landscapes”, θα έχουμε αποθηκεύσει όλες τις εικόνες που θα έχουν κύριο θέμα την απεικόνιση τοπίων. Έτσι πετυχαίνουμε να έχουμε μια πιο εύκολη και γρήγορη διαχείριση των εικόνων και των θεμάτων.

Όταν ο χρήστης, που έχει τον ρόλο του administrator, επιλέξει τον φάκελο με τις εικόνες, τότε το πρόγραμμα θα επεξεργαστεί τις εικόνες και τα στοιχεία τους και μετά θα τα αποθηκεύσει στη βάση δεδομένων μας. Στην αρχή το πρόγραμμα θα μετονομάσει κάθε εικόνα, δίνοντας της μοναδικό όνομα, ώστε να μην υπάρξουν εικόνες με την ίδια ονομασία, αλλά και για να είναι πιο εύκολη η διαχείριση των δεδομένων μας, από τη στιγμή που το σύνολο των εικόνων θα είναι αρκετά μεγάλο. Η ονομασία της κάθε εικόνας θα αποτελείται από το όνομα του φακέλου στον οποίο ανήκει και το έναν αύξοντα αριθμό. Υπενθυμίζουμε ότι η ονομασία του φακέλου θα είναι και το υψηλού επιπέδου θέμα των εικόνων που περιέχει.

Μετά την μετονομασία των εικόνων το πρόγραμμα θα χρησιμοποιεί την εφαρμογή VDE, για την εξαγωγή των χαρακτηριστικών χαμηλού επιπέδου από το σύνολο των εικόνων. Επιλέγουμε τα στοιχεία εξόδου της εφαρμογής να είσαι σε μορφή απλού κειμένου και να σώζονται σε αρχεία κειμένου (.txt). Όταν τελειώσει η διαδικασία θα έχουν παραχθεί τόσα αρχεία κειμένου όσα και οι εικόνες. Αυτά τα αρχεία θα περιέχουν τα χαρακτηριστικά της κάθε εικόνας σε μορφή διανύσματος.

Στη συνέχεια το πρόγραμμα θα αποθηκεύει τα θέματα σε έναν πίνακα στη βάση δεδομένων μας. Αυτός ο πίνακας θα περιέχει το σύνολο των θεμάτων, από την οποία οι χρήστες αργότερα θα καλούνται να επιλέξουν το θέμα με το οποίο θα κατατάσσουν τις εικόνες.

Στο τέλος το πρόγραμμα θα καταχωρεί τα παραπάνω δεδομένα σε έναν πίνακα στη βάση δεδομένων, ο οποίος θα περιέχει τα στοιχεία των εικόνων. Πιο συγκεκριμένα ο πίνακας θα περιέχει τα ακόλουθα πεδία:

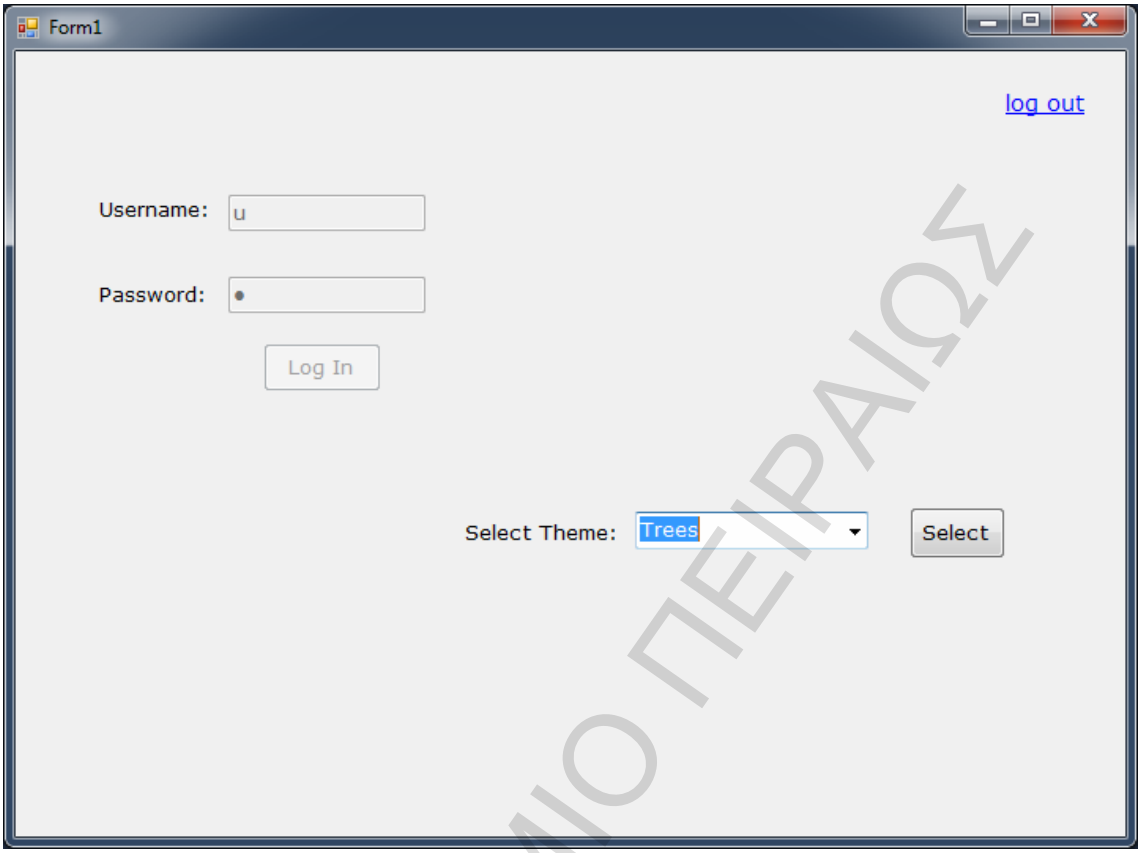
- Code, ο μοναδικός κωδικός της εικόνας,
- Name, το όνομα της εικόνας,
- Path, η τοποθεσία της εικόνας στον υπολογιστή,
- Concept, το βασικό θέμα που περιγράφει την εικόνα και
- Vector, το διάνυσμα με τα χαρακτηριστικά της εικόνας.

Για να προσθέσουμε τα χαρακτηριστικά των εικόνων στη βάση δεδομένων, επεξεργαζόμαστε τα αρχεία κειμένου που παρήχθησαν με την εκτέλεση της εφαρμογής VDE, και αποθηκεύουμε τα διανύσματα που περιέχουν στη βάση δεδομένων, και πιο συγκεκριμένα στο πεδίο Vector.

Ο administrator θα έχει επίσης την επιλογή να επιλέγει την προβολή των αποτελεσμάτων του συστήματος. Πατώντας το κουμπί “Go To Results” θα του εμφανίζεται μια λίστα με το σύνολο των χρηστών, που έχουν δώσει βαθμολογίες, και μια λίστα με τις θεματικές ενότητες που έχουν επιλέξει αυτοί οι χρήστες. Επιλέγοντας τον χρήστη και τη θεματική ενότητα, θα εμφανίζονται οι προβλέψεις του συστήματος για το σύνολο των εικόνων, που δεν έχει βαθμολογήσει ο χρήστης, καθώς και το σφάλμα των αποτελεσμάτων. Τα αποτελέσματα εμφανίζονται με τη μορφή διαγραμμάτων.

5.2.2 Ενέργειες των χρηστών

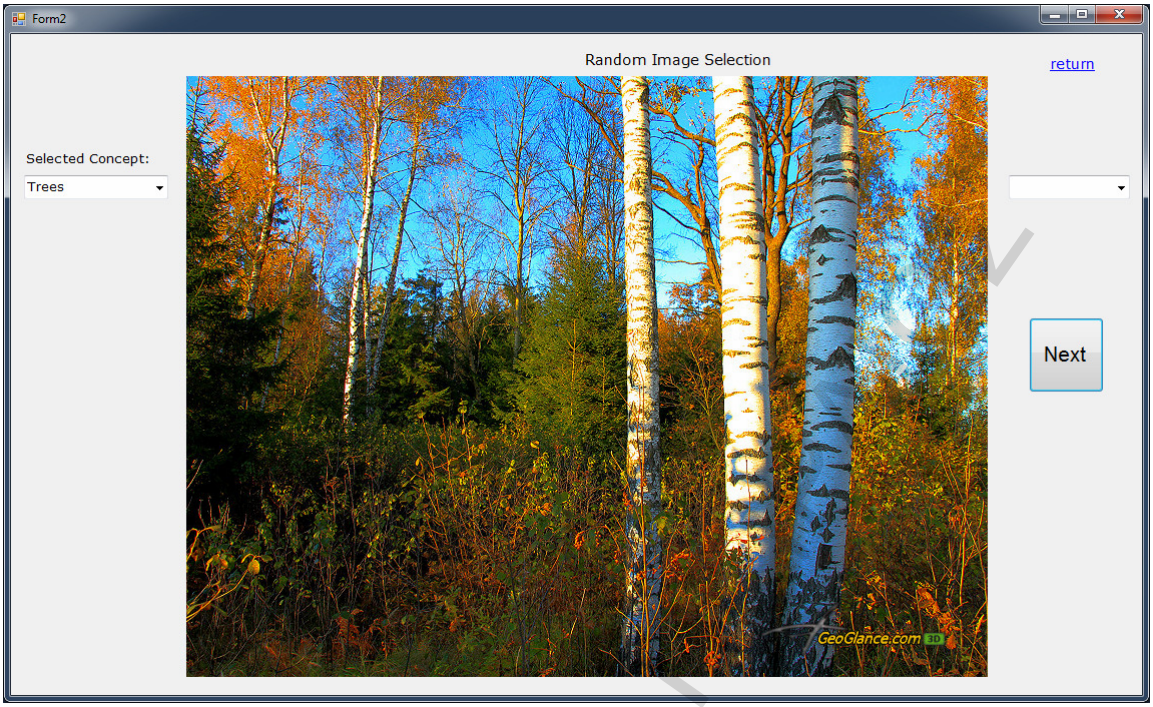
Οι χρήστες για να εισέλθουν στο πρόγραμμα θα πρέπει να δώσουν το username και το password τους. Η σελίδα του χρήστη έχει την εξής μορφή:



Εικόνα 5.3.

Η λίστα κάτω δεξιά περιέχει όλα τα υψηλού επιπέδου θέματα, στα οποία έχουμε κατατάξει τις εικόνες μας. Ο χρήστης θα επιλέγει σε κάθε session το θέμα με το οποίο θα βαθμολογεί τις εικόνες. Στην παραπάνω εικόνα έχει επιλέξει το θέμα “Cats”, οπότε θα βαθμολογήσει τις εικόνες με βάση αυτό το θέμα.

Οι εικόνες προβάλλονται στον χρήστη σε τυχαία σειρά και χωρίς να του παρουσιάζονται οι ονομασίες τους. Ο χρήστης μπορεί να βαθμολογεί την κάθε εικόνα με βάση μια κλίμακα από το ένα ως το πέντε. Θα βαθμολογεί τις εικόνες ως προς τη συνάφεια τους ως προς το θέμα που έχει επιλέξει για το συγκεκριμένο session.



Εικόνα 5.4.

Όπως βλέπουμε στην παραπάνω εικόνα, ο χρήστης έχει επιλέξει το γενικό θέμα “Trees” για να βαθμολογήσει τις εικόνες. Αριστερά από τις φωτογραφίες φαίνεται το όνομα του θέματος που έχει επιλέξει. Δεξιά από τις εικόνες είναι η λίστα με την οποία ο χρήστης θα βαθμολογεί τις φωτογραφίες. Με το κουμπί “Next” θα μπορεί να περιηγείται στις εικόνες που έχουμε αποθηκεύσει στη βάση δεδομένων. Στο πρώτο session τα θέματα των εικόνων, όπως και οι εικόνες, θα εναλλάσσονται με τυχαία σειρά, προσπαθώντας έτσι να κάνουμε πιο αντικειμενική την κρίση του χρήστη.

Ο χρήστης θα έχει τη δυνατότητα να διακόψει το session και να το συνεχίσει άλλη χρονική στιγμή. Ακόμα το πρόγραμμα αποθηκεύει τις προηγούμενες επιλογές του. Έτσι όταν προβάλλει εικόνα που ήδη έχει βαθμολογήσει, θα του παρουσιάζεται η προηγούμενη βαθμολογία. Τότε θα έχει τη δυνατότητα να αλλάξει την προηγούμενη βαθμολογία εάν το επιθυμεί.

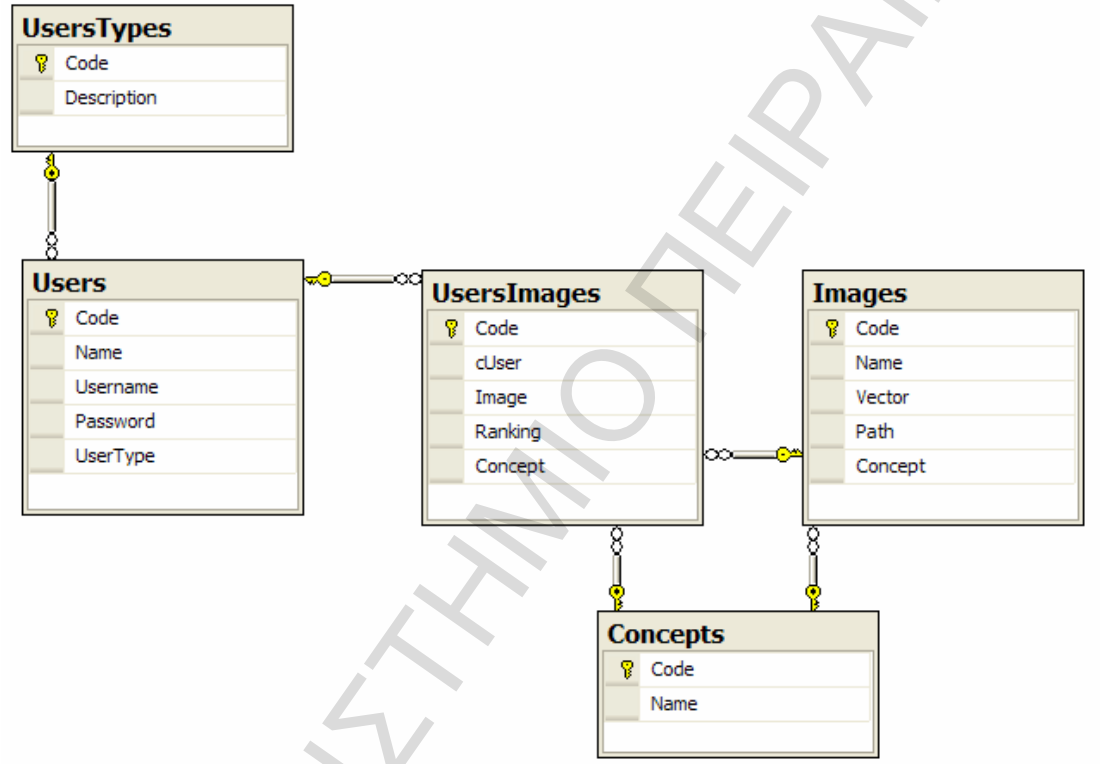
Το σύστημα αλλάζει τον τρόπο περιήγησης των εικόνων μετά την πρώτη εκπαίδευση του αλγόριθμου μηχανικής μάθησης και της εξαγωγής των πρώτων αποτελεσμάτων. Το πρόγραμμα πλέον θα εμφανίζει μόνο τις εικόνες, οι οποίες δεν έχουν βαθμολογηθεί από τον χρήστη κατά τη διάρκεια των προηγούμενων session. Οι εικόνες δεν θα εμφανίζονται σε τυχαία σειρά, όπως στο πρώτο session, αλλά με βάση τις προβλέψεις του αλγόριθμου. Δηλαδή θα εμφανίζονται οι εικόνες με τη μεγαλύτερη προς τη μικρότερη βαθμολογία που τους έχει προσδώσει το σύστημα με βάση τη συνάφειά τους ως προς το επιλεγμένο θέμα. Με αυτόν τον τρόπο θα εμφανίζονται στον χρήστη οι εικόνες, οι οποίες το πιο πιθανό να ανήκουν στην θεματική ενότητα που έχει επιλέξει.

Οι βαθμολογίες των χρηστών τις αποθηκεύουμε στον πίνακα “UsersImages” στην βάση δεδομένων. Πιο αναλυτικά ο πίνακας αποτελείται από τα εξής πεδία:

- Code, ο μοναδικός κωδικός,

- User, ο χρήστης που έβαλε την βαθμολογία,
- Image, η εικόνα,
- Ranking, η βαθμολογία και
- Concept, το γενικό θέμα με βάση το οποίο βαθμολόγησε ο χρήστης.

Η βάση δεδομένων του προγράμματός μας παρουσιάζεται στην παρακάτω εικόνα:



Εικόνα 5.5

5.3 Εκπαίδευση του αλγόριθμου

Βασικός σκοπός του προγράμματος είναι η εκπαίδευση ενός συστήματος ημι-επιτηρούμενης μάθησης, για την κατάταξη εικόνων σε υψηλού επιπέδου θέματα. Για το σύστημα μας επιλέξαμε τον αλγόριθμο Transductive Support Vector Machine (TSVM). Ο TSVM ανήκει στην κατηγορία των αλγόριθμων ημι-επιτηρούμενης μάθησης, οι οποίοι προσπαθούν να υλοποιήσουν την θεωρία διαχωρισμού σε περιοχές χαμηλής πυκνότητας, αποφεύγοντας να τοποθετήσουν τα όρια απόφασης σε περιοχές που περιέχουν unlabeled σημεία.

Η διαφορά του TSVM σε σχέση με το απλό SVM είναι ότι στον υπολογισμό του περιθωρίου περιλαμβάνει, εκτός από τα σημεία εκπαίδευσης, και τα σημεία δοκιμής. Έτσι η μεγιστοποίηση

του περιθωρίου γίνεται υπολογίζοντας την απόσταση του ορίου απόφασης σε σχέση, όχι μόνο των labeled δεδομένων του συνόλου εκπαίδευσης, αλλά και σε σχέση με τα unlabeled δεδομένα του συνόλου δοκιμής.

Οι χρήστες βαθμολογούν τις εικόνες με βάση τη σχετικότητα τους με το επιλεγμένο θέμα. Οι εικόνες που έχουν βαθμολογηθεί από τους χρήστες τις χρησιμοποιούμε σαν σύνολο εκπαίδευσης, και θα αποτελούν τα labeled δεδομένα του συστήματος μας. Σαν labels των παραδειγμάτων θέτουμε τα τις βαθμολογίες των χρηστών. Στόχος μας είναι να εκπαιδεύσουμε το σύστημα κατάλληλα ώστε να μπορεί να βαθμολογήσει πετυχημένα τις υπόλοιπες εικόνες που υπάρχουν στη βάση δεδομένων. Με αυτόν τον τρόπο να τις κατατάξει δηλαδή, στα θέματα υψηλού επιπέδου σύμφωνα με τη σχετικότητα τους ως προς τα θέματα αυτά.

Όπως έχουμε αναφέρει το TSVM περιλαμβάνει στους υπολογισμούς τους, εκτός από τα labeled και τα unlabeled δεδομένα. Έτσι θα αξιοποιήσει κατά τη διάρκεια της εκπαίδευσης και τις πληροφορίες που του παρέχουν οι εικόνες που δεν έχουν βαθμολογηθεί από τους χρήστες, δηλαδή τα δεδομένα αυτά τα οποία δεν έχουν label και τα οποία θα αποτελέσουν το σύνολο δοκιμής του αλγορίθμου. Αυτό επιτρέπει στον αλγόριθμο μάθησης να εκμεταλλευθεί πλήρως όλα τα δεδομένα που μπορεί από τις εικόνες που βρίσκονται στη βάση δεδομένων. Ακόμα πρέπει να σημειώσουμε ότι δεν μας ενδιαφέρει απαραιτήτως να βρούμε ένα γενικό κανόνα για την κατάταξη εικόνων σε γενικά θέματα, αλλά πρωτίστως βασικός μας σκοπός είναι η επιτυχημένη βαθμολόγηση των εικόνων που αποτελούν το δείγμα δοκιμής μας.

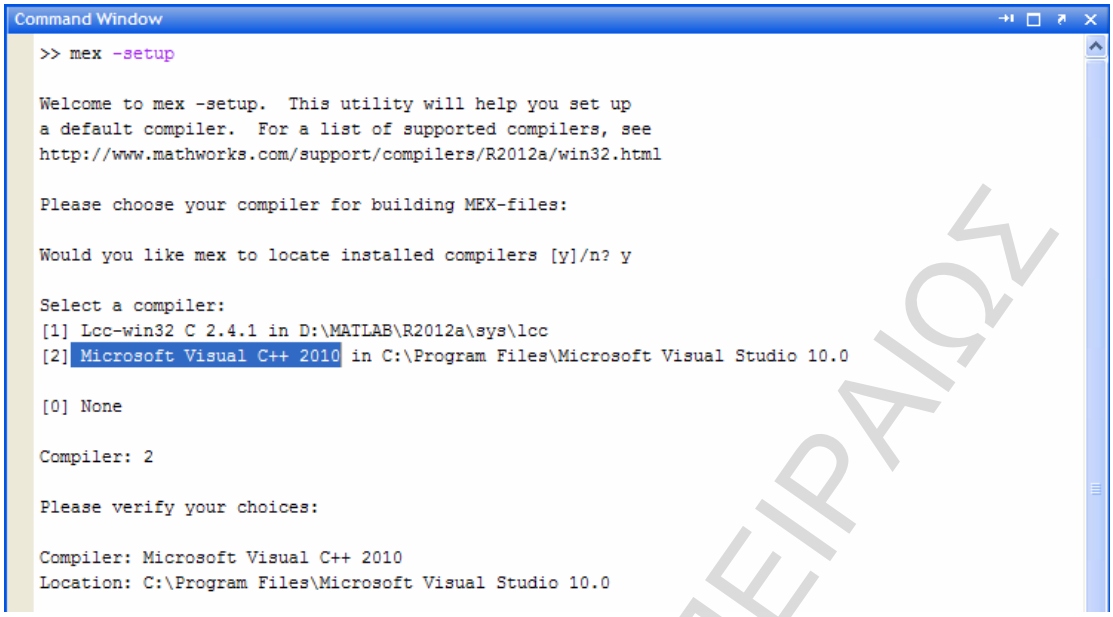
5.4 Υλοποίηση Αλγόριθμου

Στο πρόγραμμα μας για την υλοποίηση του TSVM αλγόριθμου, επιλέξαμε να χρησιμοποιήσουμε τον αλγόριθμο SVM-light. Ο SVM-light είναι ένας βελτιστοποιημένος αλγόριθμος, ο οποίος αποτελεί μια πρόταση για το πρόβλημα της βελτιστοποίησης του TSVM. Ο αλγόριθμος SVM-light εκτελεί ενός είδους τοπική αναζήτηση (local search) στις συντεταγμένες των σημείων, ξεκινώντας από τις αρχικές τιμές των σημείων.

Από θέμα αποτελέσματος, ο αλγόριθμος που υλοποιείται στον TSVM-light δεν παράγει σίγουρα μια καθολική βελτιστοποιημένη λύση, αλλά μπορεί να διαχειριστεί με επιτυχία σύνολα δοκιμών μεγέθους μέχρι 100.000 παραδείγματα. Όπως είναι κατανοητό, για το σκοπό και το μέγεθος του προγράμματος μας, τα μεγέθη αυτά μας καλύπτουν πλήρως.

Για την υλοποίηση του αλγόριθμου στο πρόγραμμά μας, επιλέξαμε την έκδοση SVM-light 6.01, το οποίο είναι επέκταση του πακέτου SVM-light και το οποίο είναι υλοποιημένο σε περιβάλλον Matlab. Για την λειτουργία του προγράμματος θα πρέπει να γίνει αρχικά εγκατάσταση των κωδικών της βιβλιοθήκης στο περιβάλλον του Matlab. Τα βήματα που πρέπει να ακολουθήσουμε είναι τα εξής:

- Θα πρέπει μέσα από το command window του Matlab να δώσουμε την εντολή “mex –setup”. Στη συνέχεια το Matlab θα μας προβάλλει τους διαθέσιμους compiler, και υπό την προϋπόθεση ότι έχουμε εγκαταστημένο το περιβάλλον του Visual Studio, επιλέγουμε τον compiler “Microsoft Visual C++ 2010”.
- Στη συνέχεια τοποθετούμε στο directory του Matlab τον κώδικα της βιβλιοθήκης του SVM-light 6.01 και μετά προσθέτουμε το path της βιβλιοθήκης με την εντολή “addpath”, π.χ. addpath('c:/matlab/svm_mex601’).
- Στο τέλος για να κάνουμε compile τον κώδικα εκτελούμε τη συνάρτηση “compilemex.m”, που βρίσκεται στο φάκελο του svm_mex601.



Εικόνα 5.6

Για να μπορούμε να χρησιμοποιήσουμε το περιβάλλον του Matlab από το πρόγραμμα μας, θα πρέπει να μπορούμε να ενσωματώσουμε κώδικα Matlab μέσα σε C# κώδικα. Για να το πετύχουμε αυτό χρησιμοποιούμε τη μέθοδο, όπου χρησιμοποιούμε το Matlab σαν αυτοματοποιημένο server μέσα από τον κώδικα C#, χρησιμοποιούμε το engine interface μέσω της COM αυτοματοποίησης. Αυτή η μέθοδος μας δίνει τη δυνατότητα να έχουμε debug από το C# πρόγραμμά μας.

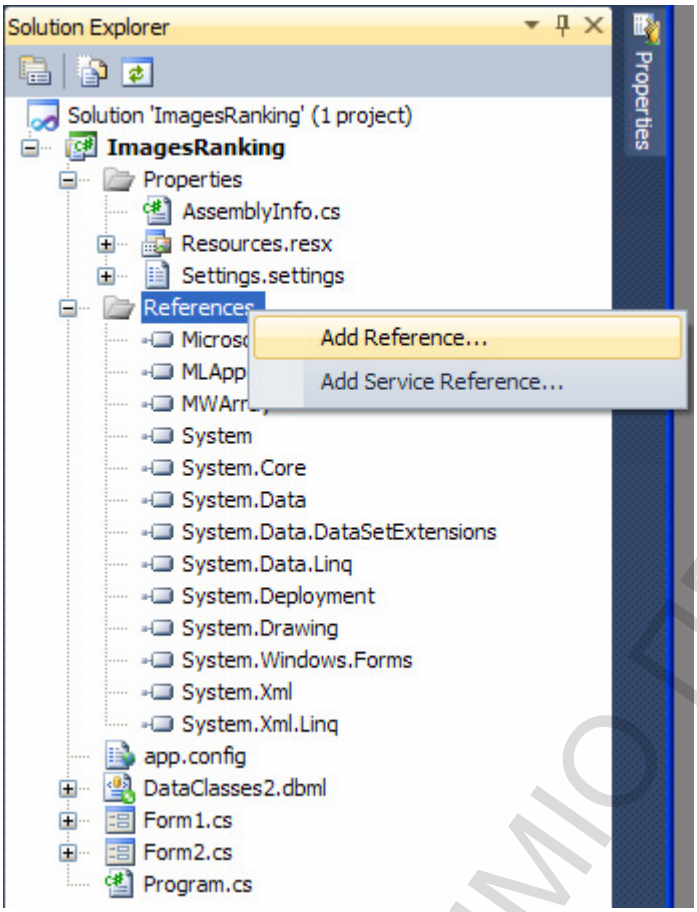
Η διαδικασία που ακολουθήσαμε είναι η εξής:

Στο Visual Studio προσθέσαμε τις παρακάτω αναφορές στο πρόγραμμα μας:

α. Την αναφορά στο Matlab Application Type Library (MLApp), το οποίο είναι COM αντικείμενο και αποτελεί μέρος του πυρήνα του Matlab. Το path της αναφοράς είναι: *matlabroot*/bin/win32/mlapp.tlb.

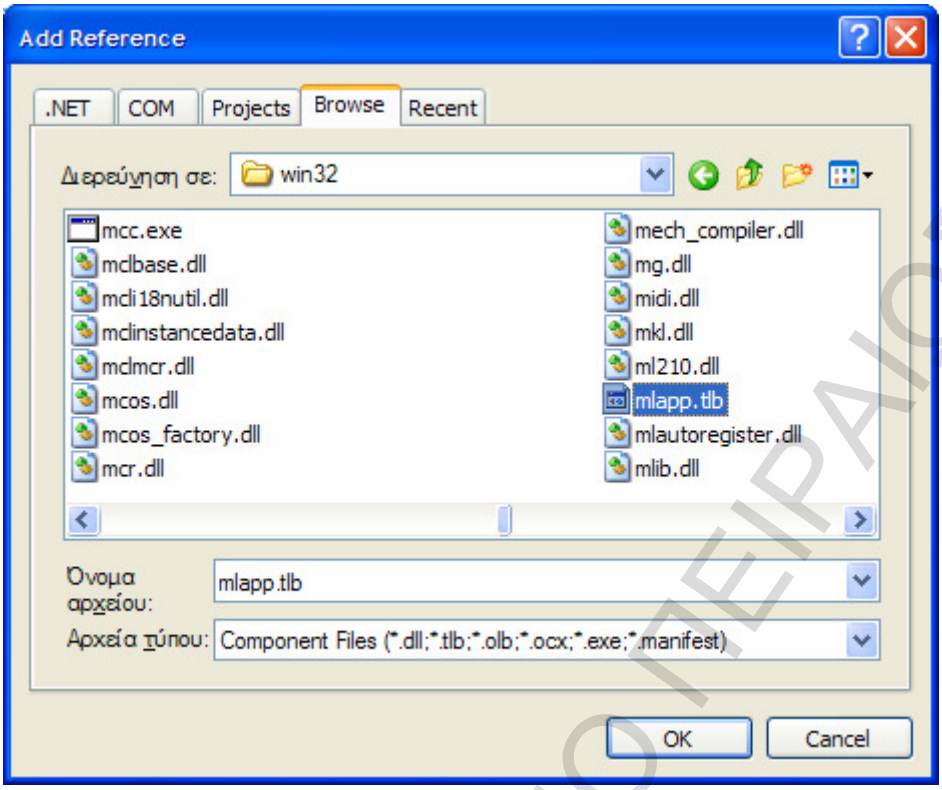
β. Την αναφορά στο MathWorks, .NET MWArray API (MWArray), το οποίο είναι .NET assembly και είναι μέρος του Builder για το .NET. Το path της αναφοράς είναι: *matlabroot*/bin/win32/MWArray.dll.

Στο Visual Studio για να προσθέσουμε αναφορές στο πρόγραμμα μας, επιλέγουμε τον φάκελο “References” που βρίσκεται στον κατάλογο στο πρόγραμμα μας, και επιλέγουμε την εντολή “Add Reference”.



Εικόνα 5.7

Στη συνέχεια επιλέγουμε την “Browse” καρτέλα και επιλέγουμε τα αρχεία που αναφέραμε προηγουμένως στις αντίστοιχες διευθύνσεις τους:



Εικόνα 5.8

5.5 Εκτέλεση Αλγόριθμου

Οι χρήστες βαθμολογώντας τις εικόνες, όπως έχουμε ήδη πει, συνεισφέρουν στο σύστημα τιμές για τα δεδομένα μας. Τα δεδομένα αυτά τα αποθηκεύουμε στη βάση δεδομένων μας για να μπορούμε να τα αξιοποιήσουμε αργότερα σαν το σύνολο εκπαίδευσης του TSVM αλγόριθμου. Έτσι όταν οι χρήστες βαθμολογήσουν τόσες εικόνες, ώστε να θεωρήσουμε ότι έχουμε ικανοποιητικό δείγμα καλούμε τον αλγόριθμο TSVM για την εξαγωγή των αποτελεσμάτων.

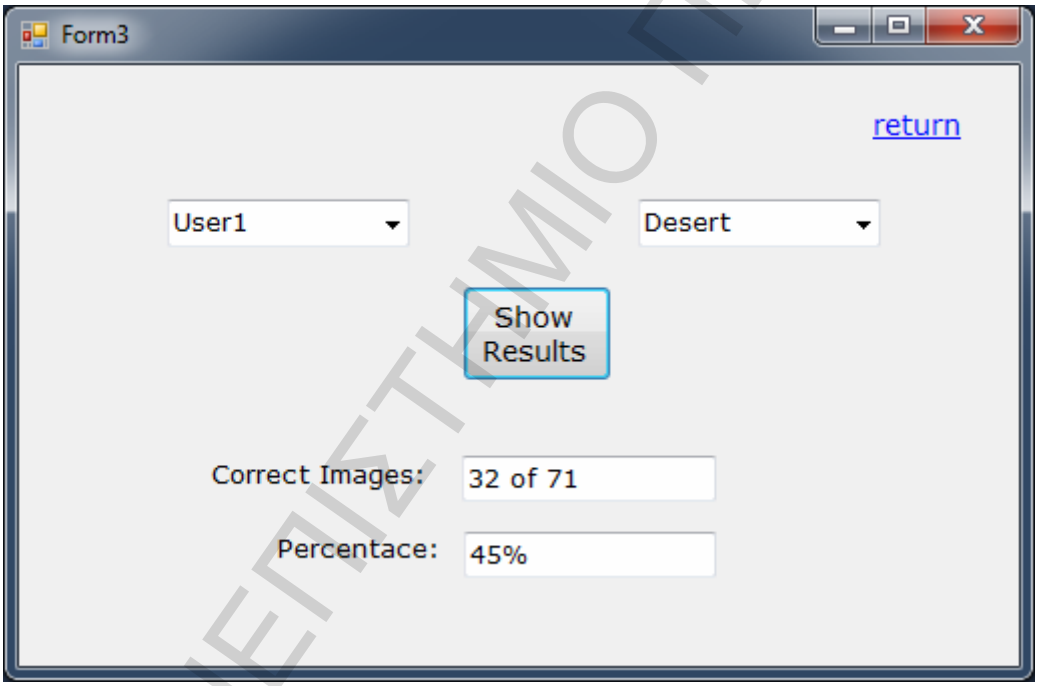
Το δικαίωμα για την εκπαίδευση του αλγόριθμου και της εξαγωγής των αποτελεσμάτων του προγράμματος μας, το έχουμε αναθέσει στο χρήστη administrator, ο οποίος θα μπορεί να διαχειρίζεται τα συγκεντρωτικά αποτελέσματα των υπόλοιπων χρηστών. Το σύνολο εκπαίδευσης του αλγόριθμου είναι το σύνολο των εικόνων, μαζί με τις βαθμολογίες όσων εικόνων έχουν βαθμολογηθεί από τους χρήστες, ενώ το σύνολο δοκιμής αποτελείται μόνο από τις εικόνες που δεν έχουν βαθμολογηθεί.

Για να εκπαιδεύσουμε τον αλγόριθμο TSVM με τα δεδομένα εκπαίδευσης, καλούμε τον αλγόριθμο SVM-light. Για να το πετύχουμε αυτό θα πρέπει ήδη να έχουμε κάνει σύνδεση του κώδικα μας με το Matlab, ώστε να μπορούμε να χρησιμοποιήσουμε τις εντολές της βιβλιοθήκης του αλγόριθμου. Εκτελούμε την εντολή ‘svmlearn’ με τα δεδομένα εκπαίδευσης και τις παραμέτρους που επιθυμούμε. Τα δεδομένα εκπαίδευσης θα πρέπει να είναι σε μορφή δύο πινάκων. Ο ένας πίνακας θα περιέχει τα χαρακτηριστικά των εικόνων τα οποία αποτελούν ένα διάνυσμα για κάθε πίνακα, ενώ ο άλλος πίνακας θα περιέχει τις βαθμολογίες των εικόνων, δηλαδή τις τιμές.

Η εντολή ‘svmlearn’ μας επιστρέφει την Matlab μεταβλητή ‘model’, η οποία περιέχει τα διανύσματα υποστήριξης και τις τιμές, οι οποίες υπολογίστηκαν κατά τις διαδικασίες βελτιστοποίησης. Χρησιμοποιώντας τη μεταβλητή αυτή και τα σύνολο δοκιμής, μπορούμε πάρουμε σαν αποτέλεσμα το ποσοστό λάθους και τις τιμές πρόβλεψης. Όπως το σύνολο εκπαίδευσης, έτσι το σύνολο δοκιμής θα πρέπει να αποτελείται από δυο πίνακες. Ο πρώτος πίνακας θα περιέχει το σύνολο των χαρακτηριστικών των εικόνων που δεν έχουν βαθμολογηθεί, ενώ ο δεύτερος θα περιέχει τυχαίες τιμές για τις βαθμολογίες των εικόνων. Η συνάρτηση που χρησιμοποιούμε για να εφαρμόσουμε το μοντέλο στο σύνολο δοκιμής είναι η ‘svmclassify’, και παίρνει σαν ορίσματα τους δυο πίνακες του συνόλου δοκιμής και την μεταβλητή ‘Model’.

5.6 Εμφάνιση Αποτελεσμάτων

Ο χρήστης administrator έχει τη πρόσβαση στη φόρμα όπου θα εμφανίζονται τα αποτελέσματα του αλγόριθμου μηχανικής μάθησης. Στη φόρμα αυτή ο χρήστης έχει τη δυνατότητα να επιλέξει το session για το οποίο θέλει να δει τα αποτελέσματα. Μπορεί μέσω λιστών να επιλέξει τις βαθμολογίες που έχει δώσει ένας συγκεκριμένος χρήστης και για τη θεματική ενότητα σύμφωνα με την οποία έχει βαθμολογήσει. Η διάταξη της φόρμας παρουσιάζεται στην παρακάτω εικόνα.



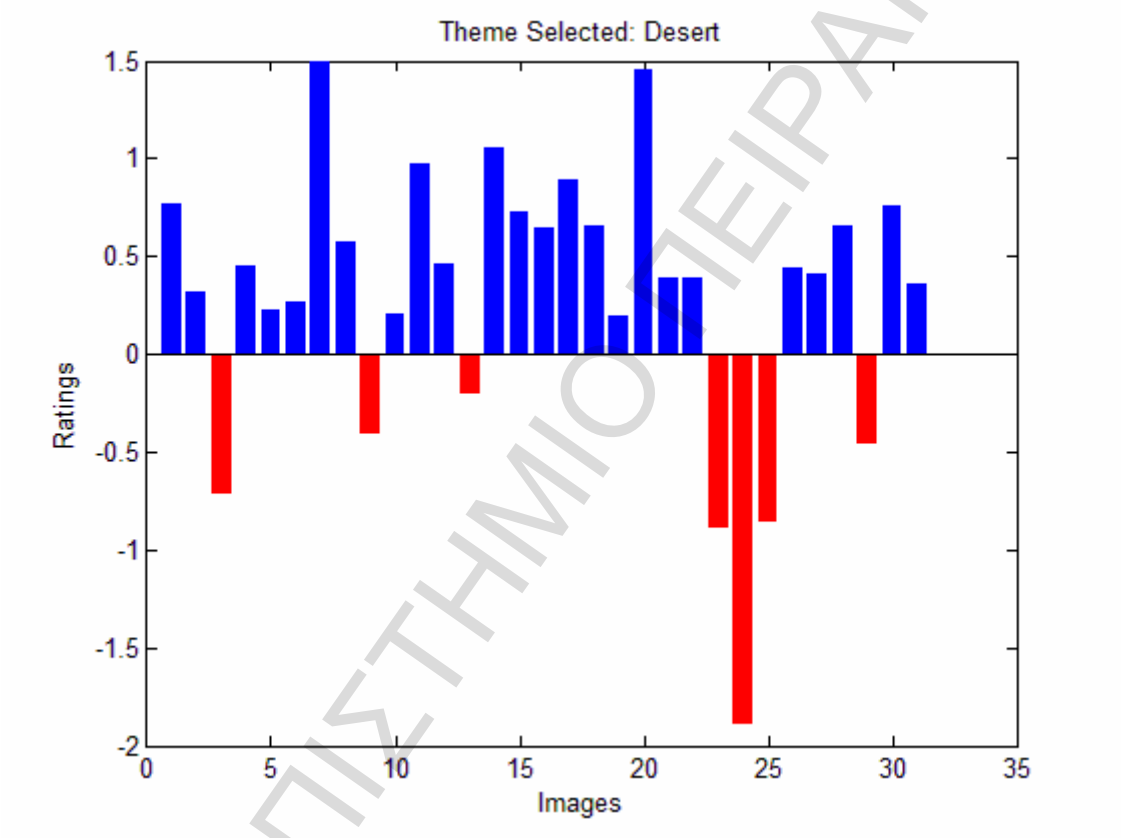
Εικόνα 5.9

Στη συγκεκριμένη περίπτωση ο administrator έχει επιλέξει να επεξεργαστεί τις βαθμολογίες του χρήστη “User1” για τη θεματική ενότητα “Desert”. Με το κουμπί “Show Results” γίνεται κλήση του αλγόριθμου μηχανικής μάθησης, το οποίο επεξεργάζεται τα σύνολο εκπαίδευσης του και επιστρέφει τις προβλέψεις για το σύνολο δοκιμής.

Το πρόγραμμα επιστρέφει πόσες εικόνες κατάφερε να ανακτήσει στο σύνολο των εικόνων της αναζητούμενης κλάσης, καθώς και το ποσοστό επιτυχίας. Σαν σύνολο εικόνων της επιλεγμένης κλάσης θεωρούμε τις εικόνες που ανήκουν στη συγκεκριμένη κλάση και δεν έχουν

βαθμολογηθεί από τον χρήστη κατά τη διάρκεια των προηγουμένων session. Στην εικόνα 5.9 βλέπουμε ότι το πρόγραμμα έχει καταφέρει να ανακτήσει 32 από τις 71 εικόνες, που ανήκουν στην αναζητούμενη θεματική ενότητα. Ακόμα το ποσοστό επιτυχίας είναι 45%.

Τα αποτελέσματα εμφανίζονται με τη μορφή γραφικών παραστάσεων, όπου απεικονίζονται οι προβλέψεις του αλγόριθμου μηχανικής μάθησης για τις εικόνες που ανήκουν στην αναζητούμενη θεματική ενότητα. Οι εικόνες στη γραφική παράσταση θα έχουν διαφορετικό χρώμα, ανάλογα με την επιτυχημένη ανάκτηση τους, ώστε να έχουμε καλύτερη απεικόνιση των αποτελεσμάτων του αλγόριθμου. Οι εικόνες όπου ανακτήσει σωστά ο αλγόριθμος, απεικονίζονται με μπλε χρώμα, ενώ οι εικόνες που δεν έχει ανακτήσει θα εμφανίζονται με κόκκινο.



Εικόνα 5.10

Πειραματική Ανάλυση

Αντικείμενο της πτυχιακής εργασίας είναι η ανάπτυξη ενός συστήματος για την ανάκτηση εικόνων με βάση το θεματικό περιεχόμενο τους. Τα πιο σημαντικά κομμάτια του συστήματος για

την αξιολόγηση των εικόνων, είναι η αντιπροσώπευση τους από τα αντίστοιχα διανύσματα χαρακτηριστικών, καθώς και η τεχνική σύστασης που θα χρησιμοποιηθεί.

Για την εξαγωγή των χαρακτηριστικών των εικόνων χρησιμοποιήσαμε τον αλγόριθμο **Visual Descriptor Extraction** (**VDE**), ο οποίος εξάγει τους οπτικούς περιγραφείς MPEG-7. Το σύνολο των χαρακτηριστικών αυτών το τροποποιήσαμε σε μορφή διανυσμάτων για να είναι δυνατή η μετέπειτα επεξεργασία τους. Αποφασίστηκε από τις εικόνες να γίνει εξαγωγή της καθολικής περιγραφής τους, ώστε να γίνει αξιολόγηση του γενικού θέματος τους.

6.1 Επιλογή Εικόνων

Το αρχικό βήμα για τη διεξαγωγή του πειράματος μας, είναι η επιλογή των εικόνων. Οι εικόνες θα αποτελέσουν το σύνολο δεδομένων, πάνω στο οποίο θα εφαρμόσουμε το πρόγραμμα μας και θα εξάγουμε τα πειραματικά αποτελέσματα. Το εργαλείο θα επεξεργαστεί κατάλληλα τα χαρακτηριστικά των εικόνων αυτών, και σε συνάρτηση με τη βαθμολόγηση των χρηστών θα εξάγει τα αντίστοιχα αποτελέσματα.

Για το πείραμα μας αποφασίσαμε το σύνολο των εικόνων να αποτελείται από εικόνες γενικού θέματος. Κάναμε αυτή την επιλογή κυρίως, γιατί έχουμε επιλέξει να εξάγουμε τα χαρακτηριστικά της καθολικής περιγραφής των εικόνων. Κατά δεύτερον προτιμήσαμε τις εικόνες γενικού θέματος, γιατί αυτές περιέχουν πιο καθαρά και διακριτά μεταξύ τους θέματα. Αυτό το γεγονός διευκολύνει τους χρήστες όσον αφορά την βαθμολόγηση των εικόνων. Ακόμα με την επιλογή εικόνων γενικού θέματος, προμηθεύουμε τον αλγόριθμο μηχανικής μάθησης με δεδομένα, τα οποία θα τον βοηθήσουν να εξάγει καλύτερα πειραματικά αποτελέσματα και να καταστήσει την βαθμολόγηση των εικόνων πιο εύκολη. Από αυτά τα πειραματικά αποτελέσματα μπορούμε να βγάλουμε πιο ξεκάθαρα συμπεράσματα κατά την πειραματική ανάλυση, αφού οι τιμές των αποτελεσμάτων θα είναι πιο διακριτές και θα έχουν μεγαλύτερες διαφορές μεταξύ τους.

Για τη διεξαγωγή του πειράματος επιλέξαμε έξι κατηγορίες θεμάτων. Προτιμήθηκαν θέματα τα οποία είναι άμεσα αναγνωρίσιμα και καθαρά διακριτά μεταξύ τους, για τους λόγους που αναφέραμε προηγουμένως. Έτσι, ανάμεσα στα άλλα θέματα, έχουμε επιλέξει σαν θέματα των εικόνων τον ωκεανό και την έρημο. Αυτά τα θέματα έχουν χαρακτηριστικά τα οποία μπορούμε εύκολα να διαχωρίσουμε σαν χρήστες, αλλά και διευκολύνουν τον αλγόριθμο να εξάγει πιο σωστά αποτελέσματα. Για παράδειγμα οι εικόνες που έχουν σαν θέμα τον ωκεανό, έχουν σαν κύριο χαρακτηριστικό ότι κυριαρχεί σχεδόν στο σύνολο της εικόνας το χρώμα μπλε, ενώ στις εικόνες που έχουν θέμα την έρημο κυριαρχεί το κίτρινο χρώμα.

Ακόμα οι εικόνες που επιλέξαμε, και ανήκουν στο ίδιο θέμα, προσπαθήσαμε να έχουν όσο τον δυνατόν μεγαλύτερη συνάφεια μεταξύ τους. Για παράδειγμα ένα θέμα που επιλέξαμε είναι τα ανθρώπινα πρόσωπα. Σε αυτήν την κατηγορία επιλέξαμε εικόνες στις οποίες τα πρόσωπα καλύπτουν όλο το κάδρο, ενώ προσπαθήσαμε στις εικόνες αυτές να φαίνεται κυρίως η μπροστινή όψη των προσώπων και όχι η πλαϊνή. Με αυτόν τον τρόπο οι εικόνες που ανήκουν στο ίδιο θέμα μοιράζονται πολλά κοινά χαρακτηριστικά και θα αποτελούν πιο σωστά δεδομένα επεξεργασίας για τον αλγόριθμο μηχανικής μάθησης.

Μια άλλη παράμετρος, με την οποία επιλέξαμε τις φωτογραφίες, είναι αυτές να περιέχουν την πλειονότητα των στοιχείων που χαρακτηρίζουν το αντίστοιχο θέμα. Στο βαθμό φυσικά που κάτι τέτοιο είναι εφικτό. Γενικά προσπαθήσαμε οι εικόνες να περιέχουν στο μεγαλύτερο μέρος της φωτογραφίας και ευκρινώς το αντίστοιχο θέμα. Για παράδειγμα στις εικόνες με θεματική ενότητα τα ανθρώπινα πρόσωπα, επιλέξαμε παραδείγματα στα οποία το πρόσωπο διακρίνεται καθαρά και χωρίς οπτικά εμπόδια. Δηλαδή δεν συμπεριλάβαμε εικόνες στις οποίες η περιοχή των ματιών κρύβεται από μαύρα γυαλιά ή δεν είναι εμφανές το κάτω μέρος του προσώπου. Ο λόγος είναι ότι τέτοια χαρακτηριστικά είναι κρίσιμα για την ταξινόμηση των εικόνων στα θέματα τους. Ο αλγόριθμος θα “αναζητεί” τέτοια κρίσιμα χαρακτηριστικά για να αναγνωρίσει τις εικόνες με βάση το θέμα τους. Παρέχοντας στο σύστημα ορθά δεδομένα, αυξάνουμε τις πιθανότητες τα πειραματικά αποτελέσματα να έχουν μικρότερο σφάλμα.

Για λόγους συνέπειας και καλύτερης παρουσίασης επιλέξαμε εικόνες με την ίδια ανάλυση και της ίδιας συμπίεσης. Όλες οι εικόνες, τις οποίες θα βαθμολογούν οι χρήστες, έχουν διαστάσεις 800x600 pixel. Ακόμα για τεχνικούς λόγους, όλες οι εικόνες είναι της μορφής JPEG. Με αυτή την μέθοδο συμπίεσης καταφέρνουμε να έχουμε μεγάλο πλήθος εικόνων με σχετικά μικρό μέγεθος στον χώρο μνήμης. Είναι δεδομένο ότι είναι προτιμητέο το πλήθος των εικόνων να είναι αρκετά μεγάλο για την καλύτερη λειτουργία του αλγόριθμου μηχανικής μάθησης. Ακόμα προτιμούμε τις εικόνες με περιορισμένο μέγεθος για να είναι πιο ομαλή η περιήγηση και η βαθμολόγηση από τους χρήστες, και να μην παρατηρούνται καθυστερήσεις στο φόρτωμα των εικόνων.

6.2 Δεδομένα Εισόδου

Το πρόγραμμα μας, όπως έχουμε αναφέρει, δίνει τη δυνατότητα στο χρήστη administrator να επεξεργαστεί τις βαθμολογίες που έχουν καταχωρήσει οι υπόλοιποι χρήστες. Ο administrator μπορεί να πλοηγηθεί στη φόρμα του προγράμματος, όπου εμφανίζονται τα αποτελέσματα, και να επιλέξει τις τιμές που έχει δώσει ο κάθε χρήστης στις εικόνες, με βάση κάθε φορά τη θεματική ενότητα που έχει επιλέξει. Όπως είναι φυσικό για κάθε χρήστη εμφανίζονται μόνο τα θέματα σύμφωνα με τα οποία έχει δώσει βαθμολογίες.

Όταν ο administrator επιλέξει το χρήστη και τη θεματική ενότητα, σύμφωνα με την οποία έχει βαθμολογήσει τις εικόνες, μπορεί να δει τα αποτελέσματα του συστήματος πατώντας το αντίστοιχο κουμπί. Το σύστημα τότε επεξεργάζεται αυτά τα δεδομένα εισόδου και επιστρέφει στο χρήστη τους υπολογισμούς του αλγόριθμου μηχανικής μάθησης για τον συγκεκριμένο χρήστη και τη συγκεκριμένη θεματική ενότητα.

Ο αλγόριθμος μηχανικής μάθησης που χρησιμοποιούμε για το πρόγραμμα μας είναι ο Transductive Support Vector Machine (TSVM). Τα δεδομένα που προμηθεύουμε το σύστημα είναι δύο ζευγάρια. Το πρώτο είναι ο πίνακας εκπαίδευσης και το διάνυσμα εκπαίδευσης. Ο πίνακας περιέχει τα διανύσματα χαρακτηριστικών των εικόνων, ενώ το διάνυσμα τις αντίστοιχες βαθμολογίες τους, δηλαδή τα label τους. Το άλλο ζευγάρι είναι ο πίνακας και το διάνυσμα δοκιμής.

Κατά τη διαδικασία της βαθμολόγησης οι χρήστες βαθμολογούν τις εικόνες σε κλίμακα από 1 εώς 5. Τις τιμές αυτές τις μετατρέπουμε κατάλληλα για το διάνυσμα εκπαίδευσης, στο οποίο οι τιμές κυμαίνονται από το -2 μέχρι το +2. Αποφασίσαμε αυτήν την μετατροπή για να είναι τα πειραματικά αποτελέσματα πιο ευδιάκριτα για μας. Έτσι, για τις εικόνες που δεν έχουν σχέση με τη θεματική ενότητα, θα έχουν αρνητικές τιμές, ενώ αντίστοιχα οι εικόνες που ανήκουν στη θεματική ενότητα θα έχουν θετικό πρόσημο οι τιμές τους. Με αυτόν τον τρόπο όταν θα επεξεργαζόμαστε συνολικά τα αποτελέσματα θα έχουμε μια πιο ευκρινή εικόνα για το σύνολο των αποτελεσμάτων.

Ο αλγόριθμος TSVM είναι transductive, γεγονός που σημαίνει ότι για την κατάταξη των εικόνων από το σύστημα, θα πρέπει ο πίνακας εκπαίδευσης να περιέχει τα διανύσματα χαρακτηριστικών όλων των εικόνων, και όχι μόνο αυτών που έχουν βαθμολογηθεί, όπως συμβαίνει στον αλγόριθμο inductive SVM. Το διάνυσμα εκπαίδευσης θα περιέχει, αντίστοιχα, τιμές για όλες τις εικόνες. Αφού οι τιμές στο διάνυσμα εκπαίδευσης κυμαίνονται από το -2 μέχρι το 2, αποφασίσαμε να προσδώσουμε στις εικόνες που δεν έχουν βαθμολογηθεί, την τιμή μηδέν, δηλαδή το μέσο της κλίμακας βαθμολόγησης. Ο πίνακας για την δοκιμή του αλγόριθμου μηχανικής μάθησης θα περιέχει τα διανύσματα χαρακτηριστικών για όλες τις εικόνες, που δεν έχουν βαθμολογηθεί. Το αντίστοιχο διάνυσμα τιμών αποφασίσαμε να έχει τιμές μηδέν.

6.3 Εκτέλεση Αλγορίθμου

Ο αλγόριθμος αποτελείται από δύο μέρη. Τη διαδικασία μάθησης (svmlearn) και τη διαδικασία κατάταξης (svmclassify). Η διαδικασία ταξινόμησης μπορεί να χρησιμοποιηθεί για να εφαρμοστεί το μοντέλο μάθησης σε νέα παραδείγματα. Στο πρόγραμμα μας εφαρμόσαμε τον

αλγόριθμο στη διαδικασία μάθησης με την παράμετρο Gamma στο 0.3 και την παράμετρο ορίου λάθους στο 0.5. Η εντολή θα έχει την μορφή:

```
model = svmlearn(X, Y,' -g 0.3 -c 0.5');
```

όπου X,Y ο πίνακας και το διάνυσμα εκπαίδευσης αντίστοιχα.

Αυτή η εντολή δημιουργεί μια νέα μεταβλητή Matlab. Αυτή η δομή περιέχει όλα τα διανύσματα υποστήριξης, καθώς και όλες τις τιμές που έχουν υπολογιστεί από τις διαδικασίες βελτιστοποίησης.

Στη συνέχεια το πρόγραμμα καλεί την εντολη 'svmclassify' για τη διαδικασία ταξινόμησης. Η εντολή αυτή παίρνει ορίσματα τον πίνακα και το διάνυσμα δοκιμής, καθώς και τη μεταβλητή model, που παράχθηκε από την προηγούμενη εντολή svmlearn. Η εντολή έχει τη μορφή:

```
[err, predictions] = svmclassify(Xt,Yt,model);
```

όπου Xt, Yt ο πίνακας και το διάνυσμα δοκιμής αντίστοιχα.

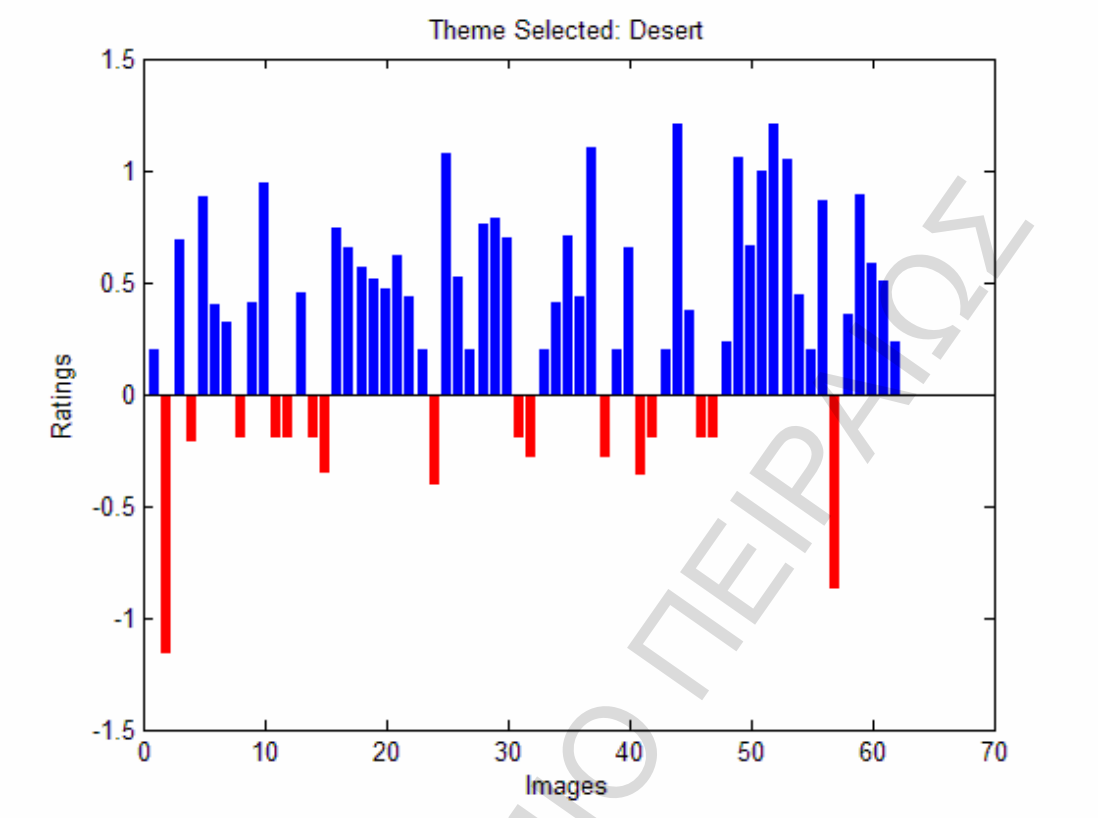
Η εντολή 'svmclassify' επιστρέφει δύο αποτελέσματα. Επιστρέφει το διάνυσμα με τα αποτελέσματα των υπολογισμών για τον πίνακα δοκιμής, ο οποίος περιέχει τα διανύσματα των χαρακτηριστικών για τις εικόνες που θέλουμε να βαθμολογήσει το σύστημα, και το μέσο σφάλμα των αποτελεσμάτων αυτών σε σχέση με το διάνυσμα δοκιμής, το οποίο περιέχει τιμές που έχουμε σε αυτές τις εικόνες.

6.4 Πειραματικά Αποτελέσματα

Καθώς το δείγμα των εικόνων, που χρησιμοποιούμε για τη διαδικασία της εκπαίδευσης και τη διαδικασία ανάκτησης, είναι αρκετά μεγάλο, για την καλύτερη ανάλυση των αποτελεσμάτων αναπαριστούμε τα αποτελέσματα, που λάβαμε από τον αλγόριθμο TSVM, σε γραφικές παραστάσεις. Ο αριθμός των εικόνων είναι τέτοιος, ώστε θα ήταν δύσκολο να μπορέσουμε να μελετήσουμε τα αποτελέσματα χωρίς τη βοήθεια γραφημάτων.

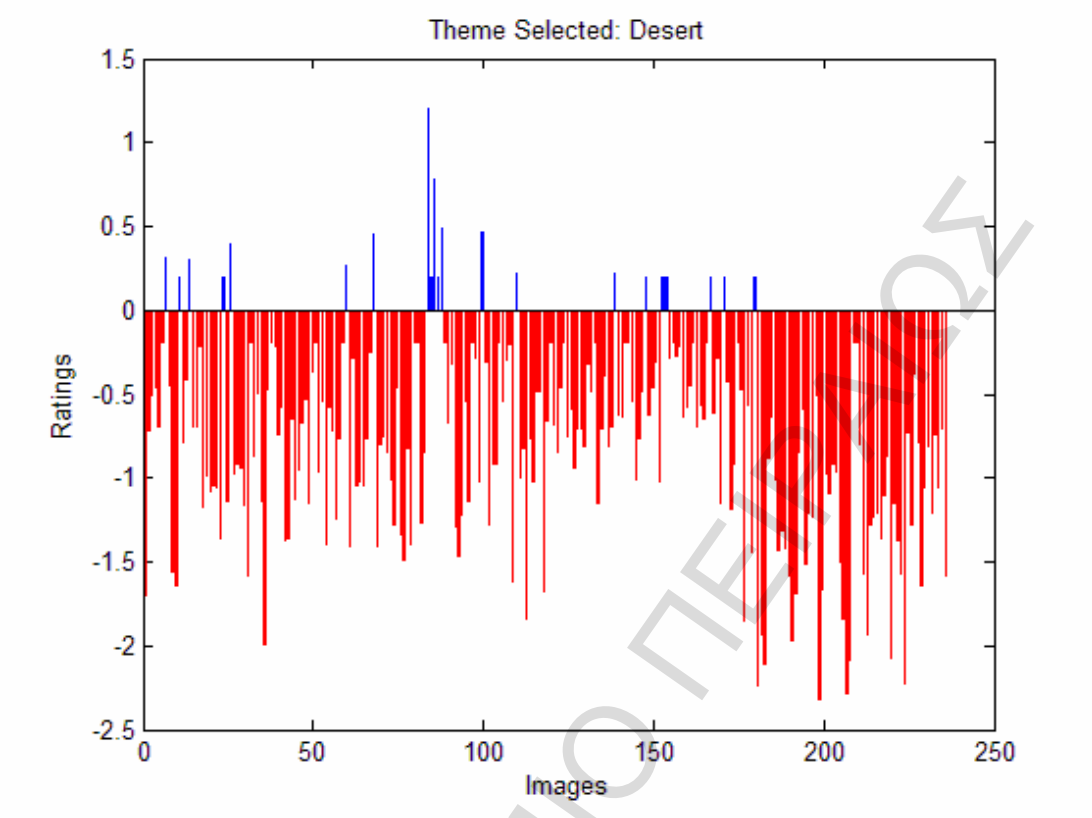
Τα αποτελέσματα που λάβαμε κατά την πειραματική δοκιμή του προγράμματος μας κρίνονται θετικά. Στην πλειονότητα τους οι προβλέψεις του προγράμματος μας για τη θεματική ενότητα στην οποία ανήκουν οι εικόνες που επεξεργάζεται, παρατηρούμε ότι είναι στη σωστή κατεύθυνση. Έτσι για τις περισσότερες εικόνες, οι οποίες ανήκουν στη θεματική ενότητα για την οποία βαθμολόγησε ο χρήστης, το πρόγραμμα τις έχει ανακτήσει σωστά. Αντίστοιχα για τις εικόνες, οι οποίες δεν ανήκουν σε αυτήν την ενότητα, το πρόγραμμα τους έχει προσδώσει αρνητικές τιμές.

Στην παρακάτω εικόνα απεικονίζονται οι τιμές των προβλέψεων του αλγόριθμου, για το θέμα “Desert”.



Εικόνα 6.1

Στην παρακάτω εικόνα για το ίδιο θέμα, προβάλλονται οι προβλέψεις για τις εικόνες, οι οποίες δεν ανήκουν σε αυτό το θέμα:



Εικόνα 6.2

Από τις δύο αυτές γραφικές παραστάσεις παρατηρούμε ότι το πρόγραμμα μας, πράγματι κάνει προβλέψεις, που στο σύνολο τους, είναι σωστές. Τα σφάλματα στη διαδικασία ανάκτησης είναι σχετικά λίγα σε αριθμό, σε σχέση με το σύνολο των εικόνων.

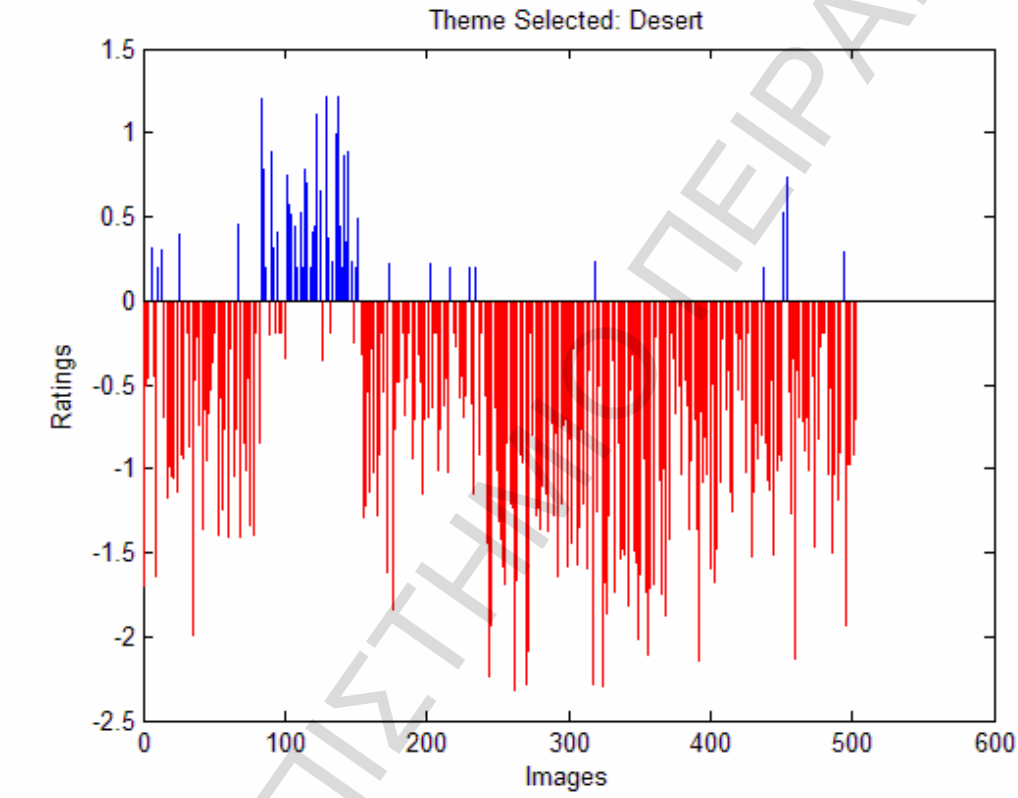
Όπως παρατηρούμε και από τις γραφικές παραστάσεις, ενώ ο αλγόριθμος, σε γενικές γραμμές, έχει κατατάξει τις εικόνες στις σωστές κλάσεις, οι βαθμολογίες που έχει δώσει ποικίλουν στις τιμές τους. Αυτή τη διακύμανση στη βαθμολογία είναι κάτι που πρέπει να τη θεωρούμε αναμενόμενη. Αυτό που θεωρείται το πιο σημαντικό είναι το γεγονός ότι η πλειονότητα των εικόνων που ανήκουν στην αναζητούμενη θεματική ενότητα, έχουν ανακτηθεί από το σύστημα, πράγμα το οποίο αποτελούσε και το σκοπό του προγράμματός μας.

Μια πολύ σημαντική παρατήρηση για τα αποτελέσματα του προγράμματος, είναι ότι οι βαθμολογίες είναι πιο συνεπείς για τις εικόνες, για τις οποίες δεν ανήκουν στη θεματική ενότητα που βαθμολογείται. Σε αυτές τις εικόνες το πρόγραμμα έχει δώσει, σχεδόν, στο σύνολο τους αρνητικές τιμές. Φυσικά υπάρχουν και θετικές τιμές για αυτές τις εικόνες, αλλά σε σχέση με το σύνολο των δειγμάτων είναι ελάχιστες, και ανήκουν στα όρια του στατιστικού σφάλματος. Αυτές οι διαφοροποιήσεις μπορούν να οφείλονται και στις εικόνες του δείγματος, οι οποίες ενδέχεται να μην απεικονίζουν το θέμα καθαρά και με συνέπεια. Τέλος μπορούν να οφείλονται και σε λάθη των βαθμολογιών των δειγμάτων εκπαίδευσης, δηλαδή οι χρήστες να έχουν δώσει βαθμολογίες θετικές σε εικόνες, που δεν ανήκουν σε αυτό το θέμα.

Αντίθετα παρατηρούμε ότι το πρόγραμμα έχει μεγαλύτερη δυσκολία στη αναγνώριση των εικόνων που ανήκουν στη θεματική ενότητα που έχουμε επιλέξει. Όπως στην προηγούμενη περίπτωση, έτσι και σε αυτή, η πλειονότητα των προβλέψεων κινείται σε σωστά πλαίσια, αλλά είναι μεγαλύτερος ο αριθμός των δειγμάτων, ο οποίος έχει πάρει αρνητική βαθμολογία. Τα αποτελέσματα αυτά μπορούμε να πούμε ότι ήταν αναμενόμενα, εξαιτίας του γεγονότος ότι και εμπειρικά για τον άνθρωπο είναι πιο εύκολο να ξεχωρίσει εικόνες, οι οποίες δεν ανήκουν σε ένα

θέμα, παρά να τις κατατάξει σε αυτό. Ακόμα υπάρχει πάντα το ενδεχόμενο, όπως αναφέραμε παραπάνω, οι εικόνες του δείγματος να μην είναι καθαρά αντιπροσωπευτικές του θέματος ή οι αξιολογήσεις των χρηστών να περιέχουν ανθρώπινα σφάλματα. Σε γενικές γραμμές όμως, τα αποτελέσματα είναι τα επιθυμητά και δείχνουν ότι το πρόγραμμα έχει πάρει τις σωστές αποφάσεις για την κατάταξη των εικόνων.

Στην παρακάτω εικόνα απεικονίζονται οι προβλέψεις του προγράμματος για το σύνολο των εικόνων, των οποίων δεν έχουν βαθμολογηθεί από τους χρήστες. Παρατηρούμε το γεγονός που επισημάνθηκε παραπάνω, ότι δηλαδή οι προβλέψεις είναι πιο συνεπείς για τις εικόνες οι οποίες δεν ανήκουν στη θεματική ενότητα, που έχουμε επιλέξει.



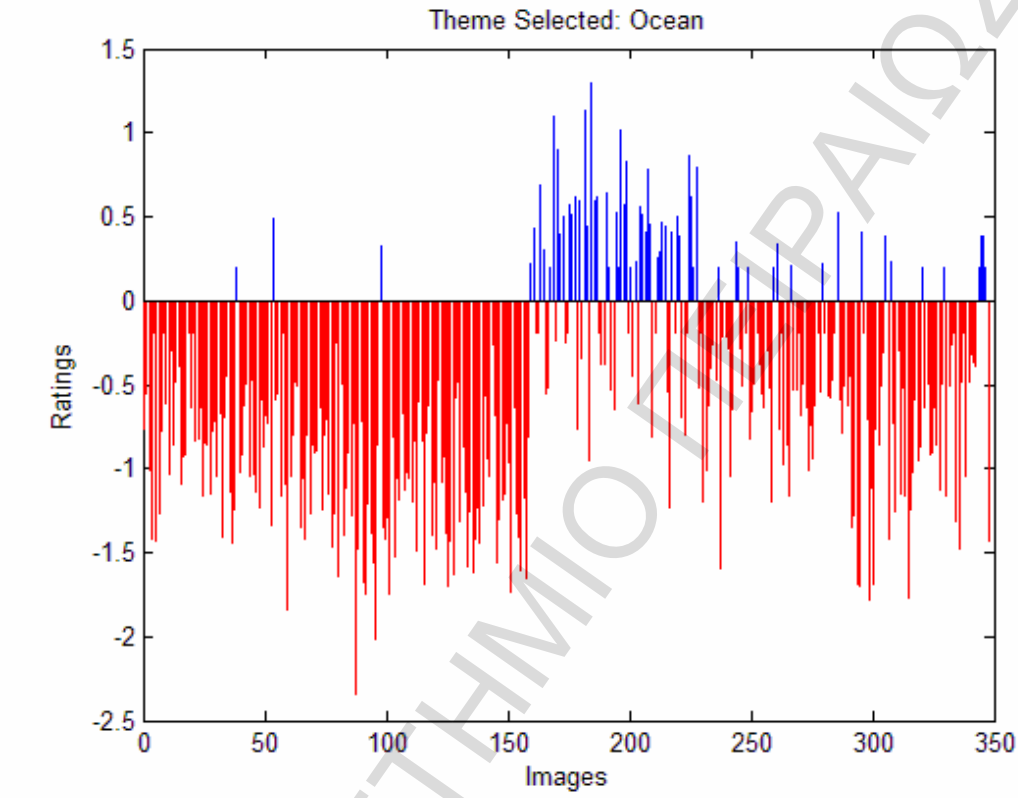
Εικόνα 6.3

Μια ακόμα σημαντική παρατήρηση έχει να κάνει με τη συμπεριφορά του προγράμματος σε σχέση με τη φύση του θέματος. Παρατηρούμε ότι αν το επιλεγμένο θέμα έχει σαφές περιεχόμενο, τότε το σφάλμα των προβλέψεων του προγράμματος είναι μικρότερο, και οι βαθμολογίες είναι σε μεγαλύτερο ποσοστό σωστές. Αντίθετα αν το περιεχόμενο του θέματος δεν είναι τόσο συγκεκριμένο, υπάρχουν μεγαλύτερα ποσοστά στις λάθος προβλέψεις. Και σε αυτή την περίπτωση η πλειονότητα των προβλέψεων κινούνται στη σωστή κατεύθυνση, αλλά το ποσοστό σφάλματος είναι μεγαλύτερο από το αντίστοιχο των περιπτώσεων όπου το θέμα έχει πιο συγκεκριμένο περιεχόμενο.

Για παράδειγμα να αναφέρουμε ότι όταν το επιλεγμένο θέμα είναι το “Animals”, παρατηρούμε ότι ενώ σε γενικές γραμμές οι προβλέψεις είναι σωστές, υπάρχουν και μεγάλος αριθμός

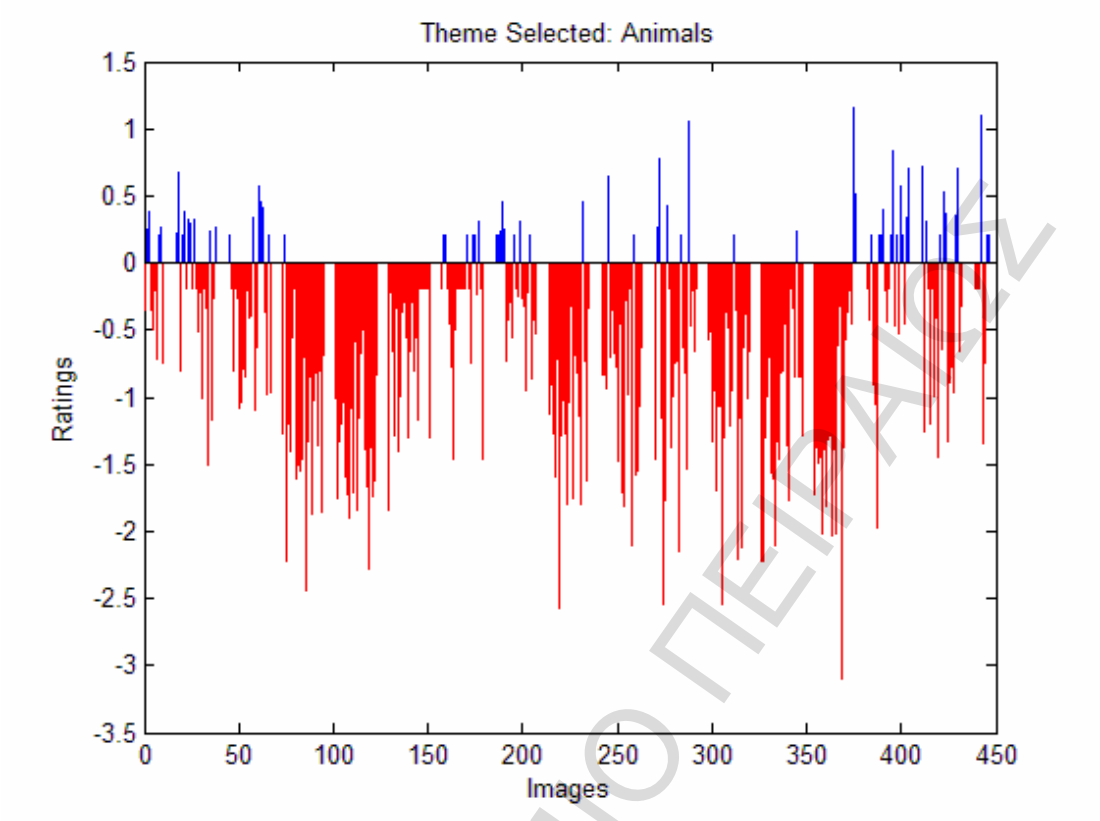
λανθασμένων. Στο θέμα “Ocean” παρατηρούμε ότι υπάρχει μεγαλύτερο ποσοστό επιτυχίας στις προβλέψεις, από το θέμα “Animals”. Αυτές οι αυξομειώσεις στις σωστές προβλέψεις είναι αναμενόμενες, αφού είναι κατανοητό ότι οι εικόνες με πιο σαφές και ενιαίο περιεχόμενο είναι πιο εύκολο να κατηγοριοποιηθούν, και κατά επέκταση να βαθμολογηθούν σωστά. Αξίζει να σημειωθεί πάντως, ότι όσο αφορά τις προβλέψεις για τις εικόνες που δεν ανήκουν στο σύστημα, το πρόγραμμα έχει σε ίδιο βαθμό επιτυχημένες προβλέψεις.

Στο παρακάτω γράφημα απεικονίζονται οι αποφάσεις για το θέμα “Ocean”.



Εικόνα 6.4

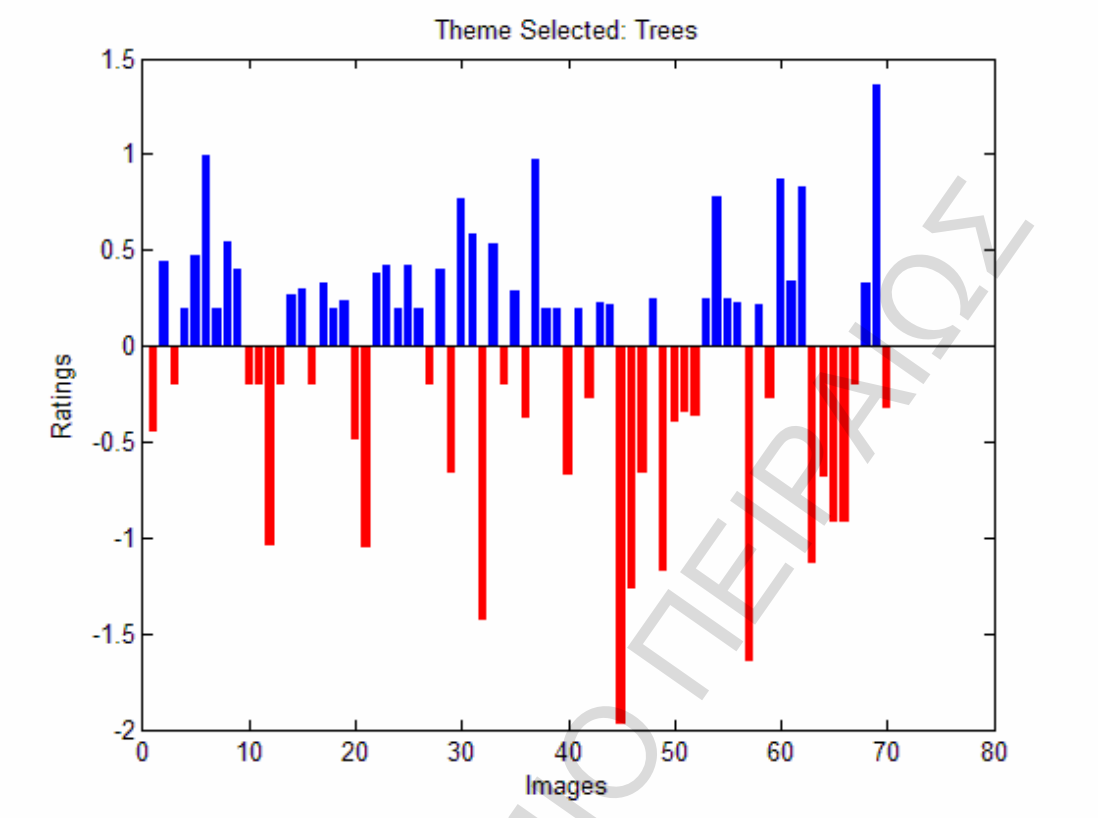
Στο παρακάτω γράφημα απεικονίζονται οι αποφάσεις για το θέμα “Animals”.



Εικόνα 6.5

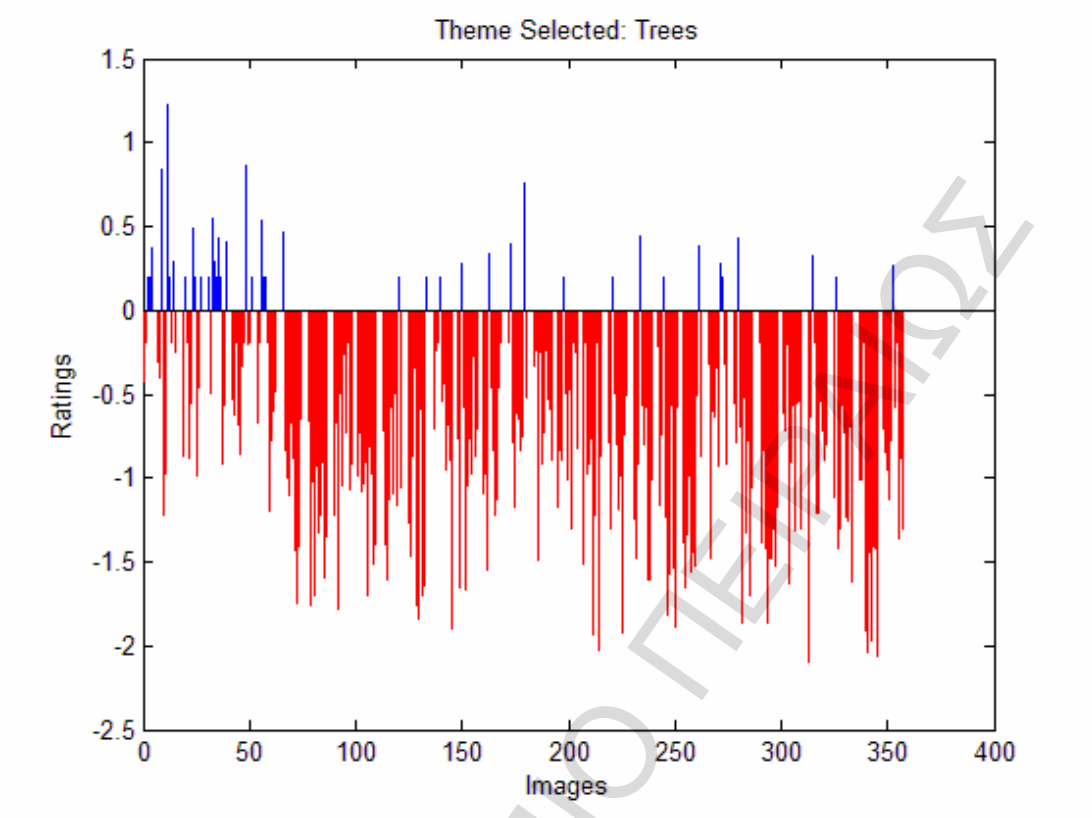
Άλλη μια σημαντική παρατήρηση είναι ότι το πρόγραμμα έχει σωστή απόδοση στην ανάκτηση των εικόνων, ακόμα και στις περιπτώσεις όπου το σύνολο των βαθμολογημένων εικόνων είναι μικρό και ο αλγόριθμος έχει εκπαιδευτεί μόνο μια φορά. Σκοπός της εργασίας είναι η δημιουργία ενός συστήματος, το οποίο θα υλοποιεί τη διαδικασία σταδιακής βαθμολόγησης των εικόνων. Είναι αναμενόμενο λοιπόν, τα αποτελέσματα στις περιπτώσεις αυτές να έχουν μεγαλύτερη απόκλιση από τις πραγματικές τιμές, σε σχέση με τις περιπτώσεις, όπου το σύνολο εκπαίδευσης είναι μεγαλύτερο και η βαθμολόγηση έχει γίνει σε περισσότερα από ένα στάδια. Από αυτό συμπεραίνουμε ότι ο αλγόριθμος έχει καλύτερη απόδοση όταν τον τροφοδοτούμε με μεγάλα σύνολα εκπαίδευσης και φυσικά όταν πραγματοποιείται σταδιακή βαθμολόγηση των εικόνων. Όμως σε γενικές γραμμές ο αλγόριθμος έχει σωστή απόδοση ακόμα και με μικρό δείγμα τιμών, γεγονός που καταδεικνύει τη δύναμη του αλγόριθμου που μπορεί και σε αυτές τις περιπτώσεις να παράγει αποτελέσματα με μεγάλο ποσοστό επιτυχίας

Στην παρακάτω εικόνα εμφανίζονται τα αποτελέσματα για μια τέτοια πειραματική μέτρηση, όπου ο αριθμός των βαθμολογιών από τους χρήστες είναι μικρότερος σε σχέση με τα προηγούμενα παραδείγματα και βασίζεται σε μία μόνο εκπαίδευση του αλγόριθμου.



Εικόνα 6.6

Παρατηρούμε ότι το πρόγραμμα στην περίπτωση αυτή, ενώ έχει ανακτήσει πολλές εικόνες σωστά, δεν έχει την ίδια απόδοση όπως στις προηγούμενες περιπτώσεις. Όπως θα δούμε όμως στην επόμενη εικόνα, ο αλγόριθμος έχει πολύ καλύτερη συμπεριφορά για την ταξινόμηση των εικόνων, οι οποίες δεν ανήκουν στο επιλεγμένο θέμα.



Εικόνα 6.7

Σαν συνολική εικόνα, παρατηρούμε ότι ο αλγόριθμος, σε γενικές γραμμές, ανακτά με μεγάλο ποσοστό επιτυχίας εικόνες που ανήκουν στην αναζητούμενη κλάση. Όπως αναφέραμε υπάρχουν και οι περιπτώσεις, όπου κάτω από συγκεκριμένες συνθήκες τα αποτελέσματα μας παρουσιάζουν μεγαλύτερο σφάλμα σε σχέση με τις υπόλοιπες δοκιμές. Αλλά ακόμα και σε αυτές τις περιπτώσεις, η απόδοση του αλγόριθμου κρίνεται ικανοποιητική.

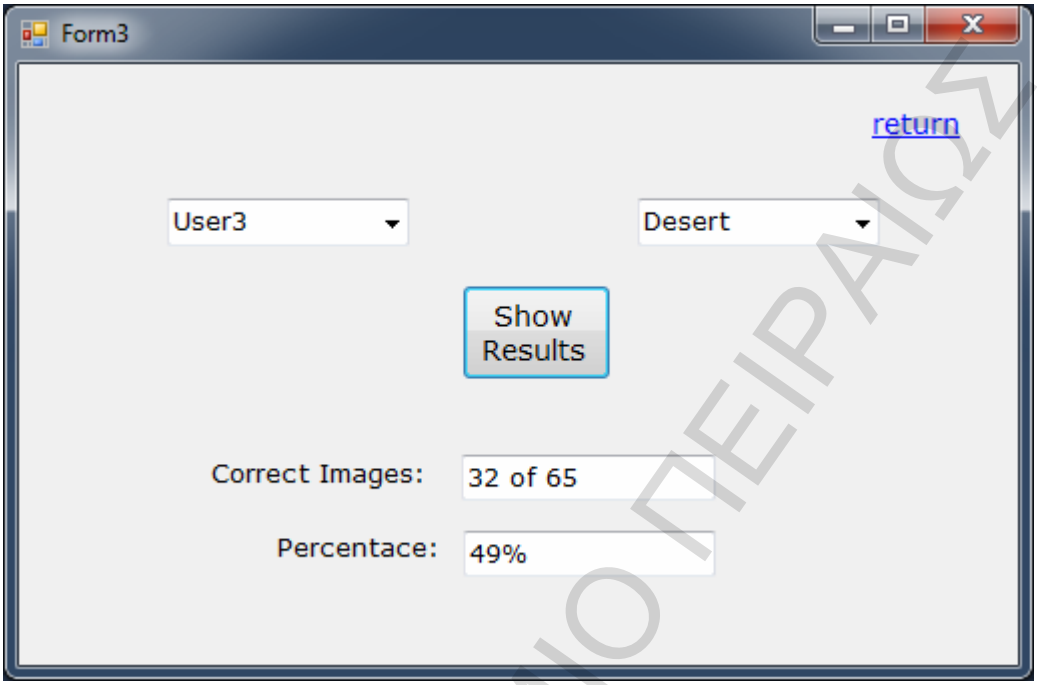
6.5 Παράδειγμα Σταδιακής Βαθμολόγησης Εικόνων

Παρακάτω θα περιγράψουμε τη διαδικασία της σταδιακής βαθμολόγησης των εικόνων, μέσα από μια σειρά δοκιμών στο πρόγραμμά μας. Για την πιο αναλυτική περιγραφή των αποτελεσμάτων, θα εστιάσουμε σε ένα παράδειγμα, όπου θα επιλέγουμε μια συγκεκριμένη θεματική ενότητα και θα παρουσιάσουμε τα αποτελέσματα του συστήματος για όλα τα στάδια της διαδικασίας. Σαν θεματική ενότητα, επιλέξαμε την κλάση "Desert", για τον λόγο ότι αποτελεί ενότητα, στην οποία παρατηρήσαμε ότι το σύστημα μας έχει θετική συμπεριφορά.

Όπως έχουμε αναφέρει στο πρώτο στάδιο του συστήματος, ο χρήστης έχει τη δυνατότητα να βαθμολογήσει μια σειρά από εικόνες σε κλίμακα από το ένα έως το πέντε, με βάση το περιεχόμενό τους ως προς την επιλεγμένη θεματική ενότητα. Κατά τη διάρκεια του πρώτου session, οι εικόνες εμφανίζονται στον χρήστη σε τυχαία σειρά από το σύνολο των θεματικών ενοτήτων. Στο παράδειγμα μας ο χρήστης βαθμολογεί τις εικόνες ως προς τη συνάφεια τους με την κλάση "Desert".

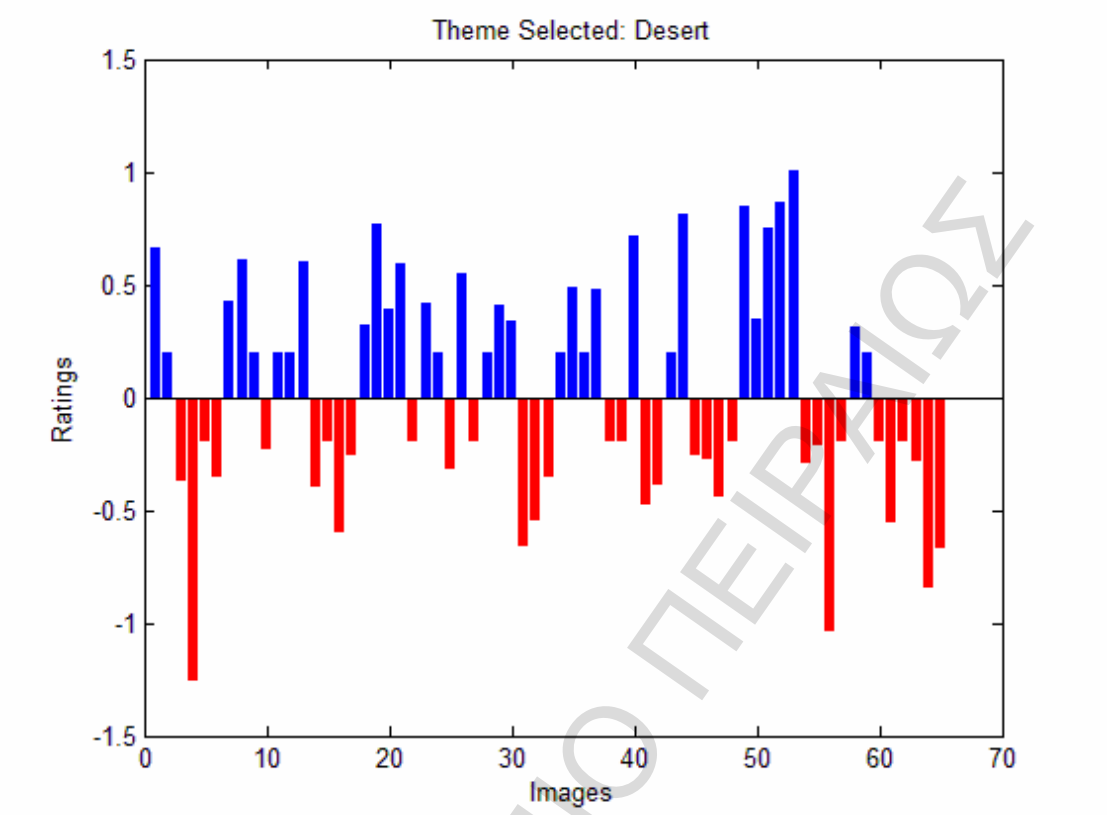
Μετά το πρώτο στάδιο βαθμολόγησης από το χρήστη, καλούμε τον αλγόριθμο μηχανικής μάθησης και μας επιστρέφει τις προβλέψεις του για τις εικόνες. Τα πρώτα αποτελέσματα του

συστήματος εμφανίζονται στην παρακάτω γραφική αναπαράσταση. Συνοπτικά τα αποτελέσματα εμφανίζονται στην παρακάτω εικόνα.



Εικόνα 6.8

Παρατηρούμε ότι ήδη από το πρώτο session, ο αλγόριθμος έχει υψηλό ποσοστό επιτυχίας. Από τις 65 εικόνες που ανήκουν στην κατηγορία "Desert" και δεν τις έχει βαθμολογήσει ο χρήστης, το σύστημα κατατάσσει τις 32 από αυτές στη σωστή κατηγορία. Το ποσοστό επιτυχίας για το συγκεκριμένο session είναι 49%, ποσοστό αρκετά υψηλό. Οι αναλυτικές προβλέψεις του αλγόριθμου για τις εικόνες του επιλεγμένου θέματος παρουσιάζονται στην παρακάτω γραφική παράσταση. Όπως έχουμε αναφέρει με μπλε χρώμα αναπαριστώνται οι εικόνες που έχει ανακτήσει σωστά ο αλγόριθμος, και με κόκκινο τις εικόνες που δεν έχει ανακτήσει.



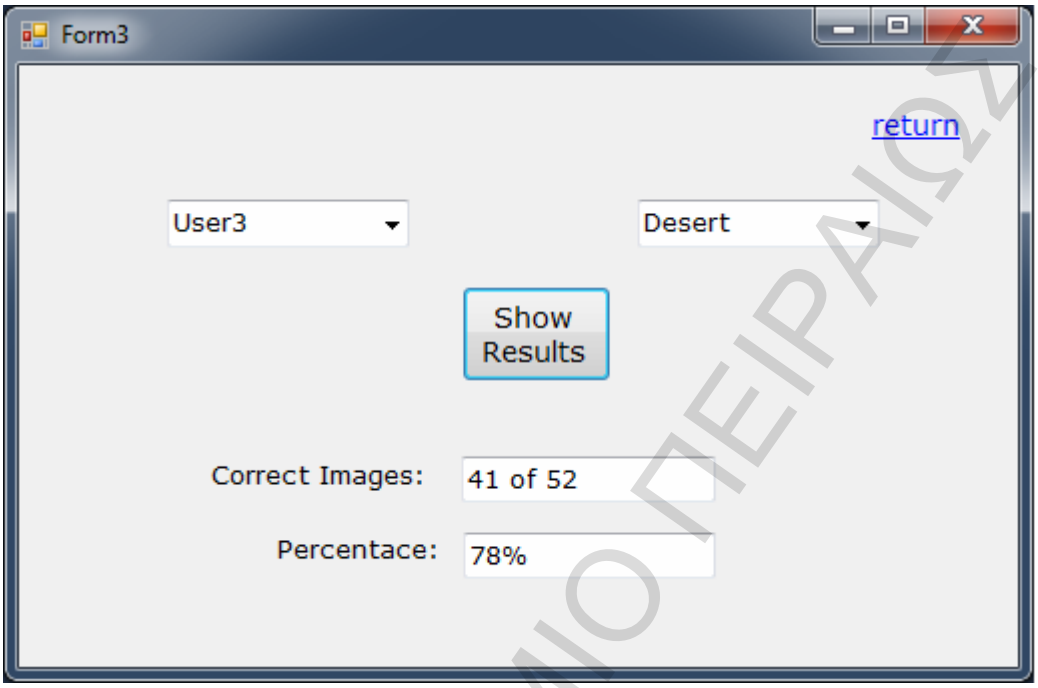
Εικόνα 6.9

Από το δεύτερο session οι εικόνες θα εμφανίζονται στην χρήστη όχι με τυχαία σειρά, αλλά με βάση της βαθμολόγησης του αλγόριθμου από το πρώτο session, εμφανίζοντας πρώτες τις εικόνες με τη μεγαλύτερη βαθμολογία. Με αυτόν τον τρόπο είναι πιο πιθανό να παρουσιαστούν στον χρήστη εικόνες που πράγματι ανήκουν στην επιλεγμένη κλάση. Στην παρακάτω εικόνα παρουσιάζονται οι τρεις πρώτες εικόνες, που το σύστημα θεωρεί ότι ανήκουν στην κατηγορία “Desert”.

Εικόνα 6.10

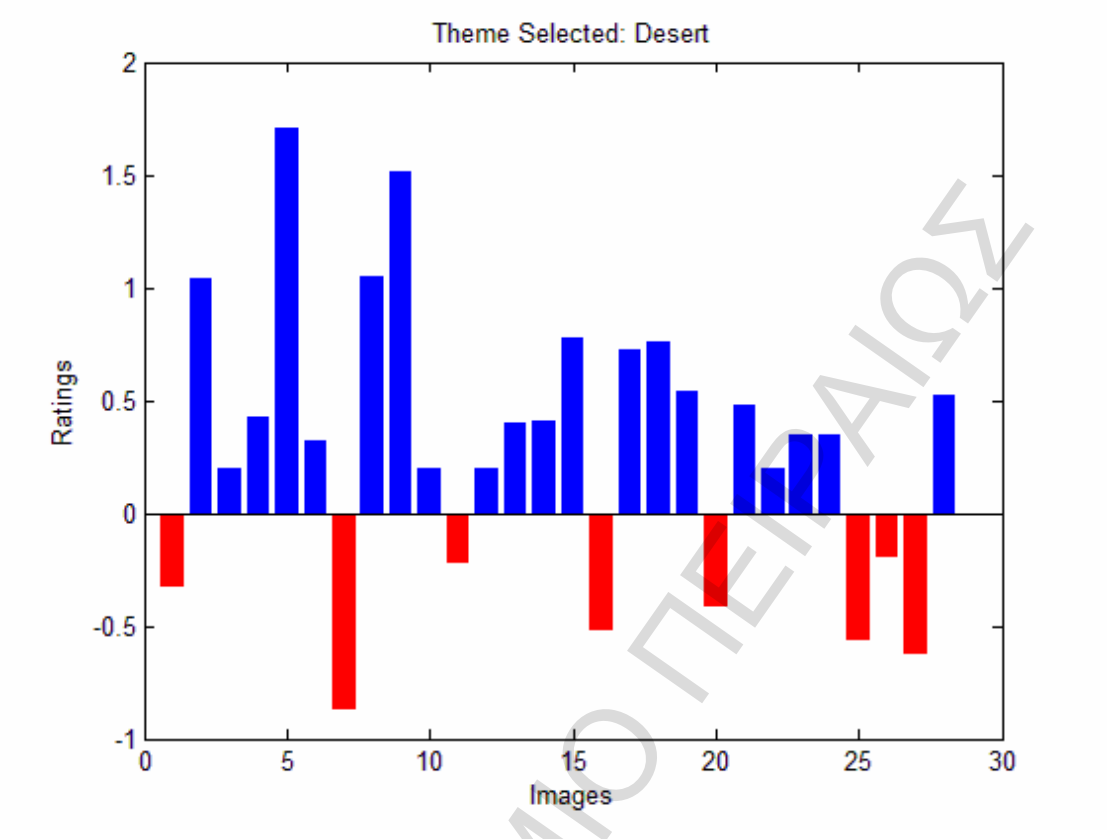
Όπως παρατηρούμε από τις εικόνες το σύστημα ήδη έχει εκπαιδευτεί σωστά από το πρώτο session, και οι εικόνες που μας παρουσιάζει ανήκουν πράγματι στο επιλεγμένο θέμα, δηλαδή αυτό της ερήμου.

Μετά τη δεύτερη περιήγηση του χρήστη στις εικόνες και τη βαθμολόγηση τους, καλούμε ξανά τον αλγόριθμο μηχανικής μάθησης και προσθέτουμε στο προηγούμενο σύνολο εκπαίδευσης τις νέες τιμές. Αφού ο αλγόριθμος εκπαιδευτεί για δεύτερη φορά μας επιστρέφει τα νέα αποτελέσματα, καθώς και το καινούργιο ποσοστό επιτυχούς ανάκτησης των εικόνων.



Εικόνα 6.11

Στην παραπάνω εικόνα παρατηρούμε ότι οι προβλέψεις του αλγόριθμου έχουν πολύ υψηλό ποσοστό επιτυχίας. Από τις 52 εικόνες που δεν έχει βαθμολογήσει ο χρήστης και ανήκουν στην κατηγορία “Desert”, το σύστημα έχει ανακτήσει τις 41. Σε ποσοστό επιτυχίας το σύστημα έχει ποσοστό ίσο με 78%. Στην παρακάτω γραφική παράσταση παρουσιάζονται λεπτομερώς οι προβλέψεις του συστήματος για τις εναπομείναντες εικόνες.



Εικόνα 6.12

Είναι εμφανές ότι τα αποτελέσματα των δοκιμών μας στο συγκεκριμένο παράδειγμα, είναι κάτι παραπάνω από ικανοποιητικά. Είναι αξιοσημείωτο ότι στο παράδειγμα που επιλέξαμε να παρουσιάσουμε, το σύστημα δεν χρειάστηκε πάνω από δυο εκπαιδεύσεις για να φτάσει σε ένα πολύ υψηλό ποσοστό επιτυχίας στην ανάκτηση εικόνων, που ανήκουν στην αναζητούμενη κλάση. Πρέπει να σημειώσουμε ότι δεν είχαμε τα ίδια ικανοποιητικά αποτελέσματα σε όλες τις δοκιμές που εφαρμόσαμε στο πρόγραμμά μας. Άλλωστε ένας λόγος που επιλέξαμε το συγκεκριμένο θέμα ήταν το γεγονός, ότι είχαμε παρατηρήσει ότι το σύστημα μας είχε θετική συμπεριφορά με τη συγκεκριμένη θεματική ενότητα. Όπως έχουμε αναφέρει και προηγουμένως, σε μερικές θεματικές ενότητες η απόδοση του συστήματος δεν είναι τόσο καλή. Όμως ακόμα και σε αυτές τις περιπτώσεις, σύμφωνα με τις δοκιμές μας, το ποσοστό επιτυχούς ανάκτησης των εικόνων μπορεί να φτάσει σε επιθυμητά επίπεδα. Αυτό μπορεί να επιτευχθεί με την επέκταση της διαδικασίας της σταδιακής βαθμολόγησης των εικόνων σε περισσότερα στάδια, από ότι στην περίπτωση του παραδείγματος μας.

Επίλογος

Στην εργασία ασχοληθήκαμε με την ανάπτυξη ενός συστήματος, το οποίο επιχειρεί την ανάκτηση εικόνων με βάση το περιεχόμενό τους, χρησιμοποιώντας τεχνικές ημι-επιτηρούμενης μηχανικής μάθησης. Σκοπός της εργασίας είναι η επιτυχημένη διαδικασία σταδιακής βαθμολόγησης των εικόνων από το σύστημα σε σχέση με ένα συγκεκριμένο θέμα, εκμεταλλευόμενο την πληροφορία που έχει λάβει από τους χρήστες. Για την εκπαίδευση του συστήματος επιλέξαμε σαν αλγόριθμο μηχανικής μάθησης τον Transductive Semi Vector Machine (TSVM) αλγόριθμο.

Η λειτουργία του συστήματος μπορεί να συνοψιστεί στα παρακάτω βήματα. Αρχικά επιλέγεται το σύνολο των εικόνων, το οποίο θα αποτελεί τη συλλογή των δεδομένων εκπαίδευσης και δοκιμής του συστήματος. Το εργαλείο επεξεργάζεται τα δεδομένα αυτά, σύμφωνα με τις ανάγκες του σκοπού μας. Στη συνέχεια γίνεται εξαγωγή των χρήσιμων μεταβλητών από τη συλλογή των δεδομένων. Πιο συγκεκριμένα το εργαλείο εξάγει τα, χαμηλού επιπέδου, οπτικά χαρακτηριστικά τους και τα επεξεργάζεται κατάλληλα, ώστε να μπορούν να αξιοποιηθούν στη συνέχεια. Το επόμενο βήμα είναι η συλλογή των παραδειγμάτων με τιμή, τα οποία θα αποτελέσουν τα δεδομένα εκπαίδευσης του συστήματος μας. Η ανάθεση των τιμών αυτών πραγματοποιείται από τους χρήστες του συστήματος. Οι χρήστες μπορούν να περιηγηθούν στη συλλογή των εικόνων και να τις βαθμολογούν, με βάση τη συνάφεια τους ως προς ένα συγκεκριμένο θέμα. Οι τιμές αυτές σε συνδυασμό με τις μεταβλητές των εικόνων, θα αποτελέσουν το συνολικό δείγμα του αλγόριθμου μάθησης. Για την εργασία μας έχουμε επιλέξει τον αλγόριθμο Transductive Semi Vector Machine (TSVM).

Το πρόγραμμα μας επιστρέφει σαν αποτέλεσμα, τις προβλέψεις του αλγόριθμου για τη συλλογή των δεδομένων που δεν έχουν τιμή, δηλαδή τις εικόνες εκείνες που δεν έχουν βαθμολογήσει οι χρήστες και ανήκουν στο ζητούμενο θέμα. Ακόμα θα εμφανίζει στο χρήστη το ποσοστό των εικόνων που έχει ανακτήσει σωστά από την αναζητούμενη κλάση. Στις επόμενες περιηγήσεις το σύστημα θα εμφανίζει στους χρήστες μόνο τις εικόνες που δεν έχουν βαθμολογήσει. Οι εικόνες αυτές θα εμφανίζονται με βάση την τιμή που τους έχει προσδώσει ο αλγόριθμος, από την μεγαλύτερη στην μικρότερη. Με αυτόν τον τρόπο το σύστημα θα εμφανίζει στους χρήστες τι εικόνες αυτές, που σύμφωνα με τα αποτελέσματα του, είναι πιο πιθανό να ανήκουν στο επιλεγμένο θέμα. Στο τέλος κάθε βαθμολόγησης, υπάρχει η δυνατότητα της νέας εκπαίδευσης του αλγόριθμου, με τις νέες βαθμολογίες να προστίθενται στο σύνολο εκπαίδευσης του αλγόριθμου, ο οποίος θα εξάγει τις νέες προβλέψεις για τις βαθμολογίες των εικόνων και το νέο ποσοστό επιτυχούς ανάκτησης τους από την αναζητούμενη θεματική ενότητα.

Σε γενικές γραμμές τα αποτελέσματα του εργαλείου κρίνονται ικανοποιητικά. Στην αξιολόγηση των αποτελεσμάτων του προγράμματος μας, παρατηρούμε ότι οι τιμές που έχει προσδώσει στη συλλογή των δεδομένων, κινούνται συνολικά προς στη σωστή κατεύθυνση. Το πρόγραμμα δηλαδή έχει τη δυνατότητα να βαθμολογεί εικόνες σε σχέση με μια επιλεγμένη θεματική ενότητα. Το ποσοστό λάθους στην βαθμολόγηση των εικόνων κρίνεται ότι βρίσκεται σε επιθυμητό επίπεδο και στην πλειονότητα τους οι προβλέψεις του συστήματος δεν απέχουν από τις πραγματικές βαθμολογίες των εικόνων. Καταλήγουμε στο συμπέρασμα, από τα αποτελέσματα των πειραματικών μετρήσεων, ότι το πρόγραμμά μας έχει τη δυνατότητα, με αυτοματοποιημένη διαδικασία και αποτελεσματικό τρόπο, να ανακτά εικόνες με βάση του περιεχόμενό τους.

Παρατηρούμε ότι το ποσοστό λάθους στην ανάκτηση των εικόνων κυμαίνεται ανάλογα με το περιεχόμενο τους και τις βαθμολογίες των χρηστών. Είναι εμφανές ότι ο αλγόριθμος έχει καλύτερη συμπεριφορά όταν η συλλογή αποτελείται από εικόνες, όπου το περιεχόμενο τους είναι αντιπροσωπευτικό της θεματικής ενότητας στην οποία ανήκουν. Ακόμα παρατηρούμε ότι οι βαθμολογίες των χρηστών, όσον αφορά την ορθότητα και την ακρίβεια τους, επηρεάζουν τις προβλέψεις του συστήματος, γεγονός το οποίο είναι αναμενόμενο, αφού οι βαθμολογίες αυτές αποτελούν το σύνολο εκπαίδευσης του αλγόριθμου μηχανικής μάθησης.

Βιβλιογραφία

[1] **Xiaojin Zhu and Andrew B. Goldberg.** Introduction to Semi-Supervised Learning. (2009)
[2] **Olivier Chapelle, Bernhard Scholkopf and Alexander Zien.** Semi-Supervised Learning. (2006)
[3] **Xiaojin Zhu.** Semi-Supervised Learning Literature Survey. (2008)
[4] **A. Gammerman, V. Vovk and V. Vapnik.** Learning by Transduction
[5] **Dmitry Pechyony.** Theory and Practice of Transductive Learning. (2008)

[6] **Olivier Bousquet.** Transductive Learning: Motivation, Model, Algorithms. (2002)
[7] **Thorsten Joachims.** Transductive Inference for Text Classification using Support Vector Machines.
[8] **Junhui Wang, Xiaotong Shen and Wei Pan.** On Transductive Support Vector Machines.
[9] **Lorenzo Bruzzone, Mingmin Chi, Mattia Marconcini.** A Novel Transductive SVM for Semi-supervised Classification of Remote-Sensing Images
[10] **Evaggelos Spyrou, Giorgos Tolias, Phivos Mylonas and Yannis Avrithis.** Concept detection and keyframe extraction using a visual thesaurus. (2008)
[11] **Thorsten Joachims**. Estimating the Generalization Performance of an SVM Efficiently.
[12] **Ralf Klinkenberg and Thorsten Joachims.** Detecting Concept Drift with Support Vector Machines. (2000)