

Μεθοδολογίες Μηχανικής Μάθησης Βασισμένες σε Τεχνητά Ανοσοποιητικά Συστήματα

Διονύσιος Ν. Σωτηρόπουλος

Πανεπιστήμιο Πειραιώς

October 27, 2010

Πίνακας Περιεχομένων I

- 1 Κίνητρα Διατριβής
- 2 Μηχανική Μάθηση
 - Ορισμοί
 - Μάθηση μέσω Παραδειγμάτων
- 3 Το Πρόβλημα της Ταξικής Ανισορροπίας
 - Ορισμός
 - Η Φύση του Προβλήματος της Ταξικής Ανισορροπίας
- 4 Φυσικά Ανοσοποιητικά Συστήματα
 - Ανοσοποιητικό Σύστημα
 - Υπολογιστική Θεώρηση Προσαρμόσιμου Ανοσοποιητικού Συστήματος
- 5 Τεχνητά Ανοσοποιητικά Συστήματα
 - Ορισμός και Εφαρμογές Τεχνητών Ανοσοποιητικών Συστημάτων
 - Γενικό Πλαίσιο Ανάπτυξης Τεχνητών Ανοσοποιητικών Συστημάτων
 - Μοντέλο Χώρου Σχημάτων
 - Σύνολο Δεδομένων
 - Ομαδοποίηση Μουσικών Διανυσμάτων Χαρακτηριστικών
 - Αυτόματη Ταξινόμηση Μουσικολογικού Είδους
 - Σύσταση Μουσικών Δεδομένων
 - Μέθοδοι Σύστασης με Δομή Καταρράκτη
 - Αξιολόγηση Συστήματος Σύστασης
- 6 Συνεισφορά Διατριβής και Μελλοντική Εργασία
- 7 Βιβλιογραφία

Κίνητρα Διατριβής I

- Το ερευνητικό πεδίο της παρούσας διδακτορικής διατριβής είναι αυτό του *Βιολογικά Εμπνευσμένου Υπολογισμού (Biologically Inspired Computing)*, προϊόν της μακρόχρονης και διμερούς αλληλεπίδρασης μεταξύ των επιστήμων της Πληροφορικής και της Βιολογίας.
- Η παρούσα διδακτορική διατριβή στοχεύει στην αξιοποίηση του γνωστικού πεδίου της Βιολογίας, συγκεκριμένα της Ανοσολογίας, ως μιας έγκυρης μεταφοράς για τη δημιουργία αφηρημένων, υψηλού επιπέδου, αναπαραστάσεων των βιολογικών συστατικών και λειτουργιών.
- Μια μεταφορά λειτουργεί ως μια απεικόνιση μεταξύ δύο επιστημονικών πεδίων, με σκοπό να ερμηνεύσει ένα σύνολο ιδεών και πεποιθήσεων κατά τέτοιο τρόπο ώστε να το καταστήσει εφαρμόσιμο σε μια διαφορετική περιοχή από αυτήν στην οποία έγινε η αρχική διατύπωση τους.

Κίνητρα Διατριβής II

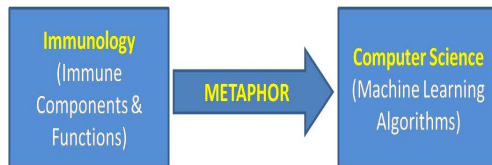


Figure: Η Ανοσολογία ως πηγή μεταφορών για την ανάπτυξη αλγορίθμων Μηχανικής Μάθησης.

Γενικά Χαρακτηριστικά Βιολογικών Συστημάτων που τα καθιστούν ως έγκυρη πηγή μεταφορών:

- 1 **Αρχιτεκτονική:** αναφέρεται στη μορφή και τη δομή των βιολογικών συστημάτων.
- 2 **Λειτουργικότητα:** αντιστοιχεί στην συμπεριφορά των βιολογικών συστημάτων.
- 3 **Μηχανισμοί:** χαρακτηρίζουν τη συνεργασία και την αλληλεπίδραση μεταξύ των θεμελιωδών συστατικών των βιολογικών συστημάτων.
- 4 **Οργάνωση:** αναφέρεται στους τρόπους με τους οποίους οι δραστηριότητες των βιολογικών συστημάτων μπορούν να εκφραστούν μέσω της δυναμικής του συνόλου.

Κίνητρα Διατριβής III

Τα τελευταία χρόνια η πρόοδος που σημειώθηκε στις επιστήμες της βιολογίας και της μοριακής γενετικής οδήγησε σε ραγδαία αύξηση την κατανόησή μας για τον τρόπο συμπεριφοράς του ανοσοποιητικού συστήματος.

Η γνώση αυτή αποκάλυψε πολλούς από τους κύριους λειτουργικούς μηχανισμούς του, οι οποίοι αποδείχθηκαν πολύ ενδιαφέροντες όχι μόνο από βιολογική αλλά και από υπολογιστική άποψη.

Γενικά Χαρακτηριστικά Ανοσοποιητικού Συστήματος (ΑΣ) που το καθιστούν ως έγκυρη πηγή μεταφορών:

- 1 *Πολυπλοκότητα*: το ΑΣ των σπονδυλωτών αποτελεί ένα από τα πιο πολύπλοκα σωματικά συστήματα η πολυπλοκότητα του οποίου είναι πολλές φορές συγκρίσιμη με αυτή του εγκεφάλου (Εγκέφαλος: 10^{10} νευρώνες, Ανοσοποιητικό Σύστημα: 10^{12} λεμφοκύτταρα).
- 2 *Αυτοοργάνωση*: Κύριο χαρακτηριστικό του ΑΣ είναι ότι δεν υπάρχει κάποιο κεντρικό όργανο για τον συντονισμό της λειτουργίας του.
- 3 *Αναγνώριση Προτύπων*: Το ΑΣ έχει την ικανότητα αναγνώρισης μιας σχεδόν ανεξάντλητης γκάμας αντιγονικών προτύπων χωρίς να είναι απαραίτητη η πρότερη έκθεσή του σε αυτά (Διάκριση Εαυτού / Μη-Εαυτού, Self / Non-Self Discrimination).

Κίνητρα Διατριβής IV

- 4** *Προβλήματα Ταξικής Ανισορροπίας:* Η Διάκριση Εαυτού / Μη-Εαυτού αποτελεί ένα εξαιρετικά ετεροβαρές πρόβλημα ταξινόμησης καθώς η κλάση Εαυτός είναι κατά πολύ μικρότερη από την κλάση Μη-Εαυτός.
- 5** *Μονοταξική Ταξινόμηση:* Η διαδικασία της Αρνητικής Επιλογής (Negative Selection) συνιστά ένα μοντέλο Μονοταξικής Ταξινόμησης καθώς η κλάση πλειοψηφίας αγνοείται πλήρως κατά τη φάση της εκπαίδευσης.
- 6** *Μάθηση:* Η Θεωρία της Επιλεκτικής Κλωνοποίησης κατατάσσει τη μαθησιακή ικανότητα του ΑΣ στα μοντέλα Ενισχυόμενης Μάθησης (Reinforcement Learning) ενώ η Θεωρία του Ανοσοποιητικού Δικτύου την αποδίδει στη δυναμική συμπεριφορά των μοριακών συνιστωσών του.
- 7** *Μνήμη:* Η Θεωρία του Ανοσοποιητικού Δικτύου ερμηνεύει την Ανοσοποιητική Μνήμη μέσω της εμφάνισης Ανοσολογικών Κύκλων. Το ΑΣ έχει την ικανότητα να διατηρεί τις εσωτερικές εικόνες των αντιγονικών προτύπων στα οποία έχει εκτεθεί στο παρελθόν (Immunological Cycles).
- 8** *Ικανότητα Γενίκευσης:* Η Θεωρία της Επιλεκτικής Κλωνοποίησης αποδίδει στην Ανοσοποιητική Μνήμη συσχετιστικό χαρακτήρα. Μέσω του φαινομένου της Ανοσοποιητικής Διάδρασης το ΑΣ έχει την ικανότητα να αναγνωρίζει δομικά συσχετισμένα αντιγονικά πρότυπα (Generalization Ability).

Μάθηση - Μηχανική Μάθηση

- Η μαθησιακή ικανότητα αποτελεί μία από τις πιο χαρακτηριστικές ιδιότητες της ευφυούς συμπεριφοράς.
- Η διαδικασία της *μάθησης* περιλαμβάνει:
 - την απόκτηση νέας δηλωτικής γνώσης;
 - την ανάπτυξη νέων δεξιοτήτων μέσω αλληλεπίδρασης ή εξάσκησης;
 - την οργάνωση της νεοαποκτηθείσας γνώσης σε γενικές, λειτουργικές αναπαραστάσεις; και
 - την ανακάλυψη νέων γεγονότων και θεωριών μέσω της παρατήρησης και του πειραματισμού.
- Ο όρος *Μηχανική Μάθηση* καλύπτει ένα ευρύ φάσμα προγραμμάτων υπολογιστών με κύριο χαρακτηριστικό τους την βελτίωση των επιδόσεών τους μέσω της εκπαίδευσης και της απόκτησης εμπειρίας.
- Η Μηχανική Μάθηση αποτελεί ένα αναπόσπαστο μέρος της *Τεχνητής Νοημοσύνης* καθώς το πρωτογενές γνώρισμα κάθε ευφυούς συστήματος είναι η ικανότητα του να μαθαίνει.

Μοντέλο Μάθησης μέσω Παραδειγμάτων (Learning from Examples) I

- Η Μάθηση μέσω Παραδειγμάτων αποτελεί το υπόδειγμα που υιοθετήθηκε προκειμένου να εκτιμηθεί η συναρτησιακή εξάρτηση που διέπει ένα δοσμένο σύνολο πειραματικών δεδομένων και ανήκει στο γενικότερο υπόδειγμα μάθησης του Μοντέλου της Πρόγνωσης ή Γενικής Συμπερασματολογίας.
- Τα θεμελιώδη συστατικά του Μοντέλου της Μάθησης μέσω παραδειγμάτων είναι τα εξής:
 - 1 Ο γεννήτορας (generator) των δεδομένων G ;
 - 2 Ο τελεστής του στόχου (target operator) ή τελεστής του επιτηρητή (*supervisor's operator*) S ;
 - 3 Η μηχανή μάθησης (learning machine) LM .

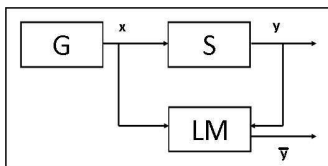


Figure: Μάθηση μέσω Παραδειγμάτων

Μοντέλο Μάθησης μέσω Παραδειγμάτων (Learning from Examples) II

- Ο γεννήτορας G αποτελεί τον κύριο περιβαλλοντικό παράγοντα η δράση του οποίου δημιουργεί τα ανεξαρτήτως και πανομοιότυπως καταναμημένα (i.i.d) τυχαία διανύσματα $x \in \mathbf{X}$, με βάση μια άγνωστη (αλλά αμετάβλητη) συνάρτηση κατανομής πιθανότητας $F(x)$.
- Τα τυχαία διανύσματα $x \in \mathbf{X}$ τροφοδοτούνται ως είσοδοι στον τελεστή του επιτηρητή, ο οποίος παρέχει τελικά την τιμή εξόδου y .
- Θα πρέπει να σημειωθεί πως δεν υπάρχει καμιά πληροφορία σχετικά με το μετασχηματισμό των διανυσμάτων εισόδου στις αντίστοιχες τιμές εξόδου.
- Η μόνη διαθέσιμη πληροφορία αφορά την ύπαρξη και τη μη-μεταβολή του τελεστή του επιτηρητή.
- Η μηχανή μάθησης παρατηρεί l ζεύγη της μορφής

$$(x_1, y_1), \dots, (x_l, y_l) \quad (1)$$

προσπαθώντας να κατασκευάσει έναν τελεστή ο οποίος έχει τη δυνατότητα να παρέχει την πρόβλεψη \bar{y}_i για την απόκριση του επιτηρητή y_i σε ένα δεδομένο διάνυσμα εισόδου x .

Μοντέλο Μάθησης μέσω Παραδειγμάτων (Learning from Examples) III

- Το κεντρικό αξίωμα της μηχανικής μάθησης αφορά τη δράση του τελεστή της επιτήρησης. Συγκεκριμένα, η παράγωγη των τιμών εξόδου γίνεται στη βάση μιας δεσμευμένης συνάρτησης κατανομής $F(y|x)$ η οποία περιλαμβάνει και την περίπτωση κατά την οποία ο τελεστής του επιτηρητή υλοποιεί μια συνάρτηση της μορφής $y = f(x)$.
- Η μηχανή μάθησης παρατηρώντας το δοσμένο σύνολο εκπαίδευσης προσπαθεί να κατασκευάσει μια προσέγγιση για την από κοινού συνάρτηση κατανομής $F(x, y) = F(x)F(y|x)$ η οποία δημιουργήσει τα παρατηρούμενα δεδομένα (Σχέση 1) ανεξαρτήτως και πανομοιότυπας.
- Το πρόβλημα της μηχανικής μάθησης ανάγεται στο πρόβλημα της επιλογής μιας κατάλληλης συνάρτησης από ένα δοσμένο σύνολο συναρτήσεων.
- Τυπικά η διαδικασία της μηχανικής μάθησης συνίσταται στην αναζήτηση εκείνης της συνάρτησης $g'(z)$ από ένα σύνολο αποδεκτών συναρτήσεων $\{g(z)\}$, $z \in Z$ σε ένα υποσύνολο Z του διανυσματικού χώρου \mathbb{R}^n η οποία ελαχιστοποιεί το παρακάτω συναρτησιοειδές καταλληλότητας:

$$R = R(g(z)) \quad (2)$$

- Στην περίπτωση κατά την οποία το σύνολο των συναρτήσεων $\{g(z)\}$, $z \in Z$ και το συναρτησιοειδές καταλληλότητας R δίνονται ρητώς, το πρόβλημα της μάθησης ανάγεται σε ένα πρόβλημα βελτιστοποίησης.

Μοντέλο Μάθησης μέσω Παραδειγμάτων (Learning from Examples) IV

- Στην πραγματικότητα το συναρτησιοειδές του ρίσκου ορίζεται μέσω μιας συνάρτησης κατανομής πιθανότητας $F(\mathbf{z})$ ορισμένης στο Z , ως ακολούθως:

$$R(g(\mathbf{z})) = \int L(\mathbf{z}, g(\mathbf{z})) dF(\mathbf{z}) \quad (3)$$

οπού η συνάρτηση απώλειας $L(\mathbf{z}, g(\mathbf{z}))$ είναι ολοκληρώσιμη για κάθε $g(\mathbf{z}) \in \{g(\mathbf{z})\}$.

Definition (Πρόβλημα της Μάθησης μέσω Παραδειγμάτων)

Το πρόβλημα της μάθησης μέσω παραδειγμάτων συνίσταται στην ελαχιστοποίηση του παρακάτω συναρτησιοειδούς του ρίσκου:

$$R(a) = \int L(\mathbf{z}, g(\mathbf{z}, \alpha)) dF(\mathbf{z}), \alpha \in \Lambda \quad (4)$$

σε ένα δοσμένο σύνολο συναρτήσεων $\{g(\mathbf{z}, \alpha), \alpha \in \Lambda\}$ στη βάση ενός δείγματος παρατηρήσεων

$$\mathbf{z}_1, \dots, \mathbf{z}_l \quad (5)$$

οι οποίες έχουν εξαχθεί με ανεξάρτητο και πανομοιότυπο τρόπο από μια άγνωστη συνάρτηση κατανομής πιθανότητας $F(\mathbf{z})$.

Μοντέλο Μάθησης μέσω Παραδειγμάτων (Learning from Examples) V

- Ανάλογα με τη συγκεκριμένη μορφή της συνάρτησης απώλειας προκύπτουν τα τρία θεμελιώδη προβλήματα της Αναγνώρισης Προτύπων:
 - 1 Πρόβλημα της Ταξινόμησης;
 - 2 Πρόβλημα της Παλινδρόμησης;
 - 3 Πρόβλημα της Εκτίμησης της Συνάρτησης Πυκνότητας Πιθανότητας (Ομαδοποίηση Δεδομένων).
- Τα προβλήματα 1 και 2 ανήκουν στην υποκατηγορία των προβλημάτων μάθησης *με επιτήρηση* ενώ το πρόβλημα 3 ανήκει στην υποκατηγορία των προβλημάτων μάθησης *χωρίς επιτήρηση*.

Definition (Πρόβλημα της Ομαδοποίησης Δεδομένων)

Έστω $X = \{x_1, \dots, x_l\}$ ένα αρχικό σύνολο δεδομένων. Μία m -ομαδοποίηση αυτού συνίσταται στη διαμέριση του συνόλου X σε m σύνολα (συστάδες / clusters) έτσι ώστε να ικανοποιούνται οι τρεις παρακατω συνθηκες:

- $C_i \neq \emptyset, i = 1, \dots, m$
- $\cup_{i=1}^m C_i = X$
- $C_i \cap C_j = \emptyset, i \neq j, i, j = 1, \dots, m$

Αρχές Ελαχιστοποίησης Εμπειρικού και Δομικού Ρίσκου

Definition (Αρχή Ελαχιστοποίησης Εμπειρικού Ρίσκου)

Η Αρχή της Ελαχιστοποίησης του Εμπειρικού Ρίσκου [Vapnik, 1995] συνίσταται στην αναδιατύπωση του προβλήματος της μηχανικής μάθησης μέσω της αντικατάστασης της συνάρτησης του πραγματικού ρίσκου (Σχέση 4) από το συναρτησιοειδές του εμπειρικού ρίσκου το οποίο δίνεται από τη σχέση:

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^l L(\mathbf{z}_i, g(\mathbf{z}_i, \alpha)) \quad (6)$$

Definition (Αρχή Ελαχιστοποίησης Δομικού Ρίσκου)

Η Αρχή της Ελαχιστοποίησης του Δομικού Ρίσκου [Vapnik, 1995] προτείνει πως η διαδικασία της μάθησης θα πρέπει να υλοποιηθεί στη βάση του αμοιβαίου συμβιβασμού μεταξύ:

- της ακριβούς προσέγγισης των εμπειρικών δεδομένων; και
- του περιορισμού της χωρητικότητας (VC dimension) του χώρου των συναρτήσεων (μηχανές μάθησης) οι οποίες χρησιμοποιούνται για την ελαχιστοποίηση του εμπειρικού ρίσκου.

Υπόδειγμα της Μεταγωγικής Συμπερασματολογίας

- Οι Vapnik και Chervonenkis διατυπώνοντας τις αρχές της Ελαχιστοποίησης του Εμπειρικού και του Δομικού Ρίσκου επέκτειναν το κλασικό Μοντέλο της Πρόγνωσης στο λεγόμενο Υπόδειγμα της Μεταγωγικής Συμπερασματολογίας.
- Σκοπός του Μεταγωγικού Υποδείγματος η εκτίμηση των τιμών μιας άγνωστης συνάρτησης πρόγνωσης σε δοσμένα σημεία ενδιαφέροντος παρά σε ολόκληρο το πεδίο ορισμού της.
- Το σκεπτικό πάνω στο οποίο βασίζεται η συγκεκριμένη προσέγγιση βασίζεται στη δυνατότητα απόκτησης ακριβέστερων λύσεων μέσω της επίλυσης απλούστερων προβλημάτων.

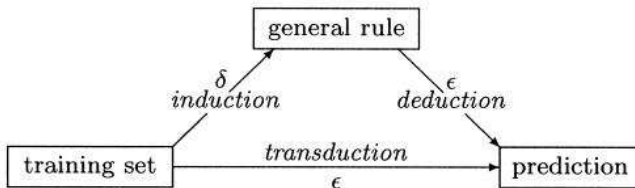


Figure: Μοντέλα Συμπερασματολογίας

Το Πρόβλημα της Ταξικής Ανισορροπίας

Definition (Πρόβλημα της Ταξικής Ανισορροπίας)

Το Πρόβλημα της Ταξικής Ανισορροπίας αφορά τις Μηχανές Επαγωγικής Μάθησης όταν παραβιάζεται η προϋπόθεση που αφορά την ομοιότητα των συναρτήσεων της εκ των πρότερων πιθανότητας, για τις κλάσεις που μετέχουν σε ένα πρόβλημα ταξινόμησης.

Το Πρόβλημα της Ταξικής Ανισορροπίας δυσχεραίνει την ταξινομητική ακρίβεια των μηχανών επαγωγικής μάθησης σε ένα ευρύ φάσμα δυνατών εφαρμογών όπως:

- ανίχνευση δόλιων τηλεφωνημάτων ή συναλλαγών
[Fawcett and Provost, 1996, Maes et al., 1993, Chan and Stolfo, 1998];
- αναγνώριση αναξιόπιστων πελατών σε τηλεπικοινωνιακές υπηρεσίες
[Ezawa et al., 1996];
- αναγνώριση κηλίδων πετρελαίου σε εικόνες που έχουν ληφθεί μέσω ραντάρ [Kubat et al., 1998];
- πρόγνωση αποτυχίας βιομηχανικών διαδικασιών ή τηλεπικοινωνιακού εξοπλισμού [Riddle et al., 1994, Weiss and Hirsh, 1998];
- διάγνωση σπάνιων ασθενειών [Laurikkala, 2001];
- ανάκτηση πληροφορίας [Lewis and Catlett, 1994];
- εκμάθηση προφοράς λέξεων [van den Bosch et al., 1997].

Η φύση του Προβλήματος της Ταξικής Ανισοροπίας I

Σύμφωνα με τους [Japkowicz and Stephen, 2002, Japkowicz, 2000]:

- Η εξαιρετικά ετεροβαρής φύση του Προβλήματος της Ταξικής Ανισοροπίας προκύπτει από το γεγονός ότι η κλάση ενδιαφέροντος [κλάση στόχος (**target class**), θετική κλάση (*positive class*), κλάση μειοψηφίας (**minority class**)] καταλαμβάνει μόνο ένα αμελητέο μέρος του συνολικού χώρου των προτύπων.
- Το Πρόβλημα της Ταξικής Ανισοροπίας εκδηλώνεται μέσω της μεροληψίας των διάφορων ταξινομητών υπέρ της κλάσης πλειοψηφίας (**majority class**) / αρνητικής κλάσης (**negative class**) / κλάσης των προτύπων εξαίρεσης (**outlier class**).
- Αυτό συμβαίνει διότι οι διάφοροι αλγόριθμοι μάθησης επιχειρούν να ελαττώσουν σφαιρικές ποσότητες όπως το συνολικό ποσοστό εσφαλμένων ταξινομήσεων χωρίς να λαμβάνουν υπόψη την κατανομή των δεδομένων.
- Συνεπώς τα πρότυπα που προέρχονται από την κατακλύζουσα κλάση ταξινομούνται σωστά ενώ τα πρότυπα που προέρχονται από την κλάση μειοψηφίας τείνουν να ταξινομούνται εσφαλμένα.
- Το συγκεκριμένο φαινόμενο επιδεινώνεται όταν αυξάνεται ο βαθμός της ανισοροπίας μεταξύ των κλάσεων η βαθμός πολυπλοκότητας της έννοιας που αναπαριστά η κλάση ενδιαφέροντος.

Η Φύση του Προβλήματος της Ταξικής Ανισοροπίας II

- Ωστόσο, το πρόβλημα αμβλύνεται όταν το μέγεθος του συνόλου εκπαίδευσης είναι επαρκώς μεγάλο η όταν το σύνολο των δεδομένων μεταξύ της αρνητικής και της θετικής είναι γραμμικώς διαχωρίσιμα.

Σύμφωνα με τους [Weiss and Provost, 2001, Weiss and Provost, 2002]:

- Το Πρόβλημα της Ταξικής Ανισοροπίας οφείλεται στην εσφαλμένη εκτίμηση της εκ των πρότερων (a priori) πιθανότητας εμφάνισης των προτύπων τόσο για την αρνητική όσο και για την θετική κλάση.
- Συνεπώς, η εκτίμηση της εκ των υστέρων πιθανότητας (a posteriori) εμφάνισης των προτύπων των δύο κλάσεων θα είναι και αυτή εσφαλμένη.
- Επομένως, η εσφαλμένα αυξημένη εκτίμηση της εκ των πρότερων πιθανότητας εμφάνισης μιας κλάσης, οδηγεί στην εσφαλμένα αυξημένη εκτίμηση της εκ των υστέρων πιθανότητας εμφάνισής της μετατοπίζοντας το κατώφλι ταξινόμησης προς το μέρος της κλάσης πλειοψηφίας.

Σύμφωνα με τους [Jo and Japkowicz, 2004, Stroulia and Matwin, 2001]:

- Το Πρόβλημα της Ταξικής Ανισοροπίας συνδέεται με την διαμέριση των προτύπων της κλάσης ενδιαφέροντος σε υποομάδες μικρότερου μεγέθους (sub-clusters / small disjuncts), οι οποίες ενδέχεται να αντανακλούν την υποδιαίρεση της αρχικής έννοιας στόχος (target concept) σε ένα σύνολο υπο-εννοιών (sub-concepts).

Ορισμοί

- Ο όρος ανοσολογία αναφέρεται στη μελέτη των αμυντικών μηχανισμών των σπονδυλωτών οργανισμών που παρέχουν αντίσταση ενάντια στις διάφορες ασθένειες.
- Το Ανοσοποιητικό Σύστημα (ΑΣ) είναι αυτό που φέρει την ευθύνη της προστασίας ενός οργανισμού ενάντια στις επιθέσεις που πραγματοποιούνται από τους διάφορους εξωγενείς μικροοργανισμούς.
- Το ΑΣ περιλαμβάνει αρκετούς αμυντικούς μηχανισμούς οι οποίοι δραστηριοποιούνται σε διαφορετικά επίπεδα, μερικοί εκ των οποίων είναι εφεδρικοί.
- Το ΑΣ παρουσιάζει τα χαρακτηριστικά της *μάθησης* και της *μνήμης*.
- Με τον όρο *παθογόνο* (**pathogen**) αναφερόμαστε σε οτιδήποτε μπορεί να προκαλέσει μια ασθένεια.
- Ο όρος *αντιγόνο* (**antigen**) αναφέρεται σε οποιοδήποτε μόριο μπορεί να διεγείρει το ΑΣ (οποιοδήποτε μόριο αναγνωρίζεται από το ΑΣ).

Δομή Ανοσοποιητικού Συστήματος

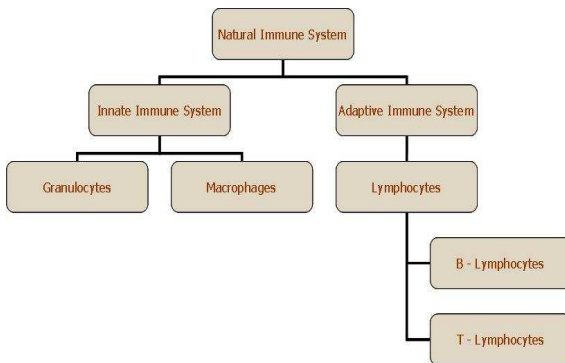


Figure: Immune System

Το Έμφυτο ΑΣ είναι άμεσα διαθέσιμο για μάχη ενώ το Προσαρμόσιμο ΑΣ επικεντρώνεται στην παραγωγή αντισωμάτων ενάντια σε καθορισμένους ιογενείς παράγοντες

Αναγνώριση Προτύπων I

- Το Προσαρμόσιμο Ανοσοποιητικό Σύστημα (ΠΑΣ) των σπονδυλωτών αποτελείται από μια μεγάλη ποικιλία μορίων, κυττάρων και οργάνων που είναι διεσπαρμένα σε όλο το σώμα.
- Η κεντρική αποστολή του ΠΑΣ συνίσταται στην ενδελεχή εξέταση του οργανισμού υπό την αναζήτηση δυσλειτουργούντων κυττάρων του ίδιου του οργανισμού (καρκινικά κύτταρα) καθώς και εξωγενή παθογόνα στοιχεία (ιοί και βακτήρια).
- Τα κύτταρα εκείνα που αρχικά ανήκουν στον οργανισμό μας και είναι μη - επιβλαβή για τη λειτουργία του χαρακτηρίζονται από τον όρο εαυτός (self ή self - antigens).
- Τα παθογόνα στοιχεία χαρακτηρίζονται από τον όρο μη - εαυτός (non - self ή non self - antigens).
- Η προστατευτική ιδιότητα του ΠΑΣ βασίζεται στην ικανότητά του να μαθαίνει να διακρίνει τα κύτταρα που ανήκουν στον οργανισμό (self cells) από αυτά που δεν ανήκουν (non self cells).
- Η διαδικασία αυτή ονομάζεται διάκριση εαυτού από τον μη - εαυτό (self / non - self discrimination).
- Τα κύτταρα εκείνα που αναγνωρίζονται ως κύτταρα εαυτού δεν προάγουν την αντίδραση του ΑΣ και για το λόγο αυτό το ΑΣ χαρακτηρίζεται ανεκτικό (tolerant) ως προς αυτά.

Αναγνώριση Προτύπων II

- Για εκείνα τα κύτταρα που δεν αναγνωρίζονται η αντίδραση του ΑΣ οδηγεί στον αφανισμό τους.
- Από την προοπτική της Αναγνώρισης Προτύπων το πιο ενδιαφέρον χαρακτηριστικό του ΑΣ είναι η ύπαρξη μορίων που επιτελούν το ρόλο του υποδοχέα στην επιφάνεια των ανοσοποιητικών κυττάρων και διαθέτουν την ικανότητα αναγνώρισης μιας σχεδόν ανεξάντλητης γκάμας αντιγονικών προτύπων.

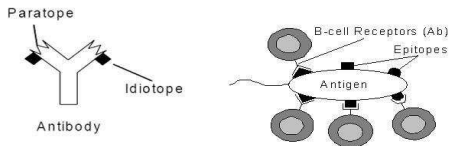


Figure: Αντισώματα και Αντιγόνα

Αναγνώριση Προτύπων III

- Τα αντισώματα αποτελούν κυτταρικούς υποδοχείς σχήματος Y που είναι προσκολλημένα στην επιφάνεια των B - Λεμφοκυττάρων με πρωτεύοντα ρόλο την αναγνώριση αντιγόνων μέσω της πρόσδεσής τους με αυτά.
- Η δύναμη και η ιδιαιτερότητα της αλληλεπίδρασης Αντιγόνου - Αντισώματος εκφράζει το μέτρο της αναγνώρισης του αντιγόνου από το αντίσωμα και ονομάζεται έλξη (affinity) ή επίπεδο συμπληρωματικότητας.
- Η αναγνώριση ενός αντιγόνου από ένα αντίσωμα βασίζεται στην συμπληρωματικότητα της τρισδιάστατης χωρικής δομής τους αλλά και της ηλεκτροχημικής τους συμπεριφοράς.
- Τα μόρια των αντισωμάτων αναγνωρίζουν ένα τμήμα του αντιγόνου που ονομάζεται *επίτοπο (epitope)*. Το τμήμα αυτό του αντιγόνου υπάρχει και στα αντισώματα και φέρει την ονομασία *ιδιότοπο(idiotope)*.
- Σε αντίθεση με τα B - Λεμφοκύτταρα τα αντιγόνα διαθέτουν διαφορετικούς τύπους επιτόπων με αποτέλεσμα να είναι δυνατή η αναγνώρισή τους από διαφορετικούς τύπους αντισωμάτων.
- Το τμήμα του αντισώματος που είναι υπεύθυνο για την σύνδεση (ταίριασμα) αντιγόνου - αντισώματος ονομάζεται παράτοπο (paratope) ή αλλιώς μεταβλητή περιοχή (variable region) καθώς εξαιτίας της μεταβλητότητάς του μπορεί να τροποποιήσει το σχήμα του για να επιτύχει μεγαλύτερο βαθμό συμπληρωματικότητας προς το επίτοπο του αντιγόνου.

Αρνητική Επιλογή I

- Ένα ζήτημα ουσιαστικής σημασίας που προκύπτει είναι το πως εξασφαλίζεται η δυνατότητα των T - Λεμφοκυττάρων να επιτελούν τη διάκριση του εαυτού από τον μη - εαυτό.
- Θα πρέπει να υπάρχει κάποιος μηχανισμός ο οποίος να εμποδίζει τα T - Λεμφοκύτταρα να αντιδρούν με τα κύτταρα του ίδιου του οργανισμού. Η κατάρρευση αυτού του μηχανισμού μάλιστα οδηγεί στην εμφάνιση των αυτοάνοσων ασθενειών.
- Ο μηχανισμός αυτός ονομάζεται ως *Διαδικασία της Αρνητικής Επιλογής (Negative Selection Process)* η οποία επιτρέπει την επιβίωση μόνο εκείνων των T - Λεμφοκυττάρων που δεν αναγνωρίζουν τα κύτταρα εαυτού.
- Τα T - Λεμφοκύτταρα παράγονται στον μυελό των οστών αλλά μετά την παραγωγή τους μεταναστεύουν στον *Θύμο αδένα (Thymus gland)*. Εκεί θα υποστούν μια διαδικασία ωρίμανσης μετά τη λήξη της οποίας και μόνο είναι επιτρεπτή η συμμετοχή τους στις ανοσοποιητικές αντιδράσεις του οργανισμού.
- Τα T - Λεμφοκύτταρα κατά τη διάρκεια της ωρίμανσης τους εκτίθενται σε πρωτεΐνες εαυτού (self proteins). Εκείνα τα T - Λεμφοκύτταρα που έχουν την ιδιότητα να αντιδρούν προς τις πρωτεΐνες αυτές εξαλείφονται από το ρεπερτόριο των T - Λεμφοκυττάρων και απομένουν τα υπόλοιπα που παραμένουν αδρανή.

Αρνητική Επιλογή II

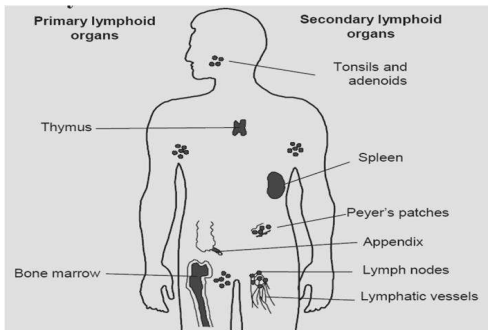


Figure: Αντισώματα και Αντιγόνα

- Αυτά τα T - Λεμφοκύτταρα χαρακτηρίζονται ως ανεκτικά ως προς τον εαυτό (*self tolerant*) και κατά συνέπεια είναι αυτά που ανοσοποιητικώς είναι ικανά (*immune competent*).

Αρνητική Επιλογή III

- Από υπολογιστική άποψη η διάκριση του εαυτού από τον μη - εαυτό είναι ιδιαίτερα σημαντική για τη δημιουργία τεχνητών συστημάτων τα οποία θα υλοποιούν τη διάκριση μεταξύ των καταστάσεων εαυτού από τις καταστάσεις μη εαυτού.
- Ως καταστάσεις εαυτού μπορούν να θεωρηθούν η κανονική συμπεριφορά ενός συστήματος ή η κανονική δικτυακή κίνηση μεταξύ δύο υπολογιστών ή το σύνολο των προτύπων που ανήκουν στην κλάση μειοψηφίας σε ένα μονοταξικό πρόβλημα ταξινόμησης.
- Η διαδικασία της αρνητικής επιλογής ενέπνευσε την ανάπτυξη του Αλγορίθμου της Αρνητικής Επιλογής (*Negative Selection Algorithm*).
- Από υπολογιστική άποψη ο Αλγόριθμος της Αρνητικής Επιλογής αποτελεί ένα εναλλακτικό υπόδειγμα αναγνώρισης προτύπων καθώς συνίσταται στην αποθήκευση πληροφοριών για το συμπληρωματικό σύνολο των προτύπων που πρόκειται να αναγνωρισθούν.

Θεωρία Ανοσοποιητικού Δικτύου I

- Το ανοσοποιητικό σύστημα μπορεί να θεωρηθεί ως ένα συντονισμένο δίκτυο κυττάρων και μορίων τα οποία αναγνωρίζουν το ένα το άλλο ακόμα και κατά την απουσία εξωγενών αντιγονικών παραγόντων (Idiotypic Network Theory [Jerne, 1974]).
- Η Θ.Α.Δ. βασίζεται στην προϋπόθεση ότι οι κυτταρικοί υποδοχείς των Β-Λεμφοκυττάρων φέρουν κάποια τμήματα (*ιδιότοπα*) τα οποία είναι αναγνωρίσιμα από άλλα ελεύθερα αντισώματα.
- Η Θ.Α.Δ προτείνει ότι το Α.Σ. παρουσιάζει μια δυναμική συμπεριφορά η οποία είναι εγγενώς συνδεδεμένη με την δυνατότητα των κύριων διαμεσολαβητών του τόσο να αναγνωρίζουν όσο και να αναγνωρίζονται.
- Βασιζόμενοι στη Θ.Α.Δ οι [Farmer et al., 1986] πρότειναν μια διαφορική εξίσωση η οποία ποσοτικοποιούσε την αρχική περιγραφική διατύπωση του Jerne.

$$\frac{dX_i(t)}{dt} = c \left[\sum_{j=1}^N d_{ji} X_i(t) X_j(t) - k_1 \sum_{j=1}^N d_{ij} X_i(t) X_j(t) + \sum_{j=1}^M d_{ji} X_i(t) Y_j(t) \right] - k_2 X_i(t) \quad (7)$$

$$\frac{dY_i(t)}{dt} = -k_3 \sum_{j=1}^M d_{ij} X_j(t) Y_i(t) \quad (8)$$

Θεωρία Ανοσοποιητικού Δικτύου II

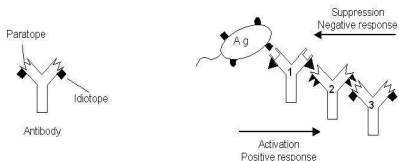


Figure: Θεωρία Ανοσοποιητικού Δικτύου.

$$d_{ij} = D(\mathbf{e}_i, \mathbf{p}_j) \quad (9)$$

- Η αναγνώριση ενός αντιγόνου από ένα αντίσωμα συνεπάγεται τη διέγερση του Α.Σ η οποία περιλαμβάνει τον πολλαπλασιασμό και την ενεργοποίηση των ανοσοποιητικών κυττάρων καθώς και την επακόλουθη έκκριση αντισωμάτων.
- Αντίθετα η αλληλοαναγνώριση δύο αντισωμάτων συνεπάγεται μια ανασταλτική απόκριση του Α.Σ. η οποία περιλαμβάνει την καταστολή του Α.Σ. μέσω της εξάλειψης εκείνων των αντισωμάτων που αναγνωρίζουν το ένα το άλλο.

Θεωρία Ανοσοποιητικού Δικτύου III

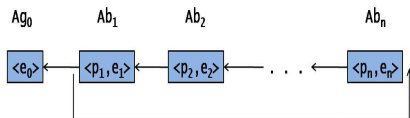


Figure: Ανοσολογικός Κύκλος

Σύμφωνα με το μοντέλο των [Farmer et al., 1986]:

- 1 η μαθησιακή ικανότητα αποτελεί ένα αναδυόμενο χαρακτηριστικό του Α.Σ. το οποίο σχετίζεται στενά με τους δυναμικούς κανόνες που διέπουν την συμπεριφορά του.
- 2 Η λειτουργία της μνήμης σχετίζεται με τη δημιουργία Ανοσολογικών Κύκλων.

Η δομή των περισσότερων δικτυακών μοντέλων για τρόπο λειτουργίας του Α.Σ. ακολουθεί το γενικότερο μοντέλο του [Perelson, 1989] που δίνεται από την παρακάτω σχέση:

$$\begin{array}{r} \text{Rate of} \\ \text{population} \\ \text{variation} \end{array} = \begin{array}{r} \text{Network} \\ \text{stimulation} \end{array} - \begin{array}{r} \text{Network} \\ \text{suppression} \end{array} + \begin{array}{r} \text{Influx of} \\ \text{new} \\ \text{elements} \end{array} - \begin{array}{r} \text{Death of} \\ \text{unstimulated} \\ \text{elements} \end{array} \quad (10)$$

Θεωρία Επιλεκτικής Κλωνοποίησης και Ωρίμανσης του Μέτρου Συμπληρωματικότητας I

Τα κυριότερα χαρακτηριστικά της Θεωρίας της Επιλεκτικής Κλωνοποίησης είναι:

- 1 Η αναπαραγωγή και η διαφοροποίηση εκείνων των αντισωμάτων που παρουσίασαν την υψηλότερη διέγερση κατά την παρουσία κάποιου αντιγόνου.
- 2 Ωρίμανση του μέτρου συμπληρωματικότητάς τους (affinity maturation) μέσω μιας διαδικασίας επιταχυνόμενης σωματικής μετάλλαξης.
- 3 Η ανοσοποιητική αντίδραση του οργανισμού δεν ξεκινά κάθε φορά από την αρχή. Η ταχύτητα και η ακρίβειά της αυξάνονται μετά την έκθεση σε κάθε αντιγόνο (Μαθηση Μέσω Παραδειγμάτων).
- 4 Αποδίδει στο Α.Σ. ένα εγγενές σχήμα ενισχυόμενης μάθησης όπου το σύστημα βελτιώνει συνεχώς την ικανότητά του να επιτελεί το σκοπό του.

Θεωρία Επιλεκτικής Κλωνοποίησης και Ωρίμανσης του Μέτρου Συμπληρωματικότητας II

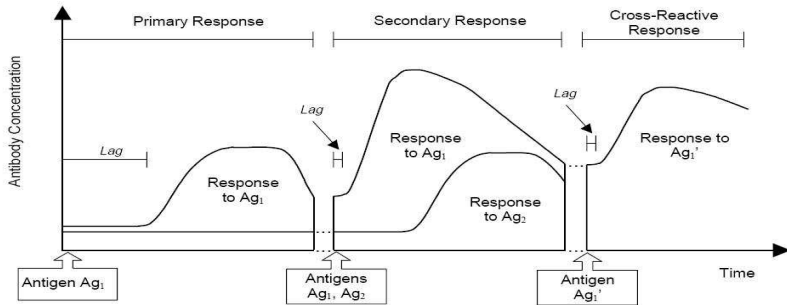


Figure: Primary, secondary and cross-reactive immune responses.

Θεωρία Επιλεκτικής Κλωνοποίησης και Ωρίμανσης του Μέτρου Συμπληρωματικότητας III

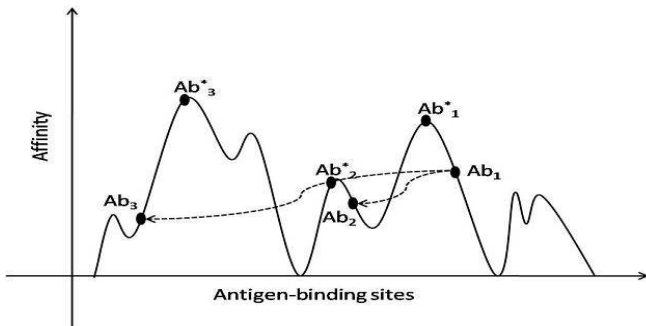


Figure: Receptor editing process.

Θεωρία Επιλεκτικής Κλωνοποίησης και Ωρίμανσης του Μέτρου Συμπληρωματικότητας IV

Η ανοσοποιητική μνήμη είναι είναι συσχετιστική. Τα αντισώματα που προσαρμόστηκαν σε ένα συγκεκριμένο τύπο αντιγόνου θα παρουσιάσουν μια ταχύτερη και αποτελεσματικότερη απόκριση όχι μόνο σε μια δεύτερη έκθεση στο ίδιο αντιγόνο αλλά και σε κάθε αντιγόνο που σχετίζεται δομικά με το πρώτο.

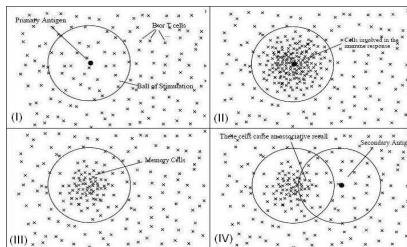


Figure: Primary, secondary and cross-reactive immune responses.

- Το φαινόμενο αυτό ονομάζεται ανοσοποιητική διάδραση (immunological cross reaction) ή διαδραστική απόκριση (cross reactive response).
- Η συσχετιστική μνήμη του ΑΣ αποτελεί στην πραγματικότητα έναν μηχανισμό γενίκευσης όπως αναφέρεται σε άλλα πεδία της αναγνώρισης προτύπων όπως αυτός των Μηχανών Διανυσμάτων Υποστήριξης.

Ορισμός και Εφαρμογές Τεχνητών Ανοσοποιητικών Συστημάτων I

Definition (Τεχνητά Ανοσοποιητικά Συστήματα)

Τα Τεχνητά Ανοσοποιητικά Συστήματα (ΤΑΣ) είναι δυνατό να οριστούν ως αφηρημένα (ή μεταφορικά) υπολογιστικά συστήματα τα οποία αναπτύχθηκαν βασισμένα σε αρχές, θεωρίες και συστατικά τα οποία εξήχθησαν από τη μελέτη του ανοσοποιητικού συστήματος.

Από υπολογιστική άποψη μερικές από τις δυνατές εφαρμογές των ΤΑΣ είναι:

- Ταξινόμηση και Ομαδοποίηση Δεδομένων.
- Ανίχνευση Ανωμαλιών και Δικτυακών Επιθέσεων.
- Βελτιστοποίηση.
- Συστήματα Αυτόματου Έλεγχου.
- Βιοπληροφορική.
- Ανάκτηση Πληροφορίας και Εξόρυξη Γνώσης.
- Μοντελοποίηση Χρηστών και Υλοποίηση Συστημάτων Σύστασης.
- Επεξεργασία Εικόνας.

Ορισμός και Εφαρμογές Τεχνητών Ανοσοποιητικών Συστημάτων II

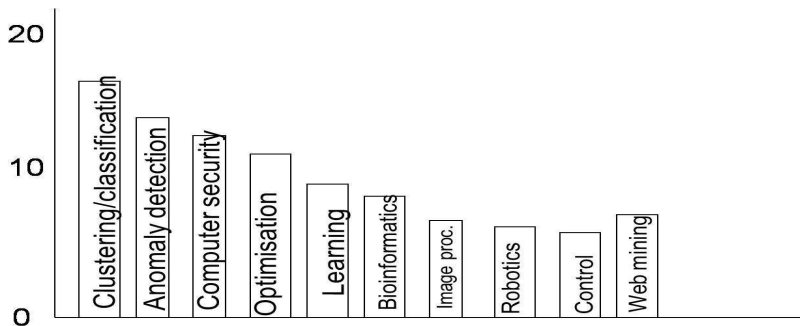


Figure: Εφαρμογές ΤΑΣ

Γενικό Πλαίσιο για την Ανάπτυξη Τεχνητών Ανοσοποιητικών Συστημάτων

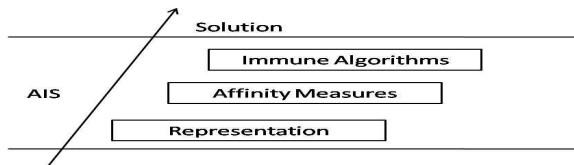


Figure: Layered Framework for AIS.

Το Γενικό Πλαίσιο Εργασίας που υιοθετήθηκε για την υλοποίηση των προτεινόμενων ΤΑΣ περιλαμβάνει:

- Μια αναπαράσταση των θεμελιωδών συστατικών του Ανοσοποιητικού Συστήματος (**Representation**).
- Ένα σύνολο μηχανισμών για την αποτίμηση των αλληλεπιδράσεων των ξεχωριστών συστατικών του συστήματος αλλά και της αλληλεπίδρασης μεταξύ αυτών και του περιβάλλοντος (Μέτρα Συμπληρωματικότητας / Εγγύτητας / **Affinity Measures**).
- Ένα σύνολο διαδικασιών προσαρμογής (*Ανοσοποιητικών Αλγορίθμων*) οι οποίες θα καθορίζουν την δυναμική συμπεριφορά των ΤΑΣ (**Immune Algorithms**).

Μοντέλο Χώρου Σχημάτων I

- Το μοντέλο του χώρου των σχημάτων στοχεύει στην ποσοτική περιγραφή των αλληλεπιδράσεων μεταξύ αντιγόνων και αντισωμάτων.
- Το σύνολο των χαρακτηριστικών που περιγράφουν ένα μόριο στο σύμπαν των στερεοχημικών σχημάτων ονομάζεται γενικευμένο σχήμα του μορίου και είναι αυτό που καθορίζει το μέτρο της *διαμοριακής έλξης* ή *συμπληρωματικότητας*.
- Το σύμπαν αυτό των στερεοχημικών σχημάτων μπορεί να αναπαρασταθεί ως ένας πολυδιάστατος διανυσματικός χώρος - σχημάτων \mathbb{S} .
- Είναι πολύ σημαντικό να σημειωθεί πως ο θεωρούμενος χώρος - σχημάτων αποτελεί μια αφαίρεση της πραγματικότητας η οποία όμως είναι αρκετά αποδοτική για τη δημιουργία και αποτίμηση υπολογιστικών προσεγγίσεων που βασίζονται στις αρχές του ΑΣ.

Definition

Το γενικευμένο σχήμα ενός μορίου είτε πρόκειται για ένα αντισώμα (**Ab**) είτε πρόκειται έναν αντιγονικό προσδιοριστή (**Ag**) μπορεί να αναπαρασταθεί ως ένα διάνυσμα συντεταγμένων (L παραμέτρων σχήματος) $\mathbf{m} = \langle m_1, \dots, m_L \rangle$ με $m_i \in \Sigma$. Επομένως, ο χώρος των σχημάτων θα είναι ο $\mathbb{S} = \Sigma^L$

Μοντέλο Χώρου Σχημάτων II

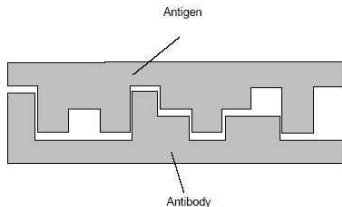


Figure: Αναγνώριση Αντιγόνου - Αντισώματος βάση της συμπληρωματικότητας της δομής τους.

Ανάλογα με τη μορφή του συνόλου Σ :

- ο χώρος-σχημάτων μπορεί να είναι συνεχής (πραγματικός) (Real Valued Shape-Space, $\Sigma = \mathbb{R}$).
- ο χώρος-σχημάτων μπορεί να είναι διακριτός (συμβολικός) (Hamming Shape-Space, $\Sigma = \{0, 1\}$)

Μέτρα Συμπληρωματικότητας - Εγγύτητας

Definition (Μέτρο Συμπληρωματικότητας - Εγγύτητας)

Ως Μέτρο Συμπληρωματικότητας - Εγγύτητας ορίζεται μία απεικόνιση της μορφής:

$$D : \mathbb{S} \times \mathbb{S} \rightarrow \mathbb{R}^+ \quad (11)$$

Η οποία ποσοτικοποιεί την αλληλεπίδραση μεταξύ δύο αντισωμάτων ή μεταξύ ενός αντισώματος και ενός αντιγόνου.

Το μέτρο συμπληρωματικότητας - εγγύτητας που χρησιμοποιήθηκε για τις ανάγκες της παρούσας διατριβής είναι η Κανονικοποιημένη Ευκλείδεια Απόσταση.

Definition

Η Κανονικοποιημένη Ευκλείδεια Απόσταση μεταξύ ενός αντισώματος $\mathbf{Ab} \in \mathbb{S}$ και ενός αντιγόνου $\mathbf{Ag} \in \mathbb{S}$ δίνεται από τη σχέση:

$$D(\mathbf{Ab}', \mathbf{Ag}') = \frac{1}{\sqrt{L}} \|\mathbf{Ab}' - \mathbf{Ag}'\| = \frac{1}{\sqrt{L}} \sqrt{\sum_{i=1}^L (Ab'_i - Ag'_i)^2} \quad (12)$$

όπου

$$Ab'_i = \frac{Ab_i - A_{i,min}}{A_{i,max} - A_{i,min}} \quad (13)$$

$$Ag'_i = \frac{Ag_i - A_{i,min}}{A_{i,max} - A_{i,min}} \quad (14)$$

ετσι ώστε:

$$0 \leq D(\mathbf{Ab}', \mathbf{Ag}') \leq 1 \quad (15)$$

Γεωμετρική Θεώρηση του Χώρου των Σχημάτων

- Σύμφωνα με το Μοντέλο του Χώρου των Σχημάτων κάθε αντίσωμα έχει τη δυνατότητα πρόσδεσης σε ένα σύνολο αντιγονικών προσδιοριστών η διανυσματική αναπαράσταση των οποίων βρίσκεται σε μια συγκεκριμένη περιοχή V_ϵ του χώρου \mathbb{S} η οποία προσδιορίζεται με βάση το προκαθορισμένο μέτρο εγγύτητας.
- Κάθε αντίσωμα τοποθετείται στο κέντρο μιας σφαιρικής περιοχής ακτίνας ϵ του χώρου \mathbb{S} έχοντας τη δυνατότητα αναγνώρισης κάθε αντιγονικού προσδιοριστή που βρίσκεται μέσα στα όρια αυτής της σφαίρας

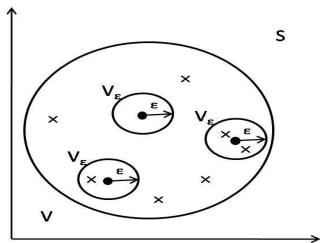


Figure: Μοντελο Χωρου Σχηματων

Σύνολο Δεδομένων I

- Το σύνολο των δεδομένων που χρησιμοποιήθηκε για τον πειραματικό έλεγχο των αλγορίθμων προέρχεται απο μια συλλογή 10 κλάσεων δυτικής μουσικής.
- Κάθε μουσικολογική κλάση περιλαμβάνει 100 στιγμιότυπα.

Table: Κλάσεις Δυτικής Μουσικής

| Class ID | Label |
|----------|-----------|
| 1 | Blues |
| 2 | Classical |
| 3 | Country |
| 4 | Disco |
| 5 | Hip-Hop |
| 6 | Jazz |
| 7 | Metal |
| 8 | Pop |
| 9 | Reggae |
| 10 | Rock |

Σύνολο Δεδομένων II

- Κάθε στιγμυότυπο είναι ένα μουσικό αρχείο ποιότητας CD το οποίο περιέχει 44.100 16bit δείγματα / δευτερόλεπτο, διάρκειας 30 δευτερολέπτων.
- Ο αρχικός χώρος των δεδομένων: $\mathbf{X} = \{0, \dots, 65535\}^{30 \times 44100}$.
- Πρωταρχικός στόχος είναι η μείωση της διάστασης του αρχικού χώρου μέσω της ψηφιακής επεξεργασίας σήματος.
- Χρησιμοποιήθηκε το λογισμικό MARSYAS προκειμένου απο κάθε μουσικό αρχείο να εξαχθεί ένα 30-διάστατο πραγματικό διάνυσμα χαρακτηριστικών: $\mathbf{X} = \mathbb{R}^{30}$.

Εξαγωγή Χαρακτηριστικών I

- Η προεπεξεργασία (preprocessing) περιλαμβάνει τον μετασχηματισμό των αρχικών δεδομένων (raw data) σε μια πιο εκλεπτυσμένη μορφή δεδομένων σε ένα υψηλότερο επίπεδο αφάιρεσης στο οποίο έχουμε μόνο την απαραίτητη και χρήσιμη πληροφορία.
- Αυτό επιτυγχάνεται μέσω μιας διαδικασίας εξαγωγής χαρακτηριστικών.
- Τα χαρακτηριστικά αντιπροσωπεύουν όλη τη σημαντική πληροφορία που περιέχεται στα αρχικά δεδομένα.

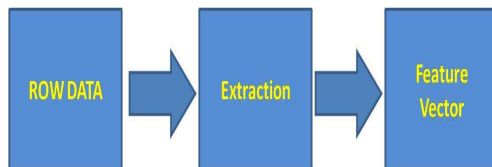


Figure: Διαδικασία Εξαγωγής Χαρακτηριστικών.

Εξαγωγή Χαρακτηριστικών II

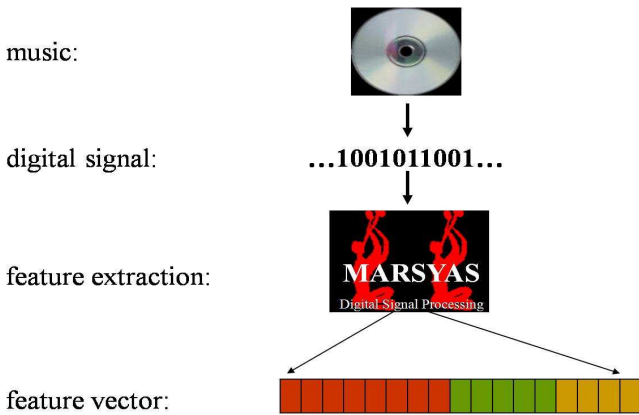


Figure: Διαδικασία Εξαγωγής Χαρακτηριστικών.

Εξαγωγή Χαρακτηριστικών III



MARSYAS

Figure: Εξαγωγή 30 χαρακτηριστικών περιεχομένου από τα μουσικά αρχεία.

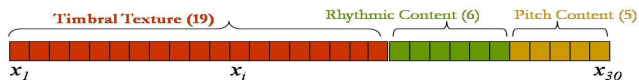


Figure: Διάγραμμα Χαρακτηριστικών 30 Διαστάσεων.

Διατύπωση Προβλήματος

Ανάπτυξη Αλγορίθμου Εκπαίδευσης ενός Τ.Α.Δ. (Artificial Immune Network) για την ανάλυση μουσικών δεδομένων με επιμέρους στόχους:

- 1 Αναγνώριση των εγγενών κλάσεων των δεδομένων.
- 2 Ποιοτική και ποσοτική περιγραφή των αναγνωριζόμενων κλάσεων των δεδομένων.
- 3 Απεικόνιση της χωρικής κατανομής των δεδομένων μέσα στις αναγνωριζόμενες κλάσεις.
- 4 Αποκάλυψη πλεονάζουσας πληροφορίας μέσα στο σύνολο των μουσικών δεδομένων.

Αναδιατύπωση Προβλήματος

Αναδιατύπωση Προβλήματος μέσα στο πλαίσιο των Τ.Α.Δ.

- 1 Το σύνολο των μουσικών διανυσμάτων χαρακτηριστικών μπορεί να θεωρηθεί ως ο δοσμένος προς αναγνώριση αντιγονικός πληθυσμός.
- 2 Τα παραγόμενα αντισώματα μπορούν να θεωρηθούν ως σημεία ενός 30-διάστατου χώρου χαρακτηριστικών τα οποία λειτουργούν ως αντιπρόσωποι ενός συγκεκριμένου τμήματος του αντιγονικού πληθυσμού.
- 3 Ο αλγόριθμος εκπαίδευσης του Τ.Α.Δ. στοχεύει στην παραγωγή ενός συνόλου αντιπροσώπων / αντισωμάτων τα οποία θα έχουν εκπαιδευθεί ώστε να αναγνωρίζουν τον δοσμένο αντιγονικό πληθυσμό παρέχοντας όμως μια εναλλακτική περισσότερο συμπαγή χωρική αναπαράστασή των αρχικών αντιγονικών προτύπων.
- 4 Στόχος είναι μια περισσότερο αποσαφηνισμένη ποσοτική και ποσοτική αναπαράσταση των εγγενών κλάσεων που υπάρχουν στα αρχικά μουσικά δεδομένα.

Τεχνητό Ανοσοποιητικό Δίκτυο

Definition

Ένα Τεχνητό Ανοσοποιητικό Δίκτυο μπορεί να οριστεί μέσω ενός γραφήματος κόμβων στο οποίο ο κάθε κόμβος αναπαριστά ένα ξεχωριστό αντισωμα μνήμης. Σε κάθε ακμή του γραφήματος αντιστοιχεί ένα συγκεκριμένο βάρος (ή ισχύς σύνδεσης) το οποίο ποσοτικοποιείται απο το μέτρο συμπληρωματικότητας / εγγύτητας που έχει υιοθετηθεί. Στη γενική περίπτωση το γράφημα που προκύπτει δεν είναι πλήρως συνδεδεμένο. Η αναπαράσταση ενός Τ.Α.Δ γίνεται μέσω του ελάχιστου δένδρου ζεύξης των παραχθέντων αντισωμάτων μνημης.

- 1 Η αναπαράσταση ενός Τ.Α.Δ γίνεται μέσω του ελάχιστου δένδρου ζεύξης των παραχθέντων αντισωμάτων μνημης.
- 2 Οι συστάδες των αντισωμάτων μνήμης που δημιουργούνται εντός του ανοσοποιητικού δικτύου μπορούν να θεωρηθούν ως εσωτερικές εικόνες των ομάδων που ανιχνεύτηκαν στο αρχικό σύνολο δεδομένων.
- 3 Γενικά ο συνολικός αριθμός των αντισωμάτων είναι μεγαλύτερος από τον αριθμό των κλάσεων που μπορούν να αναγνωρισθούν για το αρχικό σύνολο των δεδομένων αλλά πολύ μικρότερος από το σύνολο των αρχικών δεδομένων χαρακτηρίζοντας έτσι μια αρχιτεκτονική για συμπίεση δεδομένων.

Επισκόπηση Αλγορίθμου Εκπαίδευσης Τ.Α.Δ. Ι

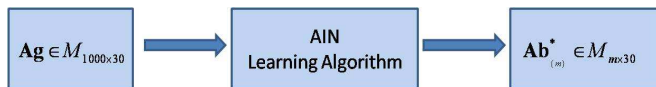


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Δ.

Ο αλγόριθμος εκπαίδευσης του Τ.Α.Δ στοχεύει στην παραγωγή ενός συνόλου αντισωμάτων μνήμης $\mathbf{Ab}_{\{m\}}^*$ με $m \ll 1000$ τέτοια ώστε:

$$\mathbf{Ab}_{\{m\}}^* = \arg \min_{\mathbf{Ab}_{\{m\}} \in \mathbb{S}^m} E(\mathbf{Ab}_{\{m\}}) \quad (16)$$

με

$$E(\mathbf{Ab}_{\{m\}}) = \frac{1}{m \cdot 1000} \sum_{i=1}^m \sum_{j=1}^{1000} D(\mathbf{Ab}_{\{m\}}^i, \mathbf{Ag}_j) \quad (17)$$

$$\text{Required Memory} = (M + N + N_c)L + N + N_c + [\zeta \cdot N_c] + m^2 \quad (18)$$

$$\text{Required Time} = O(Mm^2) \quad (19)$$

Επισκόπηση Αλγορίθμου Εκπαίδευσης Τ.Α.Δ. II

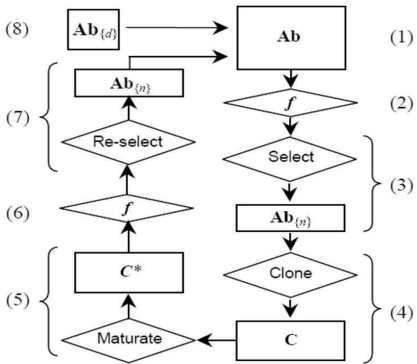


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Δ.

Παράμετροι Εισόδου Αλγορίθμου Εκπαίδευσης Τ.Α.Δ.

Ο αλγόριθμος εκπαίδευσης του Τ.Α.Δ για την ανάλυση των μουσικών δεδομένων δέχεται ως είσοδο τις παρακάτω παραμέτρους:

- N : Συνολικός αριθμός διαθέσιμων αντισωμάτων.
- n : Αριθμός των αντισωμάτων με το μεγαλύτερο μέτρο συμπληρωματικότητας / εγγύτητας προς το τρέχον αντιγονικό πρότυπο που επιλέγονται για κλωνοποίηση.
- ζ : Το ποσοστό των κλωνοποιημένων αντισωμάτων που θα εισαχθούν στο σύνολο των αντισωμάτων μνήμης.
- σ_d (Κριτήριο Απόπτωσης): Ελάχιστη επιτρεπτή απόσταση μεταξύ αντιγόνου - αντισώματος.
- σ_s (Κριτήριο Καταστολής): Ελάχιστη επιτρεπτή απόσταση μεταξύ δύο αντισωμάτων μνήμης.
- GEN : Αριθμός επαλήψεων.

Αλγόριθμος Εκπαίδευσης Τεχνητού Ανοσοποιητικού Δικτύου I

- 1 Αρχικοποίηση:** Δημιουργία ενός αρχικού πληθυσμού από τυχαία αντισώματα.
- 2 Παρουσίαση αντιγονικών προτύπων:** Για κάθε αντιγονικό πρότυπο:
 - 1 Επιλεκτική Κλωνοποίηση και Διαστολή:** Για κάθε διαθέσιμο αντισώμα του δικτύου προσδιορισμός του μέτρου συμπληρωματικότητάς του προς το τρέχων αντιγονικό πρότυπο. Επιλογή ενός υποσυνόλου αντισωμάτων υψηλής συμπληρωματικότητας προς αναπαραγωγή (κλωνοποίηση) ανάλογα προς το μέτρο της συμπληρωματικότητάς τους.
 - 2 Ωρίμανση του Μέτρου Συμπληρωματικότητας:** Μετάλλαξη του κάθε διαθέσιμου αντισώματος με ρυθμό αντιστρόφως ανάλογο προς το μέτρο της συμπληρωματικότητάς του προς το τρέχων αντιγονικό πρότυπο. Επανεπιλογή ενός αριθμού υψηλής συμπληρωματικότητας αντισωμάτων και τοποθέτησή τους σε ένα σύνολο κλώνων μνήμης.
 - 3 Μεταδυναμική:** Εξάλειψη όλων των αντισωμάτων μνήμης των οποίων το μέτρο της συμπληρωματικότητας προς το τρέχων αντιγονικό πρότυπο είναι μικρότερο από ένα προκαθορισμένο κατώφλι.
 - 4 Αλληλεπίδραση Κλώνων:** Προσδιορισμός των αλληλεπιδράσεων του δικτύου (μέτρου συμπληρωματικότητας) μεταξύ όλων των ζευγών των αντισωμάτων του συνόλου μνήμης.
 - 5 Καταστολή Κλώνων:** Εξάλειψη όλων των αντισωμάτων μνήμης των οποίων τα μέτρα συμπληρωματικότητας προς τα υπόλοιπα αντισώματα μνήμης έχουν τιμή παραπάνω από ένα προκαθορισμένο κατώφλι.
 - 6 Οικοδόμηση Δικτύου:** Ενσωμάτωση των διαθέσιμων αντισωμάτων του δικτύου με τα εναπομείναντα αντισώματα μνήμης του προηγούμενου βήματος

Αλγόριθμος Εκπαίδευσης Τεχνητού Ανοσοποιητικού Δικτύου II

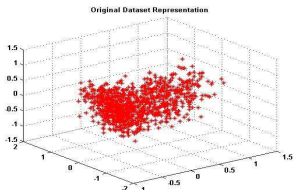
- 3 *Δικτυακές Αλληλεπιδράσεις*: Προσδιορισμός του μέτρου της συμπληρωματικότητας μεταξύ όλων των ζευγών των διαθέσιμων αντισωμάτων του δικτύου.
- 4 *Καταστολή Δικτύου*: Εξάλειψη όλων των διαθέσιμων αντισωμάτων του δικτύου οποίων τα μέτρα συμπληρωματικότητας προς τα υπόλοιπα αντισώματα έχουν τιμή παρακάτω από ένα προκαθορισμένο κατώφλι.
- 5 *Διαφοροποίηση*: Εισαγωγή στο δίκτυο ενός αριθμού νέων τυχαία παραγόμενων αντισωμάτων.
- 6 *Επανάληψη*: Επανάληψη των βημάτων 2 έως 5 μέχρι την πραγματοποίηση ενός προκαθορισμένου αριθμού επαναλήψεων.

Ανοσοποιητικές Αρχές Αλγορίθμου Εκπαίδευσης Τ.Α.Δ.

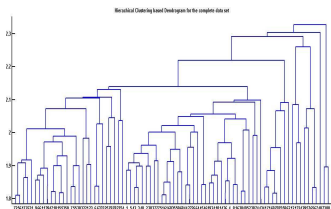
Οι βασικές αρχές των Θεωριών του Ανοσοποιητικού Δικτύου και της Επιλεκτικής Κλωνοποίησης που εφαρμόζονται κατά την ανάπτυξη του αλγορίθμου εκπαίδευσης του Τ.Α.Δ είναι:

- 1 Διατήρηση ενός συνόλου διαθέσιμων αντισωμάτων απο τα οποία θα προκύψουν τα αντισώματα μνήμης.
- 2 Διατήρηση ενός συγκεκριμένου συνόλου αντισωμάτων μνήμης.
- 3 Επιλογή και κλωνοποίηση των περισσότερο διεγερμένων αντισωμάτων.
- 4 Απαλοιφή των λιγότερο διεγερμένων αντισωμάτων.
- 5 Ωρίμανση του μέτρου συμπληρωματικότητας και επανεπιλογή των κλώνων αντιστρόφως ανάλογα προς το μέτρο της συμπληρωματικότητάς τους προς το τρέχον αντιγονικό πρότυπο.

Αποτελέσματα - Αρχική Χωρική Κατανομή Δεδομένων



(a) Original Music Dataset Representation in 3 Dimensions.

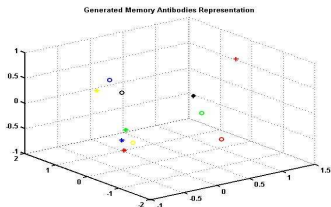


(b) Hierarchical Agglomerative Clustering-Based Dendrogram.

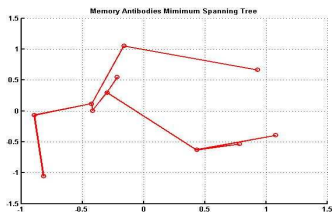
Παρουσιάζεται την 3-διάστατη χωρική κατανομή του δοσμένου προς αναγνώριση αντιγονικού πληθυσμού. Στόχος του αλγορίθμου εκπαίδευσης του Τ.Α.Δ είναι η παραγωγή ενός συνόλου αντισωμάτων μνήμης τα οποία θα αναδεικνύουν ποιοτικά και ποσοτικά το σύνολο των εγγενών κλάσεων στον αρχικό αντιγονικό πληθυσμό.

Παρουσιάζεται το δενδρόγραμμα που αντιστοιχεί στην ιεραρχική ομαδοποίηση των αρχικών δεδομένων με βάση τον αλγόριθμο της συγχωνευτικής ομαδοποίησης.

Αποτελέσματα - Χωρική Κατανομή Αντισωμάτων Μνήμης



(c) Evolved Memory Antibodies Representation in 3 Dimensions.



(d) Evolved Memory Antibodies Minimum Spanning Tree Representation.

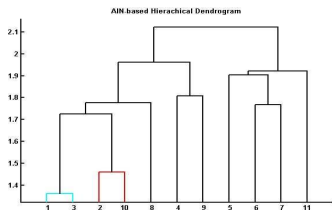
Παρουσιάζεται η χωρική κατανομή των αντισωμάτων μνήμης και το γράφημα που αντιστοιχεί στο Τ.Α.Δ. υπο τη μορφή του ελάχιστου δένδρου ζεύξης.

Παράμετροι εκπαίδευσης του Τ.Α.Δ.

Table: AIN Learning Algorithm Training Parameters

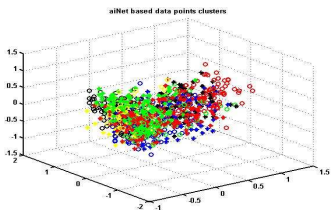
| AIN Parameter | Value |
|---------------|-------|
| N | 10 |
| n | 10 |
| ζ | 1 |
| σ_d | 0.10 |
| σ_s | 0.20 |
| GEN | 1 |

Αποτελέσματα - Ομαδοποίηση Δεδομένων: Τ.Α.Δ



(e) AIN-Based Hierarchical Dendrogram.

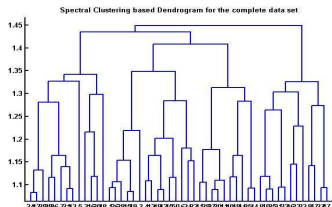
Παρουσιάζεται το δενδρόγραμμα που αντιστοιχεί στην ιεραρχική συγχωνευτική ομαδοποίηση των αντισωμάτων μνήμης που αναγνωρίζουν το αρχικό σύνολο των μουσικών δεδομένων.



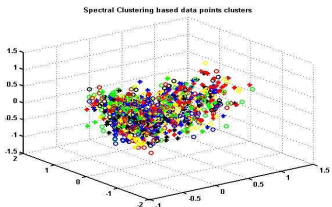
(f) AIN-Based Data Points Clusters.

Παρουσιάζεται η ποιοτική και ποσοτική περιγραφή των ομάδων που αναγνωρίστηκαν στο αρχικό σύνολο δεδομένων με βάση το Τ.Α.Δ.

Αποτελέσματα - Ομαδοποίηση Δεδομένων: Φασματική Ομαδοποίηση



(g) Spectral Clustering Based Dendrogram.

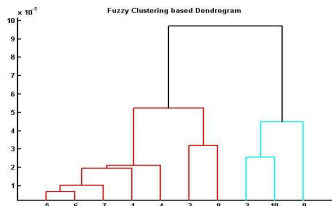


(h) Spectral Clustering Based Data Points Clusters.

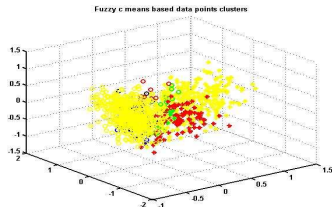
Παρουσιάζεται το δενδρόγραμμα που αντιστοιχεί στην ιεραρχική συγχωνευτική ομαδοποίηση των μουσικών δεδομένων με βάση τον αλγόριθμο της Φασματικής Ομαδοποίησης (Spectral Clustering).

Παρουσιάζεται η ποιοτική και ποσοτική περιγραφή των ομάδων που αναγνωρίστηκαν στο αρχικό σύνολο δεδομένων με βάση τον αλγόριθμο της Φασματικής Ομαδοποίησης.

Αποτελέσματα - Ομαδοποίηση Δεδομένων: Ασαφής Ομαδοποίηση



(i) Fuzzy Clustering based Dendrogram.



(j) Fuzzy c means based data points clusters.

Παρουσιάζεται το δενδρόγραμμα που αντιστοιχεί στην ιεραρχική συγχωνευτική ομαδοποίηση των μουσικών δεδομένων με βάση τον αλγόριθμο της Ασαφούς Ομαδοποίησης (Fuzzy C-Means Clustering).

Παρουσιάζεται η ποιοτική και ποσοτική περιγραφή των ομάδων που αναγνωρίστηκαν στο αρχικό σύνολο δεδομένων με βάση τον αλγόριθμο της Ασαφούς Ομαδοποίησης.

Συμπεράσματα I

- 1 Η πληροφορία που αφορά το πλήθος των εγγενών κλάσεων στο αρχικό σύνολο των μουσικών δεδομένων δεν είναι διαθέσιμη στους αλγορίθμους ομαδοποίησης.
- 2 Στόχος είναι η συγκριτική αξιολόγηση της αυτόματης οργάνωσης της δοσμένης μουσικής συλλογής σε ομάδες με βάση του αλγορίθμους:
 - Ιεραρχικής Ομαδοποίησης;
 - Φασματικής Ομαδοποίησης;
 - Ασαφούς Ομαδοποίησης;
 - Ομαδοποίησης βασισμένη στο T.A.Δ που αναπτύχθηκε.
- 3 Η ομαδοποίηση των μουσικών δεδομένων με βάση το T.A.Δ αποκαλύπτει καθαρότερα τη χωρική κατανομή των αρχικών δεδομένων.
 - Το δενδρόγραμμα που αντιστοιχεί στο σύνολο των αντισωμάτων μνήμης είναι ευκρινέστερο από το δενδρόγραμμα που αντιστοιχεί στο σύνολο των μουσικών δεδομένων.
 - Τα φύλλα του πρώτου δενδρογράμματος αντιστοιχούν στα αντισώματα μνήμης ενώ τα φύλλα του δεύτερου στα διανύσματα χαρακτηριστικών για το πλήρες σύνολο των μουσικών κομματιών.
 - Η γραφική αναπαράσταση του T.A.Δ. μέσω του ελάχιστου δένδρου ζεύξης παρέχει μια συνοπτική περιγραφή της χωρικής κατανομής των αναγνωρισθέντων ομάδων, αναδεικνύοντας την ικανότητα του T.A.Δ να αποβάλλει την πλεονάζουσα πληροφορία μέσω ενός εγγενούς μηχανισμού συμπίεσης.

Συμπεράσματα II

- 4 Τα αντισώματα μνήμης που παράγονται μπορούν να θεωρηθούν ως οι εσωτερικές εικόνες των αναγνωρισθέντων ομάδων δεδομένων στο αρχικό σύνολο.
- 5 Το πλήθος των αντισωμάτων μνήμης που παράγονται από το T.A.Δ. (11) προσεγγίζει με εξαιρετική ακρίβεια τον αριθμό των πραγματικών κλάσεων από τις οποίες προέρχονται τα μουσικά δεδομένα.
- 6 Η διαμέριση του συνόλου των μουσικών δεδομένων με βάση τα αντισώματα μνήμης του T.A.Δ. είναι συνεπέστερη σε σχέση με τη διαμέριση που προκύπτουν με βάση τα κέντρα των ομάδων που αναγνωρίζονται από τον αλγόριθμο της Φασματικής ομαδοποίησης. Η ομαδοποίηση με βάση το T.A.Δ οδηγεί σε περισσότερο ομοιογενείς ομάδες δεδομένων.
- 7 Τη μεγαλύτερη ομοιογένεια παρουσιάζουν οι ομάδες που αναγνωρίζει ο αλγόριθμος της Ασαφούς ομαδοποίησης. Αυτό συμβαίνει όμως διότι τα κέντρα των ομάδων που προκύπτουν είναι τόσο κοντά που σχεδόν ταυτίζονται. Συνεπώς για κάθε διάνυσμα χαρακτηριστικών αποδίδεται ο ίδιος βαθμός συμμετοχής σε κάθε ομάδα .

Διατύπωση Προβλήματος

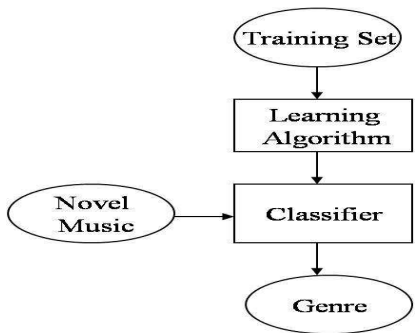


Figure: Αυτόματη Ταξινόμηση Μουσικολογικού Είδους.

Στόχος: Αυτόματη ταξινόμηση μουσικών κομματιών ως προς την μουσικολογική κλάση.

- *Χώρος Εισόδου:* Μουσικά κομμάτια.
- *Χώρος Εξόδου:* Μουσικολογική κλάση.
- *Σύνολο Εκπαίδευσης:* Κατηγοριοποιημένα μουσικά κομμάτια απο κάποιο εμπειρογνώμονα.
- *Ταξινομητής:* Τεχνητό Ανοσοποιητικό Σύστημα Αναγνώρισης.

Αναδιτύπωση Προβλήματος

Αναδιτύπωση Προβλήματος μέσα στο πλαίσιο των Τ.Α.Σ.

- Το σύνολο των μουσικών διανυσμάτων χαρακτηριστικών απο κάθε διαφορετική μουσικολογική κλάση μπορεί να θεωρηθεί ως ο δοσμένος προς αναγνώριση αντιγονικός πληθυσμός.
- Τα παραγόμενα αντισώματα μνήμης για κάθε διαφορετική μουσικολογική κλάση μπορούν να θεωρηθούν ως σημεία ενός 30 - διάστατου χώρου χαρακτηριστικών τα οποία λειτουργούν ως αντιπρόσωποι ενός συγκεκριμένου τμήματος του αντιγονικού πληθυσμού.
- Ο αλγόριθμος εκπαίδευσης ενός Τ.Α.Σ Αναγνώρισης στοχεύει στην παραγωγή ενός συνόλου αντισωμάτων μνήμης τα οποία θα έχουν την ικανότητα ταξινόμησης νέων μουσικών διανυσμάτων χαρακτηριστικών.
- Η διαδικασία εκπαίδευσης του Τ.Α.Σ Αναγνώρισης περιλαμβάνει έναν τεχνητό εξελικτικό μηχανισμό μέσα στα πλαίσια του οποίου το σύνολο των υποψήφιων αντισωμάτων μνήμης ανταγωνίζονται προκειμένου να προσαρτήσουν όσο το δυνατό μεγαλύτερο μέρος απο τους συνολικούς πόρους του συστήματος.
- Ο ανταγωνισμός για την εξασφάλιση μεγαλύτερου μέρους από τους συνολικούς πόρους του συστήματος αποτελεί τον βασικό παράγοντα άσκησης εξελικτικής πίεσης, η οποία οδηγεί στην παραγωγή αντισωμάτων υψηλής ταξινομητικής ικανότητας.

Επισκόπηση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α

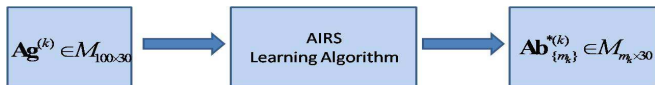


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ. Αναγνώρισης

Ο αλγόριθμος εκπαίδευσης του Τ.Α.Σ.Α στοχεύει στην παραγωγή ενός συνόλου αντισωμάτων μνήμης $\mathbf{Ab}_{\{m_k\}}^{*(k)}$ με $m_k < 100$, για κάθε κλάση $k \in [C]$ αντιγονικών προτύπων, τέτοια ώστε:

$$\mathbf{Ab}_{\{m_k\}}^{*(k)} = \arg \max_{\mathbf{Ab}_{\{m_k\}} \in \mathbb{S}^{m_k}} R^{(k)}(\mathbf{Ab}_{\{m_k\}}) \quad (20)$$

με

$$R^{(k)}(\mathbf{Ab}_{\{m_k\}}) = \frac{1}{m_k \cdot 100} \sum_{i=1}^m \sum_{j=1}^{1000} \text{stim}(\mathbf{Ab}_{\{m_k\}}^i, \mathbf{Ag}_j^{(k)}) \quad (21)$$

$$\text{Required Memory} = (M + N + m + 2N_c)L + 2N + C + M^2 \quad (22)$$

$$\text{Required Time} = O(\hat{k}ML) \quad (23)$$

όπου $\hat{k} = f(ST)$ με f γνησίως αύξουσα συνάρτηση του κατωφλίου διέγερσης ST .

Παράμετροι Εισόδου Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α

Ο αλγόριθμος εκπαίδευσης του Τ.Α.Σ Αναγνώρισης για την ταξινόμηση μουσικών διανυσμάτων χαρακτηριστικών δέχεται ως είσοδο τις παρακάτω παραμέτρους:

- *Βαθμωτό Κατωφλίου Συμπληρωματικότητας (Affinity Threshold Scalar)*: Παράμετρος ελέγχου που καθορίζει το αν ένα συγκεκριμένο υποψήφιο αντισώμα μνήμης θα συμπεριληφθεί στο σύνολο των αντισωμάτων μνήμης.
- *Ρυθμός Κλωνοποίησης (Clonal Rate)*: Καθορίζει το πλήθος των μεταλλαγμένων απογόνων που θα παραχθούν από ένα συγκεκριμένο αντισωματικό πρότυπο.
- *Ρυθμός Υπερμετάλλαξης (Hyper Mutation Rate)*: Καθορίζει το μέτρο διατάραξης των συνιστωσών του μεταλλασσόμενου αντισώματος.
- *Αριθμός Πλησιέστερων Γειτόνων (KNN)*: Το πλήθος των πλησιέστερων γειτόνων που λαμβάνονται υπόψη κατά την ταξινόμηση ενός νέου αντιγονικού προτύπου.
- *Κατώφλι Διέγερσης (Stimulation Threshold)*: Η ελάχιστη απαιτούμενη τιμή της μέσης διέγερσης για το σύνολο των διαθέσιμων αντισωμάτων.
- *Συνολικό Πλήθος Πόρων (Total Resources)*: Το συνολικό πλήθος των πόρων του συστήματος.

Αλγόριθμος Εκπαίδευσης Τεχνητού Ανοσοποιητικού Συστήματος Αναγνώρισης I

1 Αρχικοποίηση:

- 1 *Κανονικοποίηση Δεδομένων*: Όλα τα διανύσματα χαρακτηριστικών κανονικοποιούνται στο διάστημα $[0, 1]$.
- 2 *Υπολογισμός Κατώφλιου Μέτρου Συμπληρωματικότητας*: Το κατώφλι του μέτρου συμπληρωματικότητας υπολογίζεται ως η μέση τιμή των κανονικοποιημένων αποστάσεων για όλα τα ζεύγη των διανυσμάτων χαρακτηριστικών στο αρχικό σύνολο των μουσικών δεδομένων.
- 3 *Αρχικοποίηση Συνόλου Αντισωμάτων Μνήμης*: Το σύνολο των αντισωμάτων μνήμης για κάθε κλάση αρχικοποιείται αρχικοποιείται με το τρέχον αντιγονικό πρότυπο από τη ίδια κλάση προτύπων.
- 4 *Αρχικοποίηση Συνόλου Διαθέσιμων Αντισωμάτων*: Το σύνολο των διαθέσιμων αντισωμάτων για κάθε κλάση αρχικοποιείται με ένα τυχαίο διάνυσμα χαρακτηριστικών στον θεωρούμενο χώρο-σχημάτων.

2 Φάση Εκπαίδευσης (Παρουσίαση Αντιγονικών Προτύπων): Για κάθε κλάση προτύπων και κάθε αντιγονικό πρότυπο:

- 1 *Προσδιορισμός Συμβατού Αντισώματος Μνήμης*: Προσδιορισμός του αντισώματος μνήμης που παρουσιάζει τον μεγαλύτερο βαθμό διέγερσης προς το τρέχον αντιγονικό πρότυπο.

Αλγόριθμος Εκπαίδευσης Τεχνητού Ανοσοποιητικού Συστήματος Αναγνώρισης II

- 2 Παραγωγή Αντισωμάτων:** Το αντίσωμα μνήμης που παρουσιάζει τον μεγαλύτερο βαθμό διέγερσης προς το τρέχον αντιγονικό πρότυπο χρησιμοποιείται ως το αρχέτυπο για την παραγωγή ενός συνόλου από μεταλλαγμένες εκδοχές του αρχικού διανύσματος χαρακτηριστικών. Τα παραγόμενα αντισώματα θα συμπεριληφθούν στο σύνολο των διαθέσιμων αντισωμάτων. Ο ρυθμός μετάλλαξης του συμβατού αντίγονου μνήμης είναι αντιστρόφως ανάλογος του βαθμού διέγερσης προς το τρέχον αντιγονικό πρότυπο.
- 3 Διαδικασία Εκπαίδευσης:** Ενόσω ο μέσος βαθμός διέγερσης του συνόλου των διαθέσιμων αντισωμάτων είναι μικρότερος από κάποια προκαθορισμένη τιμή:
 - 1 Κατανομή Πόρων:** Για κάθε στοιχείο του συνόλου των διαθέσιμων αντισωμάτων δεσμεύεται ένα μέρος των συνολικών πόρων του συστήματος ανάλογα με το βαθμό διέγερσής του προς το τρέχον αντιγονικό πρότυπο.
 - 2 Καταστολή Διαθέσιμων Αντισωμάτων:** Απαλοιφή εκείνων των αντισωμάτων που δέσμευσαν το μικρότερο μέρος από τους συνολικούς πόρους του συστήματος.
 - 3 Παραγωγή Μεταλλαγμένων Απογόνων:** Το υποσύνολο των διαθέσιμων αντισωμάτων που έχουν εξασφαλίσει το μεγαλύτερο μέρος των πόρων του συστήματος, έχουν μια επιπλέον ευκαιρία για την παραγωγή μεταλλαγμένων απογόνων.
- 4 Προσδιορισμός Υποψήφιου Αντισώματος Μνήμης:** Ως υποψήφιο αντίσωμα μνήμης επιλέγεται εκείνο το διάνυσμα χαρακτηριστικών από το σύνολο των διαθέσιμων αντισωμάτων με το μεγαλύτερο βαθμό διέγερσης προς το τρέχον αντιγονικό πρότυπο.

Αλγόριθμος Εκπαίδευσης Τεχνητού Ανοσοποιητικού Συστήματος Αναγνώρισης III

- 5** *Εισαγωγή Αντισώματος Μνήμης:* Το υποψήφιο αντίσωμα μνήμης προστίθεται στο σύνολο των αντισωμάτων μνήμης αν ο βαθμός διέγερσής του προς το τρέχον αντιγονικό πρότυπο είναι μεγαλύτερος από αυτόν του συμβατού αντισώματος μνήμης. Στην περίπτωση που πληροί αυτή την προϋπόθεση, υπολογίζεται το μέτρο της συμπληρωματικότητάς σε σχέση με το συμβατό αντίσωμα μνήμης. Αν η συγκεκριμένη τιμή του μέτρου συμπληρωματικότητας είναι μικρότερη από κάποια προκαθορισμένη τιμή τότε το υποψήφιο αντίσωμα μνήμης αντικαθιστά το συμβατό αντίσωμα μνήμης.
- 3** *Ταξινόμηση:* Το σύνολο των παραχθέντων αντισωμάτων μνήμης χρησιμοποιείται για την ταξινόμηση νέων διανυσμάτων μουσικών χαρακτηριστικών με βάση τη λογική του πλησιέστερου γείτονα.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α Ι

Παρουσίαση Αντιγονικού Προτύπου στο Σύνολο των Αντισωμάτων Μνήμης.

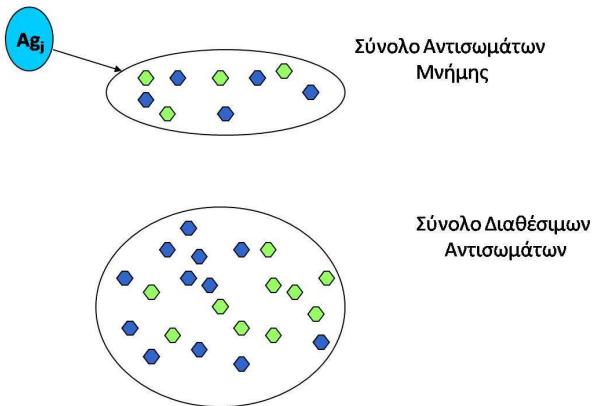


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ Αναγνώρισης Ι.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α ΙΙ

Προσδιορισμός Συμβατού Αντισώματος Μνήμης

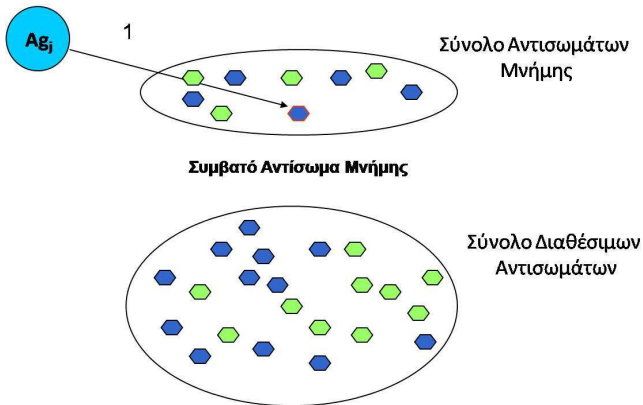


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ Αναγνώρισης ΙΙ.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α ΙΙΙ

Παραγωγή Αντισωμάτων

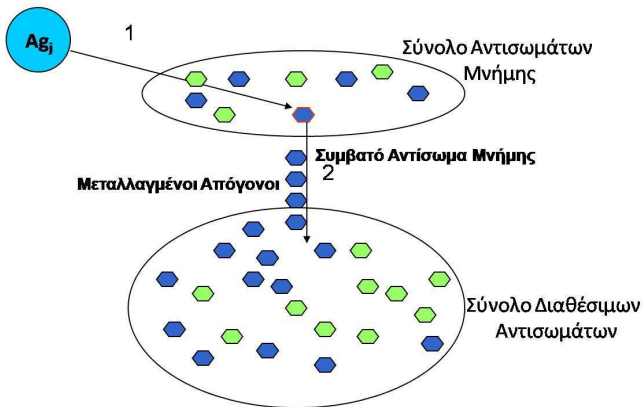


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ Αναγνώρισης ΙΙΙ.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α IV

Παρουσίαση Αντιγονικού Πρότυπου στο Σύνολο των Διαθέσιμων Αντισωμάτων

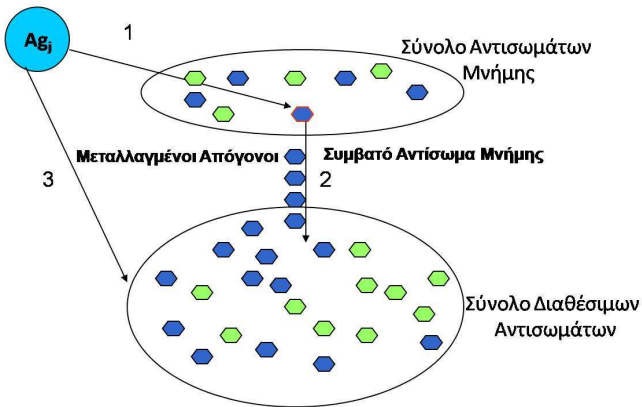


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ Αναγνώρισης IV.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α V

Διαδικασία Εκπαίδευσης - Παραγωγή Υποψήφιου Αντισώματος Μνήμης

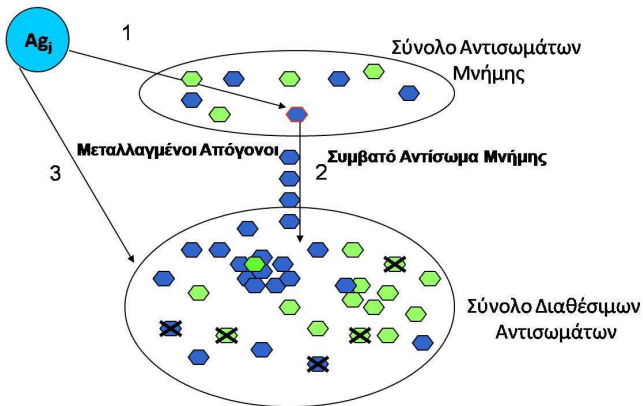


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ Αναγνώρισης V.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α VI

Σύγκριση Υποψήφιου Αντισώματος Μνήμης - Συμβατού Αντισώματος Μνήμης

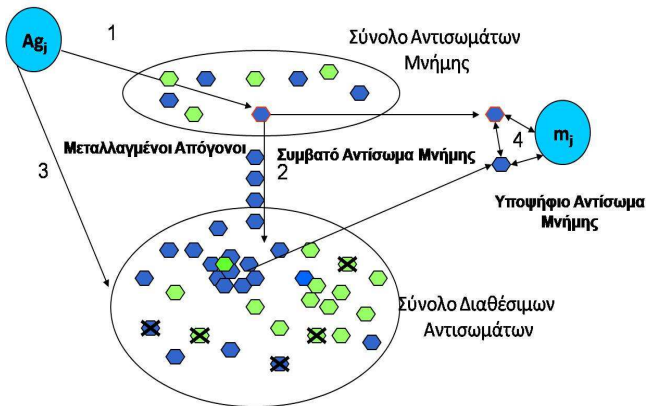


Figure: Αυτόματη Ταξινόμηση Μουσικολογικού Είδους VI.

Σχηματική Αναπαράσταση Αλγορίθμου Εκπαίδευσης Τ.Α.Σ.Α VII

Εισαγωγή Αντισώματος Μνήμης

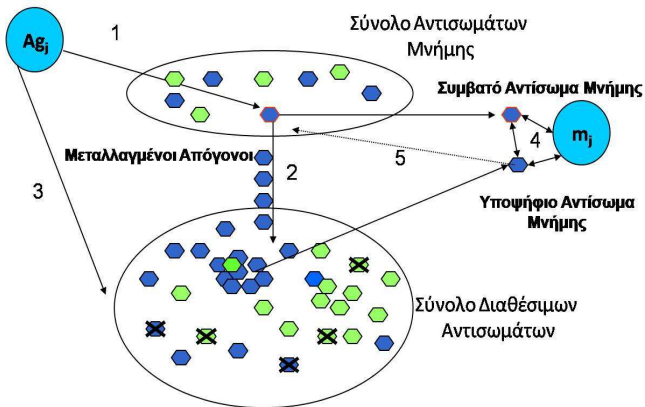


Figure: Αλγόριθμος Εκπαίδευσης Τ.Α.Σ. Αναγνώρισης VII.

Προβλήματα Ταξινόμησης

1. Ισοζυγισμένα Προβλήματα Πολυταξικής Ταξινόμησης:

- $C1$ vs $C2$;
- $C1$ vs $C2$ vs $C3$;
- $C1$ vs $C2$ vs $C3$ vs $C4$;
- $C1$ vs $C2$ vs $C3$ vs $C4$ vs $C5$;
- $C1$ vs $C2$ vs $C3$ vs $C4$ vs $C5$ vs $C6$
- $C1$ vs $C2$ vs $C3$ vs $C4$ vs $C5$ vs $C6$ vs $C7$;
- $C1$ vs $C2$ vs $C3$ vs $C4$ vs $C5$ vs $C6$ vs $C7$ vs $C8$;
- $C1$ vs $C2$ vs $C3$ vs $C4$ vs $C5$ vs $C6$ vs $C7$ vs $C8$ vs $C9$;
- $C1$ vs $C2$ vs $C3$ vs $C4$ vs $C5$ vs $C6$ vs $C7$ vs $C8$ vs $C9$ vs $C10$;

2. Ισοζυγισμένα Προβλήματα One vs All Ταξινόμησης:

- $C1$ vs $\{C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C2$ vs $\{C1 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C3$ vs $\{C1 \cup C2 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C4$ vs $\{C1 \cup C2 \cup C3 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C5$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C6$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C7$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C8 \cup C9 \cup C10\}$;
- $C8$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C9 \cup C10\}$;
- $C9$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C10\}$;
- $C10$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9\}$;

3. Μη-Ισοζυγισμένα Προβλήματα One vs All Ταξινόμησης:

- $C1$ vs $\{C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C2$ vs $\{C1 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C3$ vs $\{C1 \cup C2 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C4$ vs $\{C1 \cup C2 \cup C3 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C5$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C6 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C6$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C7 \cup C8 \cup C9 \cup C10\}$;
- $C7$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C8 \cup C9 \cup C10\}$;
- $C8$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C9 \cup C10\}$;
- $C9$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C10\}$;
- $C10$ vs $\{C1 \cup C2 \cup C3 \cup C4 \cup C5 \cup C6 \cup C7 \cup C8 \cup C9\}$;

Αποτελέσματα (Ισοζυγισμένα Πολυταξικά Προβλήματα Ταξινόμησης)

AIRS

Table: Ισοζυγισμένα Πολυταξικά Προβλήματα Ταξινόμησης: AIRS

BALANCED MULTI CLASS CLASSIFICATION PROBLEMS

AIRS CLASSIFIER

| CLASSIFICATION PROBLEM | ACCURACY | ERROR RATE |
|---|----------|------------|
| C1 vs C2 | 94% | 6% |
| C1 vs C2 vs C3 | 72.3333% | 27.6667% |
| C1 vs C2 vs C3 vs C4 | 68.5% | 31.5% |
| C1 vs C2 vs C3 vs C4 vs C5 | 61% | 39% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 | 57.6667% | 42.3333% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 | 57.8571% | 42.1429% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 vs C8 | 52.5% | 47.5% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 vs C8 vs C9 | 48.6667% | 51.3333% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 vs C8 vs C9 vs C10 | 44.3% | 55.7% |

Αποτελέσματα (Ισοζυγισμένα Πολυταξικά Προβλήματα Ταξινόμησης)

SVM

Table: Ισοζυγισμένα Πολυταξικά Προβλήματα Ταξινόμησης: SVM

BALANCED MULTI CLASS CLASSIFICATION PROBLEMS

SVM CLASSIFIER

| CLASSIFICATION PROBLEM | ACCURACY | ERROR RATE |
|---|----------|------------|
| C1 vs C2 | 92.5% | 7.5% |
| C1 vs C2 vs C3 | 75.3333% | 24.66674% |
| C1 vs C2 vs C3 vs C4 | 71% | 29% |
| C1 vs C2 vs C3 vs C4 vs C5 | 63.4% | 36.6% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 | 61.8333% | 38.1667% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 | 59.7143% | 40.2857% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 vs C8 | 57.125% | 42.875% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 vs C8 vs C9 | 53% | 47% |
| C1 vs C2 vs C3 vs C4 vs C5 vs C6 vs C7 vs C8 vs C9 vs C10 | 50% | 50% |

Αποτελέσματα (Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης: Τ.Α.Σ.Α)

Table: Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης: Τ.Α.Σ.Α

BALANCED ONE VS ALL CLASSIFICATION PROBLEMS

AIRS CLASSIFIER

| CLASSIFICATION PROBLEM | ACCURACY | ERROR RATE |
|------------------------|----------|------------|
| C1 vs ALL | 70.5 | 29.5 |
| C2 vs ALL | 90 | 10 |
| C3 vs ALL | 70.5 | 29.5 |
| C4 vs ALL | 74.5 | 25.5 |
| C5 vs ALL | 73 | 27 |
| C6 vs ALL | 81 | 19 |
| C7 vs ALL | 81.5 | 18.5 |
| C8 vs ALL | 68 | 32 |
| C9 vs ALL | 73 | 27 |
| C10 vs ALL | 63.5 | 36.5 |

Αποτελέσματα (Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης: Μ.Δ.Υ

Table: Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης: Μ.Δ.Υ

| BALANCED ONE VS ALL CLASSIFICATION PROBLEMS | | |
|--|-----------------|-------------------|
| SVM CLASSIFIER | | |
| CLASSIFICATION PROBLEM | ACCURACY | ERROR RATE |
| C1 vs ALL | 55 | 45 |
| C2 vs ALL | 100 | 0 |
| C3 vs ALL | 65 | 35 |
| C4 vs ALL | 80 | 20 |
| C5 vs ALL | 45 | 55 |
| C6 vs ALL | 90 | 10 |
| C7 vs ALL | 80 | 20 |
| C8 vs ALL | 65 | 35 |
| C9 vs ALL | 80 | 20 |
| C10 vs ALL | 80 | 20 |

Αποτελέσματα: Μη-Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης (Τ.Α.Σ)

Table: Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης: Τ.Α.Σ

UNBALANCED ONE VS ALL CLASSIFICATION PROBLEMS

AIRS CLASSIFIER

| CLASSIFICATION PROBLEM | MINORITY CLASS TP RATE | MINORITY CLASS F-MEASURE | MAJORITY CLASS TP RATE | MAJORITY CLASS F-MEASURE |
|------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|
| C1 vs ALL | 0.31 | 0.326 | 0.934 | 0.929 |
| C2 vs ALL | 0.72 | 0.727 | 0.971 | 0.97 |
| C3 vs ALL | 0.24 | 0.324 | 0.973 | 0.946 |
| C4 vs ALL | 0.36 | 0.34 | 0.916 | 0.922 |
| C5 vs ALL | 0.27 | 0.293 | 0.937 | 0.928 |
| C6 vs ALL | 0.42 | 0.519 | 0.978 | 0.958 |
| C7 vs ALL | 0.46 | 0.5 | 0.958 | 0.949 |
| C8 vs ALL | 0.46 | 0.517 | 0.964 | 0.953 |
| C9 vs ALL | 0.26 | 0.344 | 0.972 | 0.946 |
| C10 vs ALL | 0.04 | 0.073 | 0.993 | 0.946 |

Αποτελέσματα: Μη-Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης (SVM)

Table: Μη-Ισοζυγισμένα One vs All Προβλήματα Ταξινόμησης: Τ.Α.Σ

UNBALANCED ONE VS ALL CLASSIFICATION PROBLEMS

AIRS CLASSIFIER

| CLASSIFICATION PROBLEM | MINORITY CLASS TP RATE | MINORITY CLASS F-MEASURE | MAJORITY CLASS TP RATE | MAJORITY CLASS F-MEASURE |
|------------------------|---------------------------|-----------------------------|---------------------------|-----------------------------|
| C1 vs ALL | 0.01 | 0.02 | 1 | 0.948 |
| C2 vs ALL | 0.61 | 0.718 | 0.99 | 0.974 |
| C3 vs ALL | 0.03 | 0.057 | 0.998 | 0.948 |
| C4 vs ALL | 0.01 | 0.02 | 1 | 0.948 |
| C5 vs ALL | 0 | 0 | 1 | 0.947 |
| C6 vs ALL | 0.24 | 0.372 | 0.994 | 0.957 |
| C7 vs ALL | 0.31 | 0.44 | 0.989 | 0.958 |
| C8 vs ALL | 0.29 | 0.42 | 0.99 | 0.957 |
| C9 vs ALL | 0.01 | 0.02 | 0.999 | 0.947 |
| C10 vs ALL | 0 | 0 | 1 | 0.947 |

Συμπεράσματα I

Για την σύγκριση της ταξινομητικής ακρίβειας των δύο ταξινομητών χρησιμοποιήθηκε η μέθοδος της 10-πλής διεπικύρωσης.

1 Ισοζυγισμένα Προβλήματα Πολυταξικής Ταξινόμησης:

- 1 Η ταξινομητική ακρίβεια του T.A.Σ.A είναι παρόμοια με αυτή των M.Δ.Y
- 2 Η ταξινομητική ακρίβεια και των δύο προσεγγίσεων μειώνεται όσο αυξάνεται το πλήθος των προς ταξινόμηση κλάσεων.
- 3 Η ταξινομητική ακρίβεια των M.Δ.Y σε προβλήματα πολυταξικής ταξινόμησης (για παραπάνω από 2 κλάσεις) είναι μεγαλύτερη:
 - Οι M.Δ.Y αποτελούν μη-γραμμικούς ταξινομητές που λειτουργούν σε χώρους ανώτερης διάστασης σε σχέση με την διάσταση του χώρου στον οποίο γίνεται η αρχική διατύπωση του προβλήματος.
 - Το T.A.Σ.A αποτελεί έναν μη-γραμμικό ταξινομητή ο οποίος όμως λειτουργεί στον αρχικό χώρο διατύπωσης του προβλήματος.
 - Οι M.Δ.Y λειτουργούν στη βάση ενός συνόλου Διανυσμάτων Υποστήριξης (για κάθε ζεύγος κλάσεων) που αντιστοιχούν σε αρχικά δεδομένα του προβλήματος που παραβιάζουν τους περιορισμούς του υποκείμενου προβλήματος βελτιστοποίησης.
 - Το T.A.Σ.A λειτουργεί στη βάση ενός συνόλου αντισωμάτων μνήμης (για κάθε κλάση αντιγονικών προτύπων) το οποίο είναι κατά κανόνα μικρότερο σε πλήθος σε σχέση με το αρχικό σύνολο των δεδομένων.
 - Οι M.Δ.Y υπερτερούν σε προβλήματα ταξινόμησης με μεγαλύτερο πλήθος κλάσεων διότι λειτουργούν στη βάση ενός συνδυαστικού σχήματος δυαδικών ταξινομητών σε σχέση με το T.A.Σ.A που είναι από σχεδιασμού του ένας πολυταξικός ταξινομητής.
 - Ο αλγόριθμος εκπαίδευσης των M.Δ.Y έχει περισσότερες απαιτήσεις σε μνήμη ($\binom{N}{2}$ σύνολα Διανυσμάτων Υποστήριξης) σε σχέση με τον αντίστοιχο αλγόριθμο για το T.A.Σ.A (N σύνολα αντισωμάτων μνήμης).

Συμπεράσματα II

- Για παράδειγμα στο πρόβλημα των 10 κλάσεων οι Μ.Δ.Υ απαιτούν την αποθήκευση 45 συνόλων Διανυσμάτων Υποστήριξης έναντι 10 συνόλων αντισωμάτων μνήμης για το Τ.Α.Σ.Α
- 2 **Ισοζυγισμένα Προβλήματα One vs All Ταξινόμησης:**
 - Η θετική κλάση προτύπων είναι κάθε φορά μια απο τις 10 μουσικολογικές κλάσεις του αρχικού χώρου των δεδομένων.
 - Η αρνητική κλάση προτύπων είναι κάθε φορά το συμπλήρωμα (9 μουσικολογικές κλάσεις) του αρχικού χώρου των δεδομένων ως προς την θετική κλάση.
 - 1 Η ταξινομητική ακρίβεια του Τ.Α.Σ.Α είναι βελτιωμένη αισθητά σε σχέση με αυτή των Μ.Δ.Υ σε προβλήματα One vs All ταξινόμησης.
 - 2 Η ταξινομητική συμπεριφορά των Μ.Δ.Υ είναι πολύ πιο ακραία στα επιμέρους βήματα της διαδικασίας 10-πλής επικύρωσης.
 - 3 **Μη-Ισοζυγισμένα Προβλήματα One vs All Ταξινόμησης:**
 - Συνολικά Πρότυπα Θετικής Κλάσης: 100
 - Συνολικά Πρότυπα Αρνητικής Κλάσης: 900
 - Συνολικά Πρότυπα Θετικής Κλάσης (κατά τη φάση του ελέγχου ανά βήμα της 10-πλής επικύρωσης): 10
 - Συνολικά Πρότυπα Αρνητικής Κλάσης (κατά τη φάση του ελέγχου ανά βήμα της 10-πλής επικύρωσης): 90
 - Συνολικά Πρότυπα Θετικής Κλάσης προς αναγνώριση: 100
 - Συνολικά Πρότυπα Αρνητικής Κλάσης προς αναγνώριση: 900
 - 1 Το Τ.Α.Σ.Α σε ακραία προβλήματα ταξικής ανισορροπίας έχει την ικανότητα να αναγνωρίζει σε πολύ μεγαλύτερο βαθμό την κλάση μειοψηφίας (10% του συνολικού χώρου) σε σχέση με τις Μ.Δ.Υ
 - 2 Οι τιμές των True Positive Rate και F-Measure είναι συγκριτικά μεγαλύτερες για το Τ.Α.Σ.Α για το σύνολο και των 10 πειραμάτων ταξικής ανισορροπίας.

Η Σύσταση αντικειμένων ως πρόβλημα Μονοταξικής Ταξινόμησης I

- Το κύριο πρόβλημα που αντιμετωπίζει ο σχεδιασμός ενός αποδοτικού Σύστημα Σύστασης πολυμεσικών δεδομένων είναι η δυσκολία που αντιμετωπίζουν οι χρήστες να εκφράσουν τις ανάγκες τους.
- Οι χρήστες παρέχουν πληροφορία για αντικείμενα που ανήκουν στην κλάση ενδιαφέροντός τους.
- Οι χρήστες συνήθως δεν παρέχουν πληροφορία για αντικείμενα εκτός κλάσης ενδιαφέροντος εξαιτίας του κόστους που εκφράζεται με όρους χρόνου και προσπάθειας.
- Το σύνολο των αντικειμένων που ανήκουν στην κλάση ενδιαφέροντος ενός χρήστη είναι πολύ μικρό σε σχέση με το συνολικό αριθμό των αντικειμένων που είναι διαθέσιμα.
- Το γεγονός αυτό δικαιολογεί τον υψηλό βαθμό ασυμμετρίας-ανισορροπίας μεταξύ της κλάσης ενδιαφέροντος και της συμπληρωματικής που υπάρχει στο πρόβλημα της Σύστασης αντικειμένων.
- Το κόστος συλλογής πληροφορίας για τα αντικείμενα εκτός κλάσης ενδιαφέροντος είναι εξαιρετικά δαπανηρό.
- Τα αντικείμενα εκτός κλάσης ενδιαφέροντος μπορεί να έχουν ληφθεί δειγματοληπτικά με κακώς τιθέμενες κατανομές.

Η Σύσταση αντικειμένων ως πρόβλημα Μονοταξικής Ταξινόμησης II

- Συνεπώς τεχνικές μηχανικής μάθησης που βασίζονται σε πληροφορία και από τις δύο αυτές κλάσεις δεν είναι εφαρμόσιμες.
- Είναι αναγκαίες τεχνικές που να μπορούν να δημιουργήσουν ένα μοντέλο το οποίο να παρέχει:
 - είτε μια στατιστική περιγραφή της κλάσης ενδιαφέροντος
 - είτε μια περιγραφή για το σχήμα-δομή της κλάσης που παράγει τα πρότυπα εκπαίδευσης
- Η κλάση ενδιαφέροντος μπορεί να περιλαμβάνει αντικείμενα προερχόμενα από εσωτερικές υποομάδες με διαφορετική σημασιολογία.
- Για παράδειγμα η κλάση ενδιαφέροντος ενός χρήστη μπορεί να περιλαμβάνει μουσικά αρχεία προερχόμενα από 10 διαφορετικά μουσικά είδη.
- Σημαντικός παράγοντας στην επιλογή της μονοταξικής μάθησης είναι το κόστος της λανθασμένης ταξινόμησης πρέπει να είναι ανάλογο με την ανισορροπία των δύο κλάσεων (ενδιαφέροντος / συμπληρωματική).

Αλγόριθμοι Μονοταξικής Ταξινόμησης I

Definition (Αλγόριθμοι Μονοταξικής Ταξινόμησης)

Οι Αλγόριθμοι Μονοταξικής Ταξινόμησης αντιστοιχούν σε μια ιδιαίτερη κατηγορία μηχανών μάθησης και αφορούν στην επίλυση ακραίων προβλημάτων ταξικής ανισορροπίας. Συγκεκριμένα, επικεντρώνονται σε εκφυλισμένα προβλήματα δυαδικής ταξινόμησης όπου η κλάση πλειοψηφίας αγνοείται πλήρως κατά την εκπαίδευση του ταξινομητή. Αν $\mathbf{X}_T \subset \mathbf{X}$ είναι ο υποχώρος των προτύπων που αντιστοιχεί στην κλάση ενδιαφέροντος, τότε σκοπός ενός Μονοταξικού Ταξινομητή είναι η κατασκευή ενός μοντέλου $f : \mathbf{X} \rightarrow \{0, 1\}$ τέτοιο ώστε:

$$f(\mathbf{x}) = \begin{cases} 1, & \mathbf{x} \in \mathbf{X}_T; \\ 0, & \text{διαφορετικά.} \end{cases} \quad (24)$$

στη βάση ενός συνόλου δεδομένων εκπαίδευσης $\hat{\mathbf{X}}_T = \{\mathbf{x}_1, \dots, \mathbf{x}_l\} \subset \mathbf{X}_T$ τα οποία προέρχονται αποκλειστικά από την κλάση ενδιαφέροντος.

Το σημαντικότερο γνώρισμα των Μονοταξικών Ταξινομητών είναι το ότι η διαδικασία της εκπαίδευσής τους συνίσταται στην προσπάθεια αμοιβαίας ικανοποίησης δύο αντικρουόμενων στόχων:

- της ελαχιστοποίησης του ποσοστού της κλάσης ενδιαφέροντος που απορρίπτεται (**False Negative Rate / Type I Error / E_I**);

Αλγόριθμοι Μονοταξικής Ταξινόμησης II

- της ελαχιστοποίησης του ποσοστού της κλάσης πλειοψηφίας που γίνεται αποδεκτό (**False Positive Rate / Type II Error / E_{II}**).

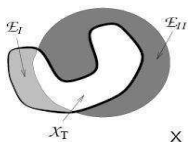


Figure: Περιοχές του χώρου των προτύπων στη Μονοταξική Ταξινόμηση

- Η διαδικασία εκπαίδευσης ενός Μονοταξικού Ταξινομητή συνιστά ένα πολυκριτήριο πρόβλημα βελτιστοποίησης το οποίο ανάγεται στην αμοιβαία ελαχιστοποίηση των ανταγωνιστικών ποσοτήτων E_I και E_{II} ($E_I \downarrow \Rightarrow E_{II} \uparrow$ και $E_{II} \downarrow \Rightarrow E_I \uparrow$).
- Η σκούρα γκριζα κυκλική περιοχή περιγράφει τον υποχώρο των προτύπων που γίνεται αποδεκτός από το μοντέλο του Μονοταξικού Ταξινομητή.
- Η ανοικτή γκριζα περιοχή περιγράφει τον υποχώρο των προτύπων που δεν γίνεται αποδεκτός από το μοντέλο του Μονοταξικού Ταξινομητή.

Μονοταξικές Μηχανές Διανυσμάτων Υποστήριξης I

- Οι Μονοταξικές Μηχανές Διανυσμάτων Υποστήριξης λειτουργούν με βάση το μοντέλο της Σφαιρικής Περιγραφής Δεδομένων (Spherical Data Description).
- Το συγκεκριμένο μοντέλο στοχεύει στη δημιουργία ενός κλειστού συνόρου γύρω από το σύνολο των δεδομένων που ανήκουν στην θετική κλάση.
- Στην ιδανική περίπτωση οι Μονοταξικές Μηχανές Διανυσμάτων Υποστήριξης (One-Class Support Vector Machines) προσπαθούν να προσδιορίσουν τις παραμέτρους (a : κέντρο, R : ακτίνα) μιας υπερσφαίρας στο χώρο $\mathbf{X} \in \mathbb{R}^n$ οι οποία θα περιέχει εντός των ορίων της το σύνολο των προτύπων της θετικής κλάσης.

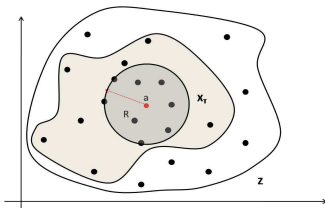


Figure: Σφαιρική Περιγραφή Δεδομένων

Μονοταξικές Μηχανές Διανυσμάτων Υποστήριξης II

- Η μαθηματική περιγραφή του μοντέλου f που υλοποιούν οι Μονοταξικές Μηχανές Διανυσμάτων Υποστήριξης δίνεται από την παρακάτω σχέση:

$$f(\mathbf{x}; \mathbf{a}, R) = \theta(R^2 - \|\mathbf{x} - \mathbf{a}\|^2) \quad (25)$$

όπου:

$$\theta(u) = \begin{cases} 1, & u \geq 0; \\ 0, & u < 0. \end{cases} \quad (26)$$

η οποία υπολογίζεται στη βάση ενός συνόλου δεδομένων εκπαίδευσης $\hat{\mathbf{X}}_T = \{\mathbf{x}_1, \dots, \mathbf{x}_l\} \subset \mathbf{X}_T$ τα οποία προέρχονται αποκλειστικά από την κλάση ενδιαφέροντος.

Ελαχιστοποίηση Εμπειρικού και Δομικού Σφάλματος

- Η απαίτηση για την παρουσία όλων των προτύπων εκπαίδευσης εντός των ορίων της υπολογιζόμενης υπερσφαίρας είναι εξαιρετικά αυστηρή και στην πράξη αντικαθίσταται από ένα σύνολο πιο χαλαρών περιορισμών της μορφής:

$$\|x_i - a\|^2 \leq R^2 + \xi_i, \quad \forall i \in [l] \quad (27)$$

όπου $\xi_i \geq 0, \forall i \in [l]$

- Η διαδικασία εκπαίδευσης των Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης ανάγεται σε ένα πολυκριτήριο πρόβλημα βελτιστοποίησης δύο αντικρουόμενων στόχων.
- Ελαχιστοποίηση του πλήθους των προτύπων που δεν προέρχονται από την θετική κλάση αλλά βρίσκονται εντός των ορίων της υπολογιζόμενης υπερσφαίρας.

$$E_{struct} = R^2 \text{ (Structural Error)} \quad (28)$$

- Ελαχιστοποίηση του πλήθους των προτύπων που προέρχονται από την θετική κλάση αλλά βρίσκονται εκτός των ορίων της υπολογιζόμενης υπερσφαίρας.

$$E_{emp} = \sum_{i=1}^l \xi_i \text{ (Empirical Error)} \quad (29)$$

Μαθηματική Ανάλυση Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης

Definition (Πρωτεύον Πρόβλημα Βελτιστοποίησης Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης)

$$\min_{R, \mathbf{a}, \boldsymbol{\xi}} \quad R^2 + C \sum_{i=1}^l \xi_i \quad (30a)$$

$$\text{s.t} \quad \|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \forall i \in [l] \quad (30b)$$

$$\text{and} \quad \xi_i \geq 0, \quad \forall i \in [l] \quad (30c)$$

Definition (Δυϊκό Πρόβλημα Βελτιστοποίησης Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης)

$$\max_{\boldsymbol{\alpha}, \boldsymbol{\gamma}} \min_{R, \mathbf{a}, \boldsymbol{\xi}} \quad L(R, \mathbf{a}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) \quad (31a)$$

$$\text{s.t} \quad \alpha_i \geq 0, \quad \forall i \in [l] \quad (31b)$$

$$\text{and} \quad \gamma_i \geq 0, \quad \forall i \in [l] \quad (31c)$$

όπου

$$L(R, \mathbf{a}, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\gamma}) = \sum_{i=1}^l \alpha_i \langle \mathbf{x}_i, \mathbf{x}_i \rangle - \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle \quad (32)$$

Επίλυση Δυϊκού Προβλήματος Βελτιστοποίησης Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης

Definition (Βέλτιστη τιμή κέντρου)

$$\mathbf{a}^* = \sum_{i=1}^l \alpha_i \mathbf{x}_i \quad (33)$$

Definition (Βέλτιστες τιμές χαλαρών μεταβλητών)

$$\forall \mathbf{x}_i \in SV, \xi_i^* = \max(0, \|\mathbf{x}_i - \mathbf{a}\|^2 - R^2) \quad (34)$$

Definition (Βέλτιστη τιμή ακτίνας)

$$R^2 = \langle \mathbf{x}_k, \mathbf{x}_k \rangle - 2 \sum_{i=1}^l \alpha_i \langle \mathbf{x}_i, \mathbf{x}_k \rangle + \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle, \forall \mathbf{x}_k \in SV^{bnd} \quad (35)$$

Definition (Διανύσματα Υποστήριξης)

$$SV = SV^{bnd} \cup SV^{out} \quad (36)$$

Definition (Διανύσματα Υποστήριξης εκτός ορίων της υπολογιζόμενης υπερσφαίρας)

$$SV^{out} = \{\mathbf{x}_i \in \hat{\mathbf{X}}_{\mathbf{T}} : \alpha_i^* = C \text{ so that } \|\mathbf{x}_i - \mathbf{a}\|^2 > R^2\} \quad (37)$$

Definition (Διανύσματα Υποστήριξης στην επιφάνεια της υπολογιζόμενης υπερσφαίρας)

$$SV^{bnd} = \{\mathbf{x}_i \in \hat{\mathbf{X}}_{\mathbf{T}} : \alpha_i^* \in (0, C) \text{ so that } \|\mathbf{x}_i - \mathbf{a}\|^2 = R^2\} \quad (38)$$

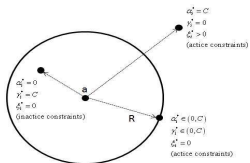


Figure: Support Vector Data Description.

Definition (Περιγραφή Δεδομένων με Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Data Description))

$$f_{SVDD}(\mathbf{z}; \mathbf{a}, R) = I(\|\mathbf{z} - \mathbf{a}\|^2 \leq R^2) \quad (39a)$$

$$= I\left(\langle \mathbf{z}, \mathbf{z} \rangle - 2 \sum_{i=1}^l \alpha_i \langle \mathbf{z}, \mathbf{x}_i \rangle + \sum_{i=1}^l \sum_{j=1}^l \alpha_i \alpha_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle\right) \quad (39b)$$

Αναδιτύπωση Προβλήματος I

Αναδιτύπωση του προβλήματος της Μονοταξικής Ταξινόμησης μέσα στο πλαίσιο των Αλγορίθμων της Τεχνητής Αρνητικής Επιλογής (Artificial Negative Selection Algorithms).

- Οι Αλγόριθμοι Τεχνητής Αρνητικής Επιλογής αποτελούν υπολογιστικές διαδικασίες οι οποίες μιμούνται τη διαδικασία της φυσικής αρνητικής επιλογής η οποία πραγματοποιείται στα ανοσοποιητικά συστήματα των σπονδυλωτών οργανισμών.
- Οι Αλγόριθμοι Τεχνητής Αρνητικής Επιλογής προτείνουν ένα εναλλακτικό υπόδειγμα Μονοταξικής Ταξινόμησης:
 - Δεν στοχεύουν στην δημιουργία ενός κλειστού συνόρου γύρω από το σύνολο των προτύπων της θετικής κλάσης.
 - Αντίθετα στοχεύουν στη δημιουργία ενός συνόλου μεταβλητών ανιχνευτών οι οποίοι θα καλύπτουν ένα επιθυμητό μέρος του υποσυνόλου των αρνητικών προτύπων.
- Ο θεωρούμενος χώρος των σχημάτων $\mathbb{S} = [0, 1]^{30}$ διαμερίζεται στα υποσύνολα των θετικών (\mathcal{S}) και αρνητικών (\mathcal{N}) προτύπων κατά τέτοιο τρόπο ώστε:
 - $\mathbb{S} = \mathcal{S} \cup \mathcal{N}$ και $\mathcal{S} \cap \mathcal{N} = \emptyset$
 - Το υποσύνολο \mathcal{S} αντιστοιχεί στην κλάση ενδιαφέροντος (κλάση πλειοψηφίας)
 - Το υποσύνολο \mathcal{N} αντιστοιχεί στην συμπληρωματική κλάση (κλάση μειοψηφίας)

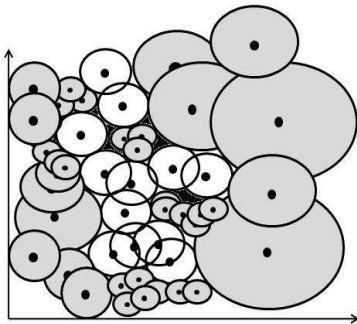
Αναδιατύπωση Προβλήματος II

- Ένας ανιχνευτής του αρνητικού χώρου των προτύπων αντιστοιχεί σε ένα σημείο ($\mathbf{d} \in \mathcal{N}$), στο οποίο αντιστοιχεί μια ακτίνα αναγνώρισης ($r(\mathbf{d}) \in [0, 1]$), έτσι ώστε να αναγνωρίζει οποιοδήποτε πρότυπο του χώρου \mathbb{S} που βρίσκεται εντός της περιοχής αναγνώρισης:

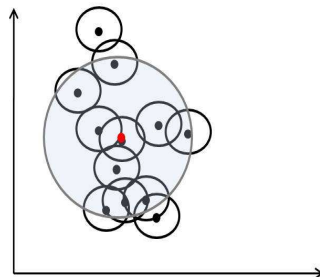
$$\mathcal{R}(\mathbf{d}, r(\mathbf{d})) = \{\mathbf{x} \in \mathbb{S} : \mathbf{D}(\mathbf{x}, \mathbf{d}) \leq r(\mathbf{d})\} \quad (40)$$

- Σε κάθε ανιχνευτή αντιστοιχεί μια διαφορετική ακτίνα αναγνώρισης.
- Σε κάθε σημείο του θετικού χώρου των προτύπων ($\mathbf{s} \in \mathcal{S}$) αντιστοιχεί η ίδια ακτίνα αναγνώρισης R_{self} ($\forall \mathbf{s} \in \mathcal{S}, r(\mathbf{s}) = R_{self}$).
- Η χρήση ανιχνευτών μεταβλητού μεγέθους δίνει τη δυνατότητα ελαχιστοποίησης του σφάλματος E_I χωρίς να συνεπάγεται την ταυτόχρονη αύξηση του σφάλματος E_{II} σε αντίθεση με την σφαιρική περιγραφή των Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης.

Αναδιατύπωση Προβλήματος III



(a) Ανιχνευτές Μεταβλητού Μεγέθους.



(b) Σφαιρική Περιγραφή Δεδομένων.

Figure: Σύγκριση Μονοταξικών Ταξινομητών

Επισκόπηση Αλγορίθμου

Definition (Τεχνητή Αρνητική Επιλογή)

Έστω $\mathcal{S} = \{s_1, \dots, s_n\} \subset \mathcal{S}$ ένα υποσύνολο του χώρου των θετικών προτύπων κάθε σημείο του οποίου είναι συσχετισμένο με μια σταθερή ακτίνα αναγνώρισης R_{self} . Στόχος του αλγορίθμου της αρνητικής επιλογής είναι παραγωγή ενός συνόλου ανιχνευτών $\mathcal{D} = \{d_1, \dots, d_m\} \subset \mathcal{N}$ για τον χώρο των αρνητικών προτύπων και αντίστοιχου συνόλου $R = \{r(\mathbf{d}) : \mathbf{d} \in \mathcal{D}\}$ ακτινών αναγνώρισης έτσι ώστε να ικανοποιούνται οι παρακάτω περιορισμοί:

1

$$\forall \mathbf{d} \in \mathcal{D}, \mathbf{d} \notin \mathcal{R}_{\mathcal{S}} \quad (41)$$

όπου $\mathcal{R}_{\mathcal{S}}$ είναι η περιοχή του χώρου των σχημάτων που αναγνωρίζεται από το υποσύνολο των θετικών προτύπων εκπαίδευσης έτσι ώστε:

$$\mathcal{R}_{\mathcal{S}} = \bigcup_{s \in \mathcal{S}} \mathcal{R}(s, r(s)) \quad (42)$$

2

$$P(\exists \mathbf{n} \in \mathcal{N} : \mathbf{n} \notin \mathcal{R}_{\mathcal{D}}) = 1 - C_0 \quad (43)$$

όπου $\mathcal{R}_{\mathcal{D}}$ η περιοχή του χώρου των σχημάτων που αναγνωρίζεται από το σύνολο των ανιχνευτών ως αρνητική έτσι ώστε:

$$\mathcal{R}_{\mathcal{D}} = \bigcup_{\mathbf{d} \in \mathcal{D}} \mathcal{R}(\mathbf{d}, r(\mathbf{d})) \quad (44)$$

και

$$C_0 = \frac{|\mathcal{D}|}{|\mathcal{N}|} \quad (45)$$

είναι είναι η εκτιμώμενη κάλυψη του αρνητικού χώρου των προτύπων.

Παράμετροι Εισόδου Αλγορίθμου Τεχνητής Αρνητικής Επιλογής

Ο αλγόριθμος της Τεχνητής Αρνητικής Επιλογής δέχεται ως είσοδο τις παρακάτω παραμέτρους:

- R_{self} : ακτίνα αναγνώρισης των θετικών προτύπων εκπαίδευσης.
- C_{self} : εκτιμώμενη κάλυψη του χώρου των θετικών προτύπων.
- C_0 : εκτιμώμενη κάλυψη του αρνητικού χώρου των προτύπων.
- T_{max} : μέγιστο πλήθος ανιχνευτών.

Αλγόριθμος Αρνητικής Επιλογής I

Ο αλγόριθμος της αρνητικής επιλογής βασίζεται στην επαναλαμβανόμενη ομοιόμορφη δειγματοληψία του χώρου των σχημάτων \mathbb{S} μέχρι να αληθεύσει μια από τις παρακάτω συνθήκες:

- 1 Το πλήθος των ανιχνευτών που έχουν παραχθεί ξεπερνά το μέγιστο επιθυμητό πλήθος T_{max} .
- 2 Η εκτιμώμενη κάλυψη του θετικού χώρου των προτύπων ξεπερνά το μέγιστη επιτρεπόμενη τιμή C_{self} .
- 3 Η εκτιμώμενη κάλυψη του αρνητικού χώρου των προτύπων ξεπερνά την ελάχιστη επιδιωκόμενη τιμή C_0 .

Η εκτίμηση του ποσοστού κάλυψης ενός συγκεκριμένου υποχώρου V (αρνητικού ή θετικού) του χώρου των σχημάτων \mathbb{S} ακολουθεί την παρακάτω συλλογιστική:

- Κατά την ομοιόμορφη δειγματοληψία k σημείων $\{x_1, \dots, x_k\}$ από το χώρο των σχημάτων \mathbb{S} , η ύπαρξη ενός και μοναδικού σημείου τέτοιου ώστε $x_i \notin V$ συνεπάγεται πως το εκτιμώμενο ποσοστό κάλυψης του υποχώρου V θα είναι $a = 1 - \frac{1}{k}$.
- Η κάλυψη του υποχώρου V κατά ένα ποσοστό a προϋποθέτει την δειγματοληψία $k = \frac{1}{1-a}$ σημείων του χώρου των σχημάτων \mathbb{S} έτσι ώστε ένα το πολύ σημείο να μην ανήκει στον υποχώρο V .

Αλγόριθμος Αρνητικής Επιλογής II

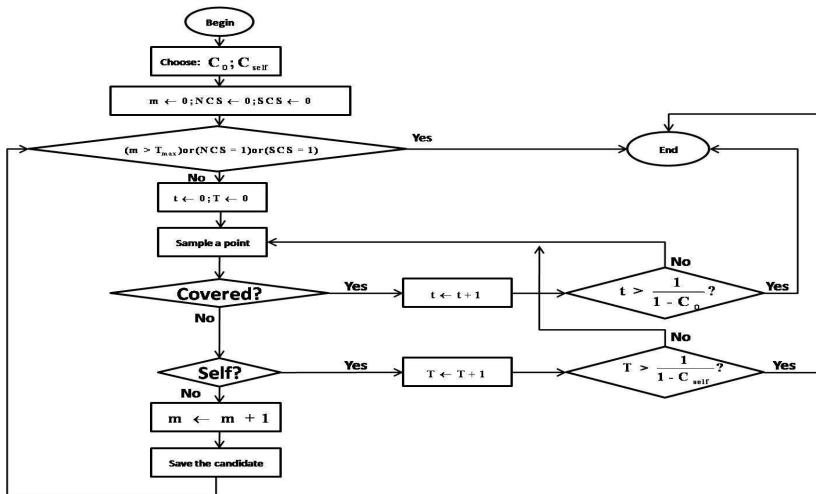


Figure: Αλγόριθμος Τεχνητής Αρνητικής Επιλογής

Επισκόπηση Συστήματος I

- 16 Χρήστες
- Βαθμοί προτίμησης $R \in \{1, 2, 3\}$.
 - 0 για αρχεία χωρίς βαθμό προτίμησης.
 - Για τους σκοπούς του πειράματος ζητήθηκε από τους χρήστες να αξιολογήσουν όλα τα αρχεία, γι' αυτό αρχεία χωρίς βαθμό προτίμησης θεωρούνται ως αρχεία εκτός κλάσης ενδιαφέροντος.
- Μέσος αριθμός προτιμήσεων ανά χρήστη ≈ 300 αρχεία.
- Έστω $U = \{u_1, \dots, u_{16}\}$ το σύνολο των χρηστών του συστήματος.
- Έστω $R(u, s)$ ο βαθμός προτίμησης που αποδίδει ο χρήστης $u \in U$ στο αντικείμενο (διάλυσμα χαρακτηριστικών) $s \in \mathbb{S}$.
- Οι προτιμήσεις κάθε χρήστη $u \in U$ ορίζουν μια διαμέριση του χώρου των σχημάτων \mathbb{S} έτσι ώστε:

$$\begin{aligned}
 C_0(u) &= \{s \in \mathbb{S} : R(u, s) = 0\} \\
 C_1(u) &= \{s \in \mathbb{S} : R(u, s) = 1\} \\
 C_2(u) &= \{s \in \mathbb{S} : R(u, s) = 2\} \\
 C_3(u) &= \{s \in \mathbb{S} : R(u, s) = 3\}
 \end{aligned}
 \tag{46}$$

και

$$\forall u \in U, \mathbb{S} = \mathbf{P}(u) \cup \mathbf{N}(u)
 \tag{47}$$

Επισκόπηση Συστήματος II

με

$$\begin{aligned} \mathbf{P}(u) &= C_1(u) \cup C_2(u) \cup C_3(u) \\ \mathbf{N}(u) &= C_0(u) \end{aligned} \quad (48)$$

- Σκοπός του προτεινόμενου Συστήματος Σύστασης είναι η υλοποίηση μιας συνάρτησης διάκρισης $\hat{R}(u, s)$ η οποία θα προβλέπει την προτίμηση του χρήστη $u \in U$ για το αντικείμενο $s \in \mathbb{S}$.
- Για το σκοπό αυτό το προτεινόμενο σύστημα υιοθετεί μια μεθοδολογία καταρράκτη η οποία αποσυνθέτει το πρόβλημα της σύστασης σε δύο επιμέρους επίπεδα:
 - Στο πρώτο επίπεδο πραγματοποιείται η διάκριση μεταξύ θετικών ($\mathbf{P}(u)$) και αρνητικών προτύπων ($\mathbf{N}(u)$) μέσω της συνάρτησης διάκρισης $f(u, s)$ ενός μονοταξικού ταξινομητή.
 - Στο δεύτερο επίπεδο πραγματοποιείται η διάκριση μεταξύ των θετικά αναγνωρισμένων προτύπων μέσω της συνάρτησης διάκρισης $g(u, s)$ ενός πολυταξικού ταξινομητή.
- Το προτεινόμενο Σύστημα Σύστασης υλοποιεί τη συνάρτηση διάκρισης $\hat{R}(u, s)$ μέσω των επιμέρους συναρτήσεων διάκρισης $f(u, s) \in \{-1, +1\}$ και $g(u, s) \in \{1, 2, 3\}$ έτσι ώστε:

$$\hat{R}(u, s) = \begin{cases} 0, & f(u, s) = +1; \\ g(u, s), & f(u, s) = -1. \end{cases} \quad (49)$$

Αρχιτεκτονική Συστήματος

- 1^ο Επίπεδο (Μονοταξική Ταξινόμηση): Αλγόριθμος Τεχνητής Αρνητικής Επιλογής vs Μονοταξικών Μηχανών Διανυσμάτων Υποστήριξης.
- 2^ο Επίπεδο (Πολυταξική Ταξινόμηση): Πολυταξικές Μηχανές Διανυσμάτων Υποστήριξης.

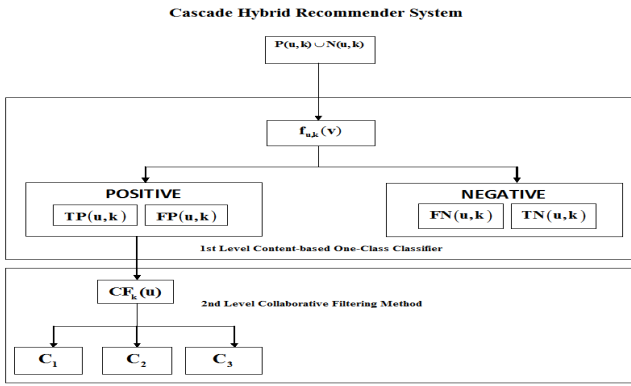


Figure: Cascade Hybrid Recommender.

Αξιολόγηση Πρώτου Επίπεδου Μεθοδολογίας Καταρράκτη Ι

Η αξιολόγηση του 1^{ου} επίπεδου περιλαμβάνει τη συγκριτική μελέτη των χρησιμοποιούμενων Μονοταξικών Ταξινομητών στη βάση των παρακάτω μέτρων:

- **MAE** (*Mean Absolute Error*): Το *Μέσο Απόλυτο Σφάλμα* αποτελεί το πιο σύνηθες μέτρο αξιολόγησης της αποτελεσματικότητας ενός Συστήματος Σύστασης.
- Το MAE που αφορά ένα συγκεκριμένο χρήστη κατά το k -οστό βήμα της διαδικασίας 10-πλής διεπικύρωσης δίνεται από τη σχέση:

$$MAE_k(u, s) = \frac{1}{|P(u, k)| + |N(u, k)|} \sum_{s \in P(u, k) \cup N(u, k)} |R_k(u, s) - \hat{R}_k(u, s)| \quad (50)$$

- **Precision**: Το μέτρο της *Ακρίβειας* εκφράζει την ποσότητα της πληροφορίας που δεν χάνεται κατά τη διαδικασία της σύστασης νέων πολυμεσικών αντικειμένων και δίνεται από την παρακάτω σχέση:

$$\overline{Precision} = \frac{\overline{TP}}{\overline{TP} + \overline{FP}} \quad (51)$$

Αξιολόγηση Πρώτου Επίπεδου Μεθοδολογίας Καταρράκτη II

- **Recall:** Το μέτρο της *Ανάκλησης Πληροφορίας* εκφράζει την ποσότητα των δεδομένων που δεν χάνονται κατά τη διαδικασία της σύστασης νέων πολυμεσικών αντικειμένων και δίνεται απο την παρακάτω σχέση:

$$\overline{Recall} = \frac{\overline{TP}}{\overline{TP} + \overline{FN}} \quad (52)$$

- **F-Measure:** Το συγκεκριμένο μέτρο αποτελεί συνδυασμό των μέτρων του Precision και του Recall και δίνεται απο την παρακάτω σχέση:

$$\overline{F1} = \frac{2 \times \overline{Precision} \times \overline{Recall}}{\overline{Precision} + \overline{Recall}} \quad (53)$$

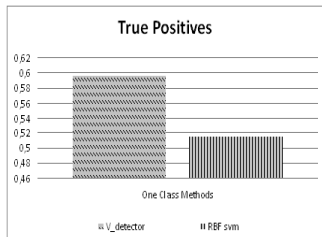
- **Quality Rate:** Το συγκεκριμένο μέτρο αποτελεί συνδυασμό των F-Measure και MAE και δίνεται απο την παρακάτω σχέση:

$$QualityRate = \frac{\overline{F1}}{\overline{MAE}} \quad (54)$$

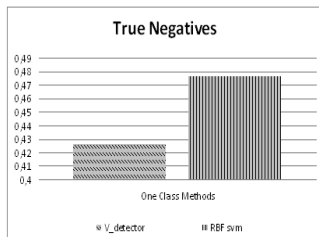
- Υψηλότερες τιμές για τις ποσότητες του Precision και του Recall υποδηλώνουν μεγαλύτερη ταξινομητική ακρίβεια.

Αξιολόγηση Πρώτου Επίπεδου Μεθοδολογίας Καταρράκτη III

- Οι τιμές του F-Measure ανήκουν στο διάστημα $[0, 1]$, όπου τιμές κοντά στο 0 υποδηλώνουν τη χειρότερη δυνατή συμπεριφορά του συστήματος ενώ τιμές κοντά στο 1 υποδηλώνουν τη καλύτερη δυνατή συμπεριφορά του συστήματος.
- Υψηλότερες τιμές για την ποσότητα του Quality Rate υποδηλώνουν καλύτερη συμπεριφορά του συστήματος καθώς συνεπάγονται αύξηση του F-Measure και μείωση του MAE.

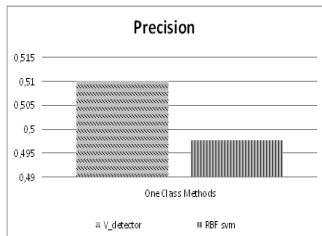


(a) True Positives

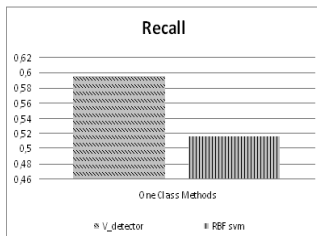


(b) True Negatives

Αξιολόγηση Πρώτου Επίπεδου Μεθοδολογίας Καταρράκτη IV

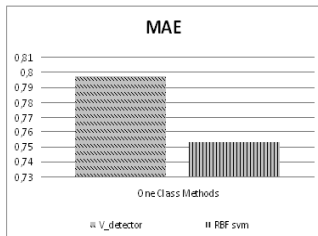


(c) Precision.eps

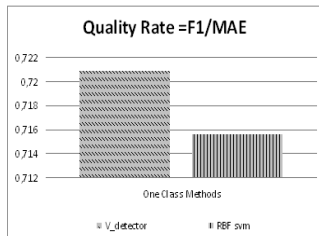


(d) Recall

Αξιολόγηση Πρώτου Επίπεδου Μεθοδολογίας Καταρράκτη V



(e) MAE



(f) Quality Rate

Αξιολόγηση Δεύτερου Επιπέδου Μεθοδολογίας Καταρράκτη I

- Η αξιολόγηση του 2^{ου} επιπέδου της μεθοδολογίας καταρράκτη προκύπτει από την εκτίμηση του συνολικού MAE του συστήματος σε σχέση με τη μεγαλύτερη και τη μικρότερη δυνατή τιμή του κατά τη φάση της πολυταξικής ταξινόμησης.
- Η μικρότερη δυνατή τιμή του MAE προκύπτει θεωρώντας πως το δεύτερο επίπεδο πολυταξικής ταξινόμησης ταξινομεί σωστά το σύνολο των ορθά αναγνωρισμένων θετικών προτύπων στο πρώτο επίπεδο.

$$\min_{u \in U} \overline{MAE} = \frac{1}{|U|} \sum_{u \in U} \frac{\overline{FNR}(u) \times \lambda(u)}{\lambda(u)+1} + \frac{\overline{FPR}(u)}{\lambda(u)+1} \quad (55)$$

- Η μεγαλύτερη δυνατή τιμή του MAE προκύπτει θεωρώντας πως το δεύτερο επίπεδο πολυταξικής ταξινόμησης αποτυγχάνει να ταξινομήσει σωστά το σύνολο των ορθά αναγνωρισμένων θετικών προτύπων στο πρώτο επίπεδο.

$$\max_{u \in U} \overline{MAE} = \frac{1}{|U|} \sum_{u \in U} \frac{3 \times \overline{FNR}(u) \times \lambda(u)}{\lambda(u)+1} + \frac{2 \times \overline{TPR}(u) \times \lambda(u)}{\lambda(u)+1} + \frac{1}{|U|} \sum_{u \in U} \frac{3 \times \overline{FPR}(u)}{\lambda(u)+1} \quad (56)$$

με

$$\lambda(u) = \frac{|P(u)|}{|N(u)|} \quad (57)$$

Αξιολόγηση Δεύτερου Επιπέδου Μεθοδολογίας Καταρράκτη II

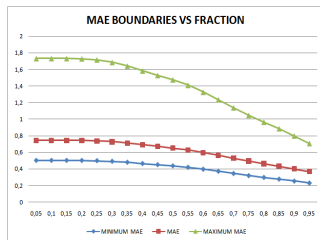


Figure: MAE Boundaries for One Class SVM.

Η συγκεκριμένη καμπύλη προκύπτει για διάφορες τιμές της παραμέτρου $v \in [0, 1]$ η οποία εκφράζει το ποσοστό των θετικών προτύπων εκπαίδευσης που προσπαθεί να συμπεριλάβει ο Μονοταξικός Ταξινομητής Μηχανών Διανυσμάτων Υποστήριξης.

Συμπεράσματα

- Ο αλγόριθμος της Τεχνητής Αρνητικής Επιλογής υπερτερεί της αντίστοιχης Μονοταξικής Μηχανής Διανυσμάτων Υποστήριξης ως προς τα μέτρα:
 - Precision
 - Recall
 - F-Measure και
 - Quality Rate
- Η Μονοταξική Μηχανή Διανυσμάτων Υποστήριξης υπερτερεί ως προς το μέτρο του MAE.
- Η Μονοταξική Μηχανή Διανυσμάτων Υποστήριξης στο 1^ο επίπεδο έχει την ικανότητα ορθής αναγνώρισης μεγαλύτερου μέρους του αρνητικού χώρου των προτύπων όπως προκύπτει από την τιμή του *True Negative Rate*.
- Η Μονοταξική Μηχανή Διανυσμάτων Υποστήριξης φιλτράρει μεγαλύτερο μέρος των μη-επιθυμητών πολυμεσικών αντικειμένων.
- Ο αλγόριθμος της Τεχνητής Αρνητικής Επιλογής αντίθετα έχει την ικανότητα να αναγνωρίζει μεγαλύτερο του θετικού χώρου των προτύπων όπως προκύπτει από την τιμή του *True Positive Rate*.
- Μεγαλύτερες τιμές των παραμέτρων *True Positive Rate*, *Precision*, *Recall*, *F-Measure* και *Quality Rate* υποδηλώνουν την ικανότητα του αλγορίθμου της Τεχνητής Αρνητικής Επιλογής να παρέχει εγκυρότερες προτάσεις σε σχέση με την Μονοταξική Μηχανή Διανυσμάτων Υποστήριξης.

Συνεισφορά Διατριβής

Η παρούσα διδακτορική διατριβή αναπτύσσει ένα εναλλακτικό υπόδειγμα Μηχανικής Μάθησης βασισμένο στο γενικότερο πλαίσιο των Τεχνητών Ανοσοποιητικών Συστημάτων.

Ανάπτυξη Ανοσοποιητικών Αλγορίθμων για τα βασικότερα προβλήματα της Μηχανικής Μαθησης.

Μαθηση Χωρίς Επιτήρηση:

- Αλγόριθμος Εκπαίδευσης ενός Τεχνητού Ανοσοποιητικού Δικτύου για την Ομαδοποίηση ενός συνόλου Μουσικών Δεδομένων.

Μαθηση Με Επιτήρηση:

- Αλγόριθμος Εκπαίδευσης ενός Τεχνητού Ανοσοποιητικού Συστήματος Αναγνώρισης για το πρόβλημα της Αυτόματης Ταξινόμησης ενός συνόλου Μουσικών Δεδομένων ως προς την Μουσικολογική Κλάση από την οποία προέρχονται.
 - *Ισοζυγισμένα Προβλήματα Ταξινόμησης:* Σύγκριση του προτεινόμενου ταξινομητή ενάντια σε κορυφαίες μεθοδολογίες ταξινόμησης όπως οι Μηχανές Διανυσμάτων Υποστήριξης (Support Vector Machines).
 - *Ετεροβαρή Προβλήματα Ταξινόμησης:* Μελέτη της συμπεριφοράς του προτεινόμενου ταξινομητή σε προβλήματα ακραίας Ταξικής Ανισορροπίας.
- Αλγόριθμος Εκπαίδευσης ενός Ανοσοποιητικού Μονοταξικού Ταξινομητή. Το πρόβλημα της Σύστασης Μουσικών Δεδομένων με βάση τις προτιμήσεις ενός συγκεκριμένου χρήστη διατυπώνεται ως ένα Πρόβλημα Μονοταξικής Ταξινόμησης.

Μελλοντική Εργασία

Ανάπτυξη αλγορίθμων εκπαίδευσης Ανοσοποιητικών Πολυταξικών Ταξινομητών με αναγωγή του προβλήματος των N -κλάσεων σε:

- $\binom{N}{2}$ ισοζυγισμένα δυαδικά προβλήματα ταξινόμησης.
- N One vs All ισοζυγισμένα δυαδικά προβλήματα ταξινόμησης.
- N One vs All μη-ισοζυγισμένα προβλήματα μονοταξικής ταξινόμησης.

Ανάπτυξη στρατηγικών συνδυασμού των επιμέρους ταξινομητών βασισμένων στη Θεωρία Παιγνίων.







Δημοσιεύσεις - Ετεροαναφορές

Ερευνητικά Άρθρα σε Διεθνή Περιοδικά μετά από Κρίση (2)

Κεφάλαια σε Βιβλία μετά από Κρίση (5)

Άρθρα σε Πρακτικά Διεθνών Συνεδρίων μετά από Κρίση Πλήρους Άρθρου (11)

Ετερο-αναφορές (13)

-  Chan, P. and Stolfo, S. J. (1998).
Toward scalable learning with non-uniform class and cost distributions: A case study in credit card fraud detection.
In In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining, pages 164–168. AAAI Press.
-  Ezawa, K., Singh, M., and Norton, S. W. (1996).
Learning goal oriented bayesian networks for telecommunications risk management.
In In Proceedings of the 13th International Conference on Machine Learning, pages 139–147. Morgan Kaufmann.
-  Farmer, J. D., Packard, N. H., and Perelson, A. S. (1986).
The immune system, adaptation, and machine learning.
Physica, 22D:187–204.
-  Fawcett, T. and Provost, F. (1996).
Combining data mining and machine learning for effective user profiling.
pages 8–13. AAAI Press.
-  Japkowicz, N. (2000).
Learning from imbalanced data sets: A comparison of various strategies.
pages 10–15. AAAI Press.
-  Japkowicz, N. and Stephen, S. (2002).
The class imbalance problem: A systematic study.
Intell. Data Anal., 6(5):429–449.