Ανάλυση δεδομένων απο χρονοσειρές

Ιωάννα Παναγιωτοπούλου

Διατριβή που υποβλήθηκε σε μερική
εκπλήρωση των απαιτήσεων για το πτυχίο
του τμήματος Μαθηματικών

Πανεπιστήμιο Αιγαίου

Σάμος, Καρλόβασι, Ελλάδα.

Οκτώβρης, 2018

Τριμελής επιτροπή:

Νάστου Παναγιώτης

Ευστάθιος Σταματάτος

Ανδρέας Παπασαλούρος

Περίληψη

Η ανάλυση δεδομένων απο χρονοσειρές έχει κάνει αισθητή την εμφάνιση της τα τελευταία χρόνια.Η αναπαράσταση εικόνας ως χρονοσειρά μπορεί να δώσει σημαντικά αποτελέσματα και να κάνει την ζωή μας πιο εύκολη. Αυτό αποτέλεσαι κεντρική πηγή έμπνευσης για την εκπόνιση αυτής της εργασίας. Ο κύριος στόχος της ειναι να επιδείξει τους αλγορίθμους που βοηθούν στην ολοκλήρωση αυτής της ιδέας, να εξηγήσει βασικούς ορισμούς. Επίσης, παρουσιάζονται παραδείγματα που έχει εφαρμοστεί αυτή η θεωρία. Αυτή η πτυχιακή γίνεται πιο συγκεκριμένη αναλύοντας το ανθρώπινο σώμα μέσα απο στάσεις της γιόγκα ώστε να εξάγει αποτελέσματα για το φύλο του ατόμου και αξιολογεί ποιός αλγόριθμος το πετυχαίνει με μεγαλύτερο ποσοστό επιτυχίας.

Summary

Data analysis from time series has made its appearance in recent years. Image representation as a time series can give significant results and make our lives easier. This is a central source of inspiration for the performance of this paper. Its main objective is to demonstrate the algorithms that help to complete this idea, to explain basic definitions. Also, there are examples that have been applied this theory. This diploma becomes more specific by analyzing the human body through yoga attitudes to extract results for the individual's gender and evaluating which algorithm succeeds with a greater success rate.

Ευχαριστίες

Το παρόν έγγραφο είναι αφιερωμένο στους αγαπημένους μου γονείς Νίκο και Κλυταιμνήστρα, στις αγαπημένες μου αδελφές Ζωή και Γεωργία.

Ευχαριστώ ιδιαιτέρως τον κύριο Μανώλη Μαραγκουδάκη, αναπληρώτη καθηγητή για την υποστήριξη του στην εκπόνηση της πτυχιακής μου εργασίας.

Πίνακας περιεχομένων

Ανάλυση δεδομένων απο χρονοσειρές 1

Περίληψη 3

Summary 4

Ευχαριστίες 5

1 Εισαγωγή 9

1.1 Εισαγωγή 9

1.2 Βασικό πλάνο 9

1.2.1 Ανάλυση του θέματος 9

1.2.2 Η σημασία αυτής της έρευνας 9

1.3 Παρόμοια εμπορικά συστήματα 10

1.4 Αναγνώριση κίνησης χεριού 10

1.5 Αναγνώριση της κίνησης του αυτοκινήτου και του περιβάλλοντος του 12

1.6 Δομή αυτής της εργασίας 14

2. Κεφάλαιο Σχετική εργασία 16

2.1 Σημαντικές έρευνες για τα πακέτα δεδομένων τους 15

2.1.1 Συμβολική Συνολική προσέγγιση (Symbolic Aggregate approximation) 15

2.1.2 Χαρακτηριστικά δυναμικής χρονικής στρέβλωσης (Dynamic Time Warping Features) 16

2.1.3 Σύνολο(σάκος) μοτίβων (Bag of Patterns) 17

2.1.4 Μαθηματικά Σχήματα στη πάροδο του χρόνου (Learning Time-series Shapelets) 19

2.1.5 Επεξεργασία χρονικής στρέβλωσης (Time Warp Edit) 20

2.1.6 Σάκος απο SFA Σύμβολα (Bag of SFA Symbols (BOSS)) 21

3 Θεωρητικό υπόβαθρο 25

3.1 Τι είναι η εξόρυξη δεδομένων (data mining) 25

3.2 Η ταξινόμιση (classification) 25
3.2.1 Οι φάσεις της ταξινόμισης 24
3.2.2 Τύποι απο τα αποτελέσματα της ταξινόμισης 27
3.3 Τα μοντέλα της ταξινόμισης 27
3.4 Αλγόριθμοι εξόριξης δεδομένων 29
3.4.1 Η επιλογή του καλύτερου αλγορίθμου 29
3.4.2 Οι 10 κάλυτεροι αλγόριθμοι 30
3.5 Τι είναι οι χρονοσειρές και η χρήση τους 32
3.6 Παραγείγματα γραφημάτων 33
4 Μετατροπή εικόνων σε χρονοσειρές για εξόριξη δεδομένων 36
4.1 Μετατροπή εικόνων σε χρονοσειρές για εξόρυξη δεδομένων 36
4.2 Πότε χρησιμοποιείται η ανάλυση απο χρονοσειρές 42
5 Η περιγραφή του συνόλου δεδομένων και η αξιολόγηση της επιτυχίας 43
5.1 Περιγραφή του συνόλου δεδομένων 43
5.2 Σύνολο Μοτίβων (Bag of Patterns) 43
5.3 Δυναμική χρονική στρέβλωση (Dynamic Time Warping) 44
Βιβλιογραφία 45

Πίνακας Εικόνων

Εικόνα 1.1: Η αναγνώριση του χεριού 11
Εικόνα 1.2: Η αναγνώριση του χεριού 12
Εικόνα 1.3: IMAPCAR..... 13
Εικόνα 1.4: Αναγνώριση εμπρόσθιου αυτοκινήτου 14
Εικόνα 2.1 : Αλγόριθμος SAX..... 17
Εικόνα 2.2: Αλγόριθμος στο DTW 18
Εικόνα 2.3:Αλγόριθμος σε BoP..... 19
Εικόνα 2.4: Αλγόριθμος LTS 21
Εικόνα 2.5: Αλγόριθμος TWE 22
Εικόνα 2.6: Αλγόριθμος BOSS.....24

Εικόνα 3.1: Χρυσό σε δολάρια 33
Εικόνα 3.2: Πωλήσεις υπολογιστών σε εκκατομύρια δολάρια για τα έτη 1986-2000 34
Εικόνα 3.3: Ο ρυθμός ανεργίας των ΗΠΑ..... 35
Εικόνα 4.1: Τα κρανία ως χρονοσειρά 36
Εικόνα 4.2: Η ανάλυση των κρανίων..... 37
Εικόνα 4.3: Η χρονοσειρά κάθε κρανίου..... 38
Εικόνα 4.4: Η ανάλυση του φύλλου..... 39
Εικόνα 4.5: Αλεξάνδρεια 40
Εικόνα 4.6: Χωρίς κλίση..... 40
Εικόνα 4.7: Η χρονοσειρά απο τη λέξη Αλεξάνδρεια..... 40
Εικόνα 4.8: Ευκλείδεια απόσταση 41
Εικόνα 4.9: DTW για χρονοσειρές..... 49
Εικόνα 5.1: Γράφημα αλγορίθμων.....45

Εικόνα 5.2:Γραφήματα BOSS-BoP.....45

Εικόνα 5.3: Συγκεντρωτικά αποτελέσματα.....45

1 Εισαγωγή

1.1 Εισαγωγή

1.2 Βασικό πλάνο

Στις επόμενες παραγράφους θα οριστεί η σημασία και η προσφορά της εύρεσης μοτίβου χρονοσειρών και θα παρασχεθούν κάποια παρόμοια συστήματα προκειμένου ο χρήστης να κατανοήσει το περιεχόμενο της διατριβής.

1.2.1 Ανάλυση του θέματος

Τα μοτίβα χρονοσειρών είναι τμήματα που επαναλαμβάνονται σε μακρά χρονική διάρκεια υπό την προϋπόθεση ότι υπάρχουν, παρέχουν ακριβείς πληροφορίες σχετικά με την υποκείμενη πηγή των χρονοσειρών. Από την αρχή της εξόρυξης δεδομένων, έχει προκαλέσει ενδιαφέρον, η ανακάλυψη του μοτίβου μιας εικόνας μέσω χρονοσειρών. Όταν τα δεδομένα είναι μεγάλα, οι πιθανότητες να επιτύχετε την ανακάλυψη μοτίβου είναι περισσότερες. Η ανάλυση χρονολογικών σειρών έχει ως στόχο να κατανοήσει τα πρότυπα που εξελίσσονται με την πάροδο του χρόνου και να τα χρησιμοποιήσει για να προβλέψει τη μελλοντική συμπεριφορά (μηνιαίες πωλήσεις, εβδομαδιαίες ποσότητες ER, καρδιακές αρρυθμίες, τιμές μετοχών ...). Τα μοντέλα χρονοσειρών είναι τα πιο απλά διαχρονικά μοντέλα και η διαμήκης μοντελοποίηση είναι σημαντική σε πολλούς τομείς, από τη μοντελοποίηση της επιχειρηματικής διαδικασίας έως την κατανόηση της εξέλιξης της νόσου ή των κοινωνικών διεργασιών μέχρι την πρόβλεψη του καιρού. Το ενδιαφέρον μας εστιάζεται στην ανάλυση μιας εικόνας σε χρονολογικές σειρές και στα δεδομένα που μπορούμε να εξάγουμε. Για παράδειγμα, θα θέλαμε να δούμε ένα πρόγραμμα που θα μετατρέψει την εικόνα σε μια χρονολογική σειρά και θα μπορεί να ταιριάζει κάθε χρονική σειρά με μόνο ένα αποτέλεσμα. Δίνοντας στο πρόγραμμα μια γυναίκα ή έναν άντρα μετά την ανάγνωση του σώματός τους, το μετατρέπουν σε μια χρονοσειρά.

1.2.2 Η σημασία αυτής της έρευνας

Η μετατροπή της εικόνας σε χρονοσειρά θα έχει κάποιο όφελος για το μέλλον μας; Αυτό είναι κάτι που ο χρόνος θα το αποδείξει, αλλά ήδη αποτελεί μέρος των πρόσφατων τεχνολογικών ανακαλύψεων. Από την πλευρά μας, αυτή η έρευνα θα βοηθήσει τους ανθρώπους να κατανοήσουν την αξία της ανάλυσης μιας εικόνας σε χρονοσειρές. Η έρευνα αυτή μπορεί να χρησιμοποιηθεί όχι μόνο για την

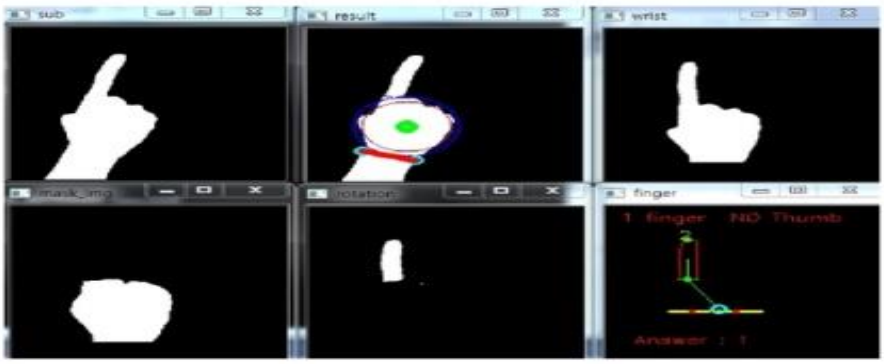
αναγνώριση του φύλου ή της στάσης στη γιόγκα, αλλά και σε οτιδήποτε μπορούμε να φανταστούμε.

1.3 Παρόμοια εμπορικά συστήματα

Αυτή η παράγραφος θα δώσει παραδείγματα δύο εμπορικών συστημάτων που έχουν παρόμοιο θέμα. Παίρνουν μια φωτογραφία και αναγνωρίζουν μερικά από τα χαρακτηριστικά της. Τα αποτελέσματα αυτής της διαδικασίας μας βοηθούν να κατανοήσουμε τα δεδομένα της εικόνας διαχωρίζοντας το σημείο στο οποίο θέλουμε να επικεντρωθούμε και του δίνουν έμφαση.

1.4 Αναγνώριση χειρονομίας

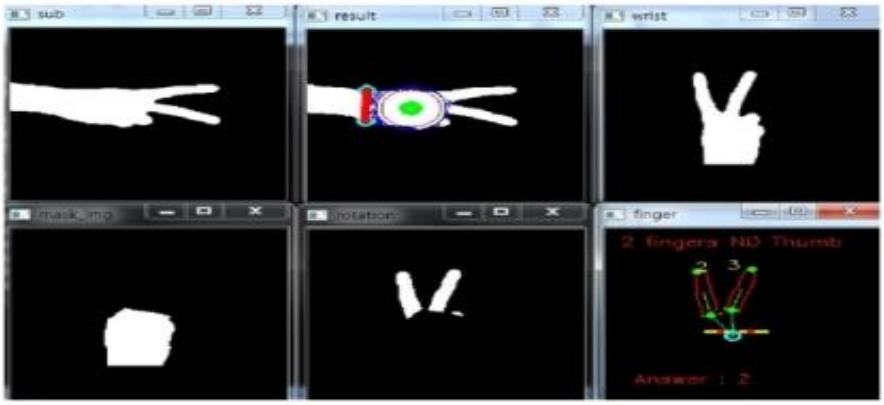
Η πρώτη έρευνα που θα περιγράψουμε αφορά την αναγνώριση χειρονομίας. Το 2014 στο περιοδικό "The Scientific World" ο Shan Zhao, δημοσίευσε το άρθρο που εξηγεί πώς αναγνωρίζει τη χειρονομία σε πραγματικό χρόνο. Ο τρόπος με τον οποίο αναγνωρίζεται η χειρονομία ακολουθεί τα επόμενα βήματα. Η περιοχή του χεριού ανιχνεύεται από το φόντο με τη μέθοδο αφαίρεσης υποβάθρου. Στη συνέχεια, η παλάμη και τα δάχτυλα είναι κατακερματισμένα. Με βάση την κατάτμηση, τα δάχτυλα στην εικόνα του χεριού ανακαλύπτονται και αναγνωρίζονται. Η αναγνώριση των χειρονομιών πραγματοποιείται με έναν απλό ταξινομητή κανόνων. Η απόδοση της μεθόδου μας αξιολογείται σε σύνολο δεδομένων 1300 εικόνων της χειρός. Τα πειραματικά αποτελέσματα δείχνουν ότι η προσέγγισή μας έχει καλές επιδόσεις και είναι κατάλληλη για εφαρμογές σε πραγματικό χρόνο. Η απόδοση της προτεινόμενης μεθόδου εξαρτάται σε μεγάλο βαθμό από το αποτέλεσμα της ανίχνευσης χεριών. Αν υπάρχουν κινούμενα αντικείμενα με χρώμα παρόμοιο με εκείνο του δέρματος, τα αντικείμενα υπάρχουν στο αποτέλεσμα της ανίχνευσης χεριών και στη συνέχεια υποβαθμίζουν την απόδοση της αναγνώρισης της χειρονομίας. Ωστόσο, οι αλγόριθμοι μηχανικής μάθησης μπορούν να διακρίνουν το χέρι από το φόντο. Ακολουθεί μια φωτογραφία που μας δείχνει τη διαδικασία που περιγράψαμε:



ΕΙΚΌΝΑ 1.1: Η αναγνώριση του χερίου

Η παραπάνω εικόνα αποτελεί μέρος της έρευνας Real-Time Hand Gesture Recognition Using Finger Segmentation

Στην παρακάτω εικόνα, τα στάδια της ανάλυσής της είναι προφανή.



ΕΙΚΌΝΑ 1.2: Η αναγνώριση του χεριού

Η εικόνα έχει παρουσιαστεί σε έρευνα με τίτλο: Real-Time Hand Gesture Recognition Using Finger Segmentation.

1.5 Αναγνώριση της κίνησης του αυτοκινήτου και του περιβάλλοντος του

Η ανάγκη για ασφαλή οδήγηση έχει αυξηθεί και έχει ως αποτέλεσμα την αναζήτηση συσκευών ασφαλείας που μπορούν να αναγνωρίσουν τις εικόνες έξω απο το αυτοκίνητο και να λειτουργήσουν όπως τα "μάτια του οχήματος", για αυτό το λόγο προσελκύουν ιδιαίτερη προσοχή. Το IMAPCAR είναι ένας επεξεργαστής που πληρεί τις προϋποθέσεις για εφαρμογές αναγνώρισης εικόνας αυτοκινήτου, επιτυγχάνοντας υψηλή απόδοση επεξεργασίας και χαμηλή ισχύ. Μπορεί να αναγνωρίσει τις ενέργειες που ακολουθούν:

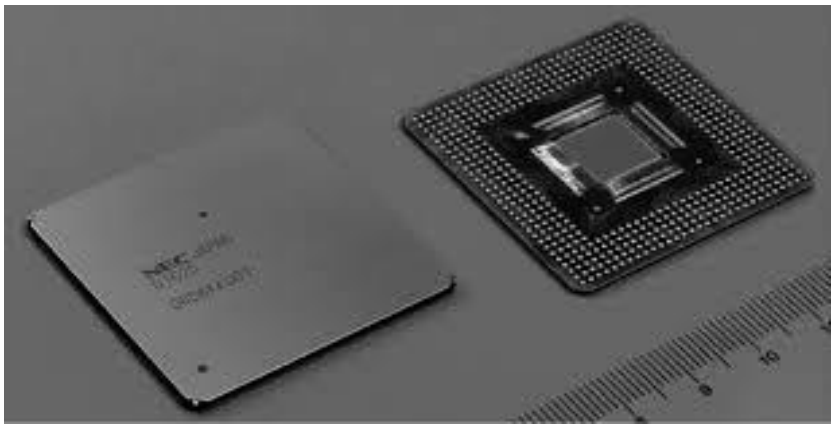
1. Αναγνώριση πεζών

2. Σύστημα διατήρησης λωρίδων

3. Αναγνώριση οχήματος προς τα εμπρός και προς τα πίσω

4. Αναγνώριση σημάτων κυκλοφορίας

Το IMAPCAR ορίζεται σήμερα ως συσκευή αναγνώρισης κίνησης αυτοκινήτου, οι απαιτήσεις σε αυτό το πεδίο είναι οι πιο προηγμένες. Παρόλα αυτά, η συσκευή μπορεί να χρησιμοποιηθεί για οποιοδήποτε σύστημα που απαιτεί αναγνωρίσεις κινητών εικόνων. Βλέπουμε πώς είναι οπτικά:



Εικόνα 1.3: IMAPCAR

Η εικόνα έχει παραχωρήθει απο την διεύθυνση www.nec.com .

Στην ακόλουθη εικόνα, θα παρατηρήσουμε μία από τις παραπάνω λειτουργίες που μπορεί να αναγνωρίσει αυτή η εφαρμογή:

ΕΙΚΌΝΑ 1.4: Αναγνώριση εμπρόσθιου αυτοκινήτου

Εικόνα απο την σελίδα www.nec.com .

1.6 Η δομή αυτής της εργασίας

Η έρευνα αυτή ακολουθεί την παρακάτω δομή:

1. Κεφάλαιο Εισαγωγή

   Αυτό το κεφάλαιο θα παρουσιάσει το θέμα και τη σημασία της έρευνας . Θα αναφερθεί σε δύο εφαρμογές που χρησιμοποιούν την αναγνώριση εικόνας για τα συμπεράσματά τους.

2. Κεφάλαιο Σχετική εργασία

Σε αυτό το σημείο παρουσιάζουμε τους πιο σημαντικούς αλγόριθμους που περιγράφουν τη χρήση τους και ένα μέρος του κώδικα τους.

3. Κεφάλαιο Μετατροπή εικόνων σε χρονοσειρές για εξόρυξη δεδομένων

4. Κεφάλαιο Θεωρητικό υπόβαθρο

5. Κεφάλαιο Η περιγραφή του συνόλου δεδομένων και η αξιολόγηση της επιτυχίας

2 Σχετική εργασία

2.1 Σημαντικές έρευνες για τα πακέτα δεδομένων τους

Σε αυτό το κεφάλαιο θα αναφερθώ σε μερικές έρευνες που έλαβαν χώρα τα τελευταία χρόνια καθώς και ορισμένες από αυτές είναι πολύ σημαντικές για τα πακέτα δεδομένων τους.

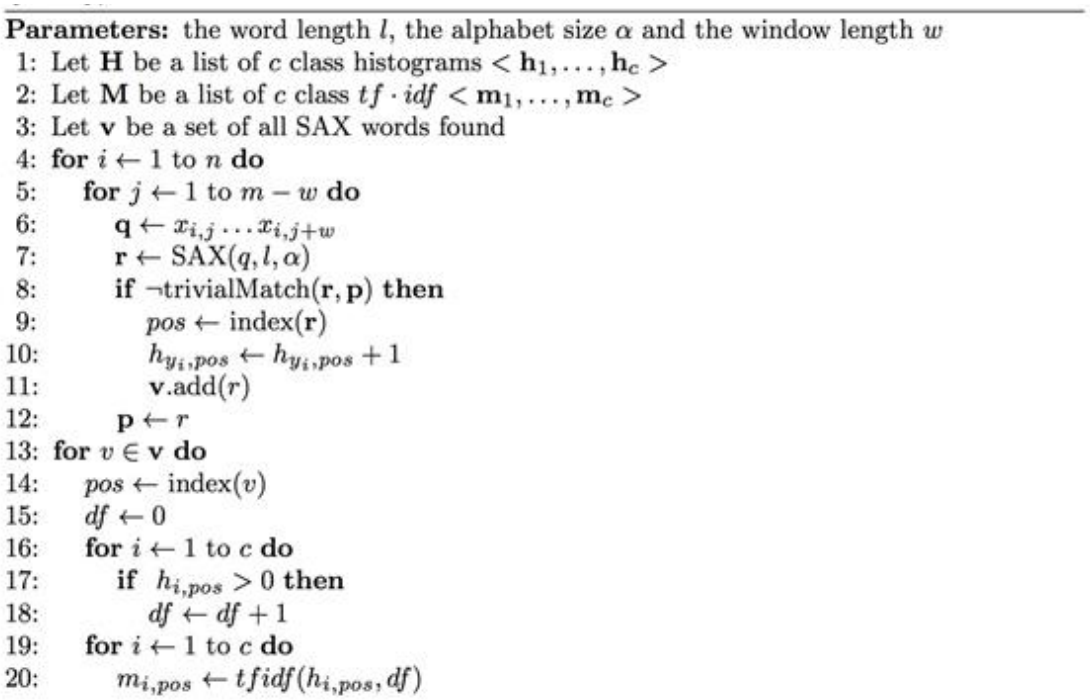
2.1.1 Συμβολική Συνολική προσέγγιση (Symbolic Aggregate approximation)

Ξεκινάμε με το γνωστο πακετο SAX (Symbolic Aggregate Approximation). Αυτή η συνάρτηση μετατρέπει μια αριθμητική σειρά χρονοσειρών σε μια σειρά γραμμάτων με συγκεκριμένο μήκος και αλφάβητο. Το SAX έχει αναπτυχθεί για να μειώσει τη διαστατικότητα μιας αριθμητικής σειράς σε μια σύντομη σειρά χαρακτήρων. Το SAX ακολουθεί μια διαδικασία δύο σταδίων: α) Συγκεντρωτική προσέγγιση στο περίπου (PAA) και β) μετατροπή μιας ακολουθίας PAA σε μια σειρά γραμμάτων. Αλλά τι γίνεται με το SAX και το Vector Space Model; Το SAXVSM συμφωνεί με την αναπαράσταση SAX που υπάρχει στο BOP με το μοντέλο διανυσματικού χώρου μόλις χρησιμοποιηθεί στην ανάκτηση πληροφοριών. Οι διαφορές μεταξύ BOP και SAXVSM είναι ότι το SAXVSM διαχωρίζει τις λέξεις σε κατηγορίες αντί για σειρές και τις εμπλουτίζει με τη συχνότητα αντίστροφων εγγράφων (tf·idf). Για το SAXVSM, η συχνότητα των όρων tf περιγράφει πόσες φορές εμφανίζεται μια λέξη σε μια τάξη και συχνότητα εγγράφων df που σημαίνει πόσες κλάσεις εμφανίζει μια λέξη. Θέτω ότι tf·idf είναι:

$$tfidf(tf, df) = \begin{cases} \log(1+tf)\log(c/df), \text{if } df>0 \\ 0, otherwise \end{cases}$$

Το c διογκώνει τον αριθμό των κλάσεων. Το SAX μετασχηματίζει μια χρονοσειρά X του μήκους l στη σειρά του αυθαίρετου μήκους w, όπου w τυπικά, χρησιμοποιεί ένα αλφάβητο A μεγέθους a> 2. Η μεθοδος αυτη στοχεύει στην εκτίμηση του μεγέθους του τμήματος και του μεγέθους του αλφαβήτου της Symbolic Aggregate Approximation (SAX). Στο SAX, τα δεδομένα χρονοσειρών χωρίζονται σε ένα σύνολο τμημάτων ίσου μεγέθους. Κάθε τμήμα αντιπροσωπεύεται από τη μέση τιμή του και χαρτογραφείται με ένα αλφάβητο, όπου ο αριθμός των υιοθετημένων συμβόλων ονομάζεται μέγεθος αλφάβητου. Και οι δύο παράμετροι ελέγχουν τη σχέση μεταξύ των δεδομένων και την ακρίβεια των εργασιών εξόρυξης χρονολογικών σειρών. Εκτός αυτού, η επιλογή βέλτιστων παραμέτρων εξαρτάται σε μεγάλο βαθμό από

διαφορετικά σύνολα εφαρμογών και δεδομένων. Στην πραγματικότητα, αυτές οι παράμετροι επιλέγονται διαδοχικά με ανάλυση ολόκληρων συνόλων δεδομένων, γεγονός που περιορίζει τη διαχείριση του τεράστιου αριθμού χρονολογικών σειρών και μειώνει την εφαρμοσιμότητα του SAX. Το μέγεθος του τμήματος εκτιμάται με βάση το θεώρημα δειγματοληψίας Shannon (autoSAXSD_S) και την προσαρμοστική ιεραρχική κατάτμηση (autoSAXSD_M). Όσον αφορά το μέγεθος του αλφαβήτου, εστιάζεται στο πώς κατανέμονται οι μέσες τιμές όλων των τμημάτων. Ο μικρός αριθμός του μεγέθους του αλφάβητου έχει οριστεί για μεγάλη διάρκεια για να διακρίνει εύκολα τη διαφορά μεταξύ των τμημάτων.Χρησιμοποιείται μια ταξινόμηση 1-NN με βάση το πόσες φορές εμφανίζονται οι λέξεις κάθε νέας περίπτωσης.



Parameters: the word length l, the alphabet size α and the window length w

Let $\mathbf{H}$ be a list of c class histograms $< \mathbf{h}_1, \ldots, \mathbf{h}_c >$
Let $\mathbf{M}$ be a list of c class $tf \cdot idf$ $< \mathbf{m}_1, \ldots, \mathbf{m}_c >$
Let $\mathbf{v}$ be a set of all SAX words found
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow 1$ to $m - w$ **do**
    $\mathbf{q} \leftarrow x_{i,j} \ldots x_{i,j+w}$
    $\mathbf{r} \leftarrow \mathrm{SAX}(q, l, \alpha)$
    if $\neg$trivialMatch$(\mathbf{r}, \mathbf{p})$ **then**
      $pos \leftarrow \mathrm{index}(\mathbf{r})$
      $h_{y_i,pos} \leftarrow h_{y_i,pos} + 1$
      $\mathbf{v}$.add(r)
    $\mathbf{p} \leftarrow r$
for $v \in \mathbf{v}$ **do**
  $pos \leftarrow \mathrm{index}(v)$
  $df \leftarrow 0$
  for $i \leftarrow 1$ to c **do**
    if $h_{i,pos} > 0$ **then**
      $df \leftarrow df + 1$
  for $i \leftarrow 1$ to c **do**
    $m_{i,pos} \leftarrow tfidf(h_{i,pos}, df)$

ΕΙΚΌΝΑ 2.1 : Αλγόριθμος SAX

Η εικόνα 2.1 προέρχεται απο έρευνα δημοσιευμένη με τίτλο Time Series Classification (http://timeseriesclassification.com/algorithmdescription.php?algorithm_id=9)

2.1.2 Χαρακτηριστικά δυναμικής χρονικής στρέβλωσης (Dynamic Time Warping Features)

Η μέθοδος DTWF(dynamic time warping features) αναμειγνύει τις αποστάσεις DTW σε δεδομένα εκπαίδευσης και ιστογράμματα SAX, επίσης, αναμιγνύει ολόκληρη τη σειρά και το λεξικό βασίζεται πιο κοντά σε ένα μόνο ταξινομητή. Μια προσομοίωση με n περιπτώσεις μεταβάλλεται πλήρως σε ένα σύνολο με n

χαρακτηριστικά, όπου το χαρακτηριστικό xij είναι το μέγιστο όριο DTW μεταξύ περίπτωσης i και περίπτωσης j. Συνεπώς, δημιουργούνται n λειτουργίες. Ανάμεσα στις περιπτώσεις, αυτές αποτελούν τις καλύτερες δυνατές αποστάσεις ορίων στο DTW. Χρησιμοποιώντας τον αλγόριθμο BOP για κάθε παράδειγμα, δημιουργούνται ιστογράμματα συχνότητας λέξεων SAX. Αυτά τα ειδικά χαρακτηριστικά συνδέονται με τα χαρακτηριστικά πλήρους και καλυτέρου DTW 2n. Το νέο σύνολο δεδομένων πραγματοποιείται με μια μηχανική συσκευή με πυρήνα πολυωνύμων με δομή είτε 1, 2 είτε 3, που καθορίζονται μέσω της ταυτότητας της αυθεντικότητας τους. Το μέγεθος του παραθύρου DTW και οι παράμετροι SAX ρυθμίζονται επίσης ξεχωριστά με διασταύρωση ταυτότητας με έναν ταξινομητή 1-NN. Οι παράμετροι που δέχεται ο κώδικας αυτός είναι: φορέα υποστήριξης (SVM) τάξης s, το μήκος της λέξης l, το μέγεθος του αλφαβήτου a, το μήκος του παραθύρου w καθως και το πλάτος του r.

Parameters: the SVM order s, SAX word length l, alphabet size α and window length w, DTW window width r
Let $\mathbf{Z}$ be a list of n cases of length $2n + a^l$, $\mathbf{z}_1 \ldots, \mathbf{z}_n$ initialised to zero.
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow i + 1$ to n **do**
    $z_{i,j} \leftarrow DTW(\mathbf{x}_i, \mathbf{x}_j)$
    $z_{j,i} \leftarrow z_{i,j}$
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow i + 1$ to n **do**
    $z_{i,n+j} \leftarrow DTW(\mathbf{x}_i, \mathbf{x}_j, r)$
    $z_{n+j,i} \leftarrow z_{i,n+j}$
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow 1$ to $m - w$ **do**
    $\mathbf{q} \leftarrow x_{i,j} \ldots x_{i,j+w}$
    $\mathbf{r} \leftarrow \text{SAX}(q, l, \alpha)$
    if $\neg$ trivialMatch$(\mathbf{r}, \mathbf{p})$ **then**
      $pos \leftarrow \text{index}(\mathbf{r})$
      $z_{i,2n+pos} \leftarrow z_{i,2n+pos} + 1$
    $\mathbf{p} \leftarrow \mathbf{r}$
SVM.buildClassifier$(\mathbf{Z}, s)$

ΕΙΚΌΝΑ 2.2: Αλγόριθμος στο DTW

Ο παραπάνω αλγόριθμος είναι κομμάτι εμπνευσμένο απο την ηλεκτρονική διεύθυνση www.timeseriesclassification.gr .

2.1.3 Σύνολο(σάκος) μοτίβων (Bag of Patterns)

Ο αλγόριθμος BoP(Bag of Patterns) είναι ένα λεξικό που κατηγοριοποιεί σε μέθοδο SAX όπως περιγράφηκε πιο πάνω. Η διάσταση της σειράς μειώνεται από το SAX με έναν τρόπο συγκεντρωτικής σύγκλισης κατά προσέγγιση (Piecewise Aggregate Approximation (PAA)), στη συνέχεια διαχωρίζει τη σειρά σε ¨αποθήκες¨ που σχηματίζονται από ίσα διαστήματα πιθανότητας της κανονικής κατανομής.Το SAX,

εξαπλώνεται σε κάθε παράθυρο για να δημιουργήσει μια λέξη στο BoP. Το Bag of Patterns αντιπροσωπεύει την εξόρυξη κειμένου και την προσέγγιση του υπολογιστή δομημένη έτσι για την εξαγωγή πληροφοριών υψηλής δομής από έγγραφο κειμένου. Η αναπαράσταση του BoP είναι αποτελεσματική για να ανιχνεύσει την ομοιότητα των μοτίβων σε διαφορετικές χρονολογικές σειρές. Το BoP χρησιμοποιείται στις χρονοσειρές για την ταξινόμηση, την ομαδοποίηση και την ανίχνευσης ανωμαλιών για να αγνοήσει τις επαναλαμβανόμενες πληροφορίες των μοτίβων και να μετρήσει την εμφάνιση των μοτίβων σε χρονοσειρές ώστε να παράγει ενα μοτίβο συχνοτήτων. Το BoP σαν ενα λεξικό που αποτελείται απο σύνολο κωδικών, παρουσιάζει μια χρονοσειρά ως ένα ιστόγραμμα του συνόλου κωδικών και υποδεικνύει τον αριθμό του οποίου συνέβη σε πολλές χρονικές σειρές. Στην πράξη, το μοντέλο αυτό χρησιμοποιείται κυρίως ως εργαλείο δημιουργίας χαρακτηριστικών. Αφού μετατρέψουμε το κείμενο σε "σακούλα λέξεων", μπορούμε να υπολογίσουμε διάφορα χαρακτηριστικά για να περιγράψουμε το κείμενο. Ο συνηθέστερος τύπος χαρακτηριστικών που υπολογίζονται από το μοντέλο BoP είναι η μακροχρόνια συχνότητα, δηλαδή ο αριθμός των φορών που εμφανίζεται ένας όρος στο κείμενο. Η χρήση του κώδικα BoP βοηθά στην ανίχνευση δομικών πληροφοριών σε χρονοσειρές καιρού. Τα δεδομένα μετατρέπονται σε μοτίβα συχνότητας. Τα μοτίβα δημιουργούνται μετατοπίζοντας την χρονοσειρά για να ανιχνεύσουν όλες τις υποπεριοχές σε ένα φάσμα μεγέθους w χρησιμοποιώντας την προσέγγιση των συρόμενων παραθύρων, τις ακολουθίες που θεωρούνται ως τοπικά τμήματα και στη συνέχεια τη μετατροπή των τοπικών τμημάτων σε μια αναπαράσταση SAX που χρησιμοποιεί ένα μέγεθος παραθύρου w για να εξαγάγει σχέδια σε τοπικές υπακολουθίες και μετρήσει τα εμφανιζόμενα πρότυπα σε χρονοσειρές για να δημιουργήσει μια πληροφορία συχνότητας. Ο αλγόριθμος χρησιμοποιεί τις εξής δύο παραμέτρους: η παράμετρος l περιγράφει το μήκος της λέξης, το μέγεθος του αλφαβήτου ειναι a.

Parameters: the word length l, the alphabet size α and the window length w

Let $\mathbf{H}$ be a list of n histograms $< \mathbf{h}_1, \ldots, \mathbf{h}_n >$
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow 1$ to $m - w$ **do**
    $\mathbf{q} \leftarrow x_{i,j} \ldots x_{i,j+w}$
    $\mathbf{r} \leftarrow \text{SAX}(q, l, \alpha)$
    if $\neg$ trivialMatch$(\mathbf{r}, \mathbf{p})$ **then**
      $pos \leftarrow$ index$(\mathbf{r})$ {*the function* index *determines the location of the word* $\mathbf{r}$ *in the count matrix* $\mathbf{h_i}$}
      $h_{i,pos} \leftarrow \mathbf{h}_{i,pos} + 1$
    $\mathbf{p} \leftarrow \mathbf{r}$

ΕΙΚΌΝΕΣ 2..3: Αλγόριθμος σε BoP

Η εικόνα 2.3 αποτελεί μέρος της ηλεκτρονικής σελίδας Time Series Classification.

2.1.4 Μαθηματικά σχήματα στη πάροδο του χρόνου (Learning Time-series Shapelets)

Ο αλγόριθμος LTS(Learning Time-series Shapelets), ως μέθοδος βασίζεται σε τοπικά χαρακτηριστικά για την ταξινόμηση των χρονοσειρών. Ο αλγόριθμος κατασκευάζει ένα δέντρο απόφασης όπου βρίσκει μορφές και στη συνέχεια χρησιμοποιεί τις χρήσιμες πληροφορίες των δυνητικών υποψηφίων που συνδέονται με όλες τις ακολουθίες των δεδομένων χρονοσειρών. Η μετατροπή σχημάτων είναι μια τεχνική που συνδυάζει σχήματα και μηχανική μάθηση. Οι συγγραφείς θεωρούν τα παραδείγματα χρονοσειρών ως διανύσματα που ορίζονται από το σύνολο ομοιοτήτων των σχημάτων για να μπορέσουν να υπαχθούν σε ταξινόμηση. Να σημειωθεί ότι ο μετασχηματισμός του σχήματος διαχωρίζει τελείως τη φάση που ψάχνει τα σχήματα από τις φάσεις δημιουργίας κανόνων ταξινόμησης. Στη συνέχεια, έχουν προταθεί πολλοί αλγόριθμοι για την αναζήτηση των καλών σχημάτων διατηρώντας υψηλή ακρίβεια προβλέψεων στην πράξη. Οι αλγόριθμοι βασίζονται στην ιδέα ότι τα διακριτικά σχήματα εμπεριέχονται στα δεδομένα κατάρτισης (training data). Ωστόσο, η παραπάνω προσέγγιση μπορεί να γίνει χωρίς τον αλγόριθμο ταξινόμισης. Ο αλγόριθμος LTS είναι μια διαφορετική προσέγγιση από αυτούς τους αλγορίθμους που βασίζονται σε υποεπιλογές. Το LTS, διαπιστώνει ότι οι μορφές που αποτελούν δευτερεύουσες σειρές στα εκπαιδευτικά δεδομένα (education data) δεν τους περιορίζουν. Χρησιμοποιώντας μια συλλογή k-δεδομένων απο τα υποψήφια δεδομένα εκπαίδευσης, υπολογίζονται τα σχήματα. Η συνάρτηση για τη διαδικασία επιλογής είναι μια λογική συνάρτηση απώλειας L που βασίζεται σε ένα μοντέλο λογικής παλινδρόμησης κάθε κατηγορίας. Αυτός ο αλγόριθμος δίνει μια οπτική υψηλού επιπέδου. Η αναζήτηση περιοριζόταν σε σχήματα με μήκος {Lmin, 2Lmin, ..., RLmin}. Όταν ολοκληρωθεί ο μισός αριθμός των επιτρεπόμενων επαναλήψεων, πραγματοποιείται έλεγχος. Αυτό ονομάζεται λάθος δεδομένων 1 (train set error 1) ή άπειρη απώλεια (infinite loss). Αυτό σημαίνει ότι το LTS δεν τερμάτισε ποτέ για κάποια προβλήματα. Για το λόγο αυτό, περιορίσαμε τον αλγόριθμο σε πέντε επανεκκινήσεις κατ 'ανώτατο όριο. Η παράμετρος K ορίζει τον αριθμό των σχημάτων, το ελάχιστο μήκος του σχήματος Lmin, η κλίμακα του μήκους των σχημάτων - η παράμετρος κανονικοποίησης ειναι R, το ποσοστό μάθησης αρχικοποιείται ώς λw, ο αριθμός των επαναλήψεων είναι η, α η μέγιστη παράμετρος.

Parameters: number of shapelets K, minimum shapelet length L^{min}, scale of shapelet length, R, regularization parameter, λ_W, learning rate, η, number of iterations, $maxIter$, and softmax parameter, α.

$\mathbf{S} \leftarrow$initializeShapeletsKMeans($\mathbf{T}$,K,R, L^{min})
$\mathbf{W} \leftarrow$initializeWeights($\mathbf{T}$,K,R)
for $i \leftarrow 1$ to $maxIter$ **do**
  $\mathbf{M} \leftarrow$ updateModel($\mathbf{T}, \mathbf{S}, \alpha, L^{min}, R$)
  $\mathbf{L} \leftarrow$ updateLoss($\mathbf{T}, \mathbf{M}, \mathbf{W}$)
  $\mathbf{W}, \mathbf{S} \leftarrow$ updateWandS($\mathbf{T}, \mathbf{M}, \mathbf{W}, \mathbf{S}, \eta, R, L^{min}, \mathbf{L}, \lambda_W, \alpha$)
  if diverged() **then**
    $i = 0$
    $\eta = \eta/3$

ΕΙΚΌΝΑ 2.4: Αλγόριθμος LTS

Η εικόνα ανήκει στην ιστοσελίδα Time Series Classification.

2.1.5 Επεξεργασία χρονικής στρέβλωσης (Time Warp Edit)

Ο Marteau (2009) προτείνει τη μέθοδο TWE (Time Warp Edit distance). Είναι ένα ελαστικό μέτρο απόστασης το οποίο, σε αντίθεση με το DTW είναι επίσης μετρικό. Επιτρέπει την κάμψη του σχήματος στον άξονα του χρόνου, περιλαμβάνει χαρακτηριστικά από το DTW και συνδυάζει την απόσταση επεξεργασίας με τα πρότυπα Lp (μετρική). Η παραποίηση του προκαθορισμένου σχήματος που ονομάζεται ακαμψία χειραγωγείται απο μια παράμετρο ν. Στο DTW ένα περιστρεφόμενο παράθυρο αναστέλλει την αναζήτηση, η ακαμψία επιβάλλει μια πολλαπλασιαστική ποινή στην απόσταση μεταξύ των αντιστοιχησμένων σημείων. Με τη μέτρηση απόστασης που είναι ισοδύναμη με μια πλήρης αναζήτηση DTW, η θέση ν = 0 δεν έχει ακαμψία. Η θέση ν = ∞ δίνει την Ευκλείδεια απόσταση.Η TWED προσδιορίζει τις λειτουργίες εισαγωγής, αφαίρεσης και αντιστοίχισης που χρησιμοποιούνται στην απόσταση επεξεργασίας, υπέρ των delete_a, delete_b και της αντιστοίχισης. Όταν ένα στοιχείο αφαιρεθεί από την πρώτη σειρά για να ταιριάζει με το δεύτερο, η λειτουργία delete_a εμφανίζεται και το delete_b είναι όταν ένα στοιχείο της δεύτερης σειράς αφαιρεθεί για να ταιριάζει με το πρώτο. Ένας υπολογισμός απόστασης Lp προτύπων χρησιμοποιείται όταν οι αντιστοιχίες υπάρχουν και η ποινή απώλειας εφαρμόζεται μέσω της παραμέτρου λ όταν οι ακολουθίες δεν ταιριάζουν, ενώ παράμετρο ακαμψίας ορίζουμε την ν.

Parameters: stiffness parameter ν, penalty value λ

Let D be an $m+1 \times m+1$ matrix initialised to zero.
$D(1,1) \leftarrow 0$
$D(2,1) \leftarrow {a_1}^2$
$D(1,2) \leftarrow {b_1}^2$
for $i \leftarrow 2$ to $m+1$ **do**
  $D(i,1) \leftarrow D(i-1,1) + |a_{i-2} - a_{i-1}|$
for $j \leftarrow 2$ to $m+1$ **do**
  $D(1,i) \leftarrow D(1,j-1) + |b_{j-2} - b_{j-1}|$
for $i \leftarrow 2$ to $m+1$ **do**
  for $j \leftarrow 2$ to $m+1$ **do**
    if $i > 2$ and $j > 2$ **then**
      $dist1 \leftarrow D(i-1,j-1) + \nu \times |i-j| \times 2 + |a_{i-1} - b_{j-1}| + |a_{i-2} - b_{j-2}|$
    else
      $dist1 \leftarrow D(i-1,j-1) + \nu \times |i-j| + |a_{i-1} - b_{j-1}|$
    if $i > 2$ **then**
      $dist2 \leftarrow D(i-1,j) + |a_{i-1} - a_{i-2}| + \lambda + \nu$
    else
      $dist2 \leftarrow D(i-1,j) + |a_{i-1}| + \lambda$
    if $j > 2$ **then**
      $dist3 \leftarrow D(i,j-1) + |b_{j-1} - b_{j-2}| + \lambda + \nu$
    else
      $dist3 \leftarrow D(i,j-1) + |b_{j-1}| + \lambda$
    $D(i,j) \leftarrow \min(dist1, dist2, dist3)$
return $D(m+1, m+1)$

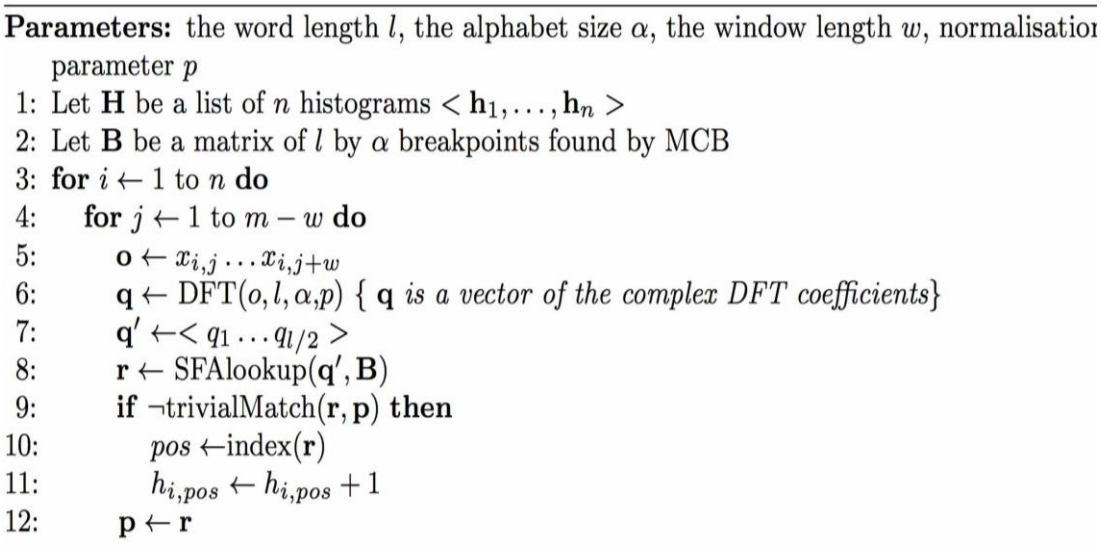
ΕΙΚΌΝΑ 2.5: Αλγόριθμος TWE

Η παραπάνω εικόνα που περιγράφει τον αλγόριθμο,ειναι κομμάτι το οποίο υπάρχει στην δημοσίευση στο διαδίκτυο Time series classification

2.1.6 Σάκος απο SFA Σύμβολα (Bag of SFA Symbols (BOSS))

Η μέθοδος BOSS χρησιμοποιεί επίσης παράθυρα για να σχηματίσει λέξεις πέρα από τις σειρές, αλλά έχει αρκετές σημαντικές διαφορές ως προς το BOP και το SAX. Πρωταρχική διαφορά μεταξύ αυτών είναι ότι το BOSS χρησιμοποιεί ένα αποκομμένο διακεκριμένο μετασχηματισμό Fourier (DFT) αντί για ένα τρόπο συγκεντρωτικής σύγκλισης κατά προσέγγιση(PAA) σε κάθε παράθυρο. Μια άλλη διαφορά είναι ότι η σειρά διακριτοποιείται μέσω μιας τεχνικής που ονομάζεται Πολλαπλή Συγκέντρωση Συντελεστή (Multiple Coefficient Binning (MCB)), αντί να χρησιμοποιεί σταθερά διαστήματα. Το MCB βρίσκει τα σημεία διάσπασης ως ένα στάδιο προεπεξεργασίας εκτιμώντας τη διανομή των συντελεστών Fourier. Αυτό γίνεται με την διάσπαση της σειράς, εκτελώντας ένα DFT, και στη συνέχεια εντοπίζοντας σημεία διακοπής για κάθε συντελεστή, έτσι ώστε κάθε κάδος να περιέχει τον ίδιο αριθμό στοιχείων. Στη συνέχεια, το BOSS περιλαμβάνει παρόμοια στάδια με το BOP όπου το παράθυρο κάθε σειράς διαμορφώνει τη διανομή λέξεων μέσω της εφαρμογής DFT. Για την ταξινόμηση πλησιέστερων γειτόνων χρησιμοποιείται μια συνάρτηση παραμετροποίησης κατά παραγγελία. Αυτή η μη συμμετρική συνάρτηση περιλαμβάνει μόνο τις αποστάσεις μεταξύ των συχνοτήτων λέξεων που συμβαίνουν στην πραγματικότητα μέσα στο πρώτο ιστόγραμμα που μεταβιβάζεται ως όρισμα. Το BOSS περιλαμβάνει επίσης μια παράμετρο που

καθορίζει αν οι υποεπιχειρήσεις είναι κανονικοποιημένες ή όχι. Η μέθοδος BOSS χρησιμοποιεί τις εξής παραμέτρους, το μήκος της λέξης l, το μέγεθος του αλφαβήτου a, το μήκος του παραθύρου w και την παράμετρο κανονικοποίησης p. Δίνοντας τίμες σε αυτές τις παραμέτρους και χρησιμοποιώντας το παρακάτω βασικό κομμάτι του BOSS μπορούμε να οδηγηθούμε σε διάφορα αποτελέσματα.



Parameters: the word length l, the alphabet size α, the window length w, normalisation parameter p

Let $\mathbf{H}$ be a list of n histograms $< \mathbf{h}_1, \ldots, \mathbf{h}_n >$
Let $\mathbf{B}$ be a matrix of l by α breakpoints found by MCB
for $i \leftarrow 1$ to n **do**
  for $j \leftarrow 1$ to $m - w$ **do**
    $\mathbf{o} \leftarrow x_{i,j} \ldots x_{i,j+w}$
    $\mathbf{q} \leftarrow \text{DFT}(o, l, \alpha, p)$ { $\mathbf{q}$ *is a vector of the complex DFT coefficients*}
    $\mathbf{q}' \leftarrow < q_1 \ldots q_{l/2} >$
    $\mathbf{r} \leftarrow \text{SFAlookup}(\mathbf{q}', \mathbf{B})$
    if $\neg$trivialMatch$(\mathbf{r}, \mathbf{p})$ **then**
      $pos \leftarrow$index$(\mathbf{r})$
      $h_{i,pos} \leftarrow h_{i,pos} + 1$
    $\mathbf{p} \leftarrow \mathbf{r}$

Εικόνα 2.6: Ο αλγόριθμος BOSS

Η παραπάνω εικόνα αποτελεί μέρος του κώδικα BOSS απο την ηλεκτρονική σελίδα (www.timeseriesclassification.gr).

3 Θεωρητικό υπόβαθρο

3.1 Τι είναι η εξόρυξη δεδομένων (data mining)

Εξόρυξη δεδομένων (data mining) ή αλλιώς Ανακάλυψη γνώσεων σε βάσεις δεδομένων (Knowledge Discovery in Databases (KDD)). Είναι κοινώς ορισμένη ως η διαδικασία ανεύρεσης χρήσιμων πληροφοριών σε μεγάλες αποθήκες δεδομένων. Οι τεχνικές εξόρυξης δεδομένων αναπτύσσονται ώστε να αποφανθούν μεγάλες βάσεις δεδομένων (ή κειμένων, εικόνων, διαδικτύου κ.λπ.) προκειμένου να βρεθούν χρήσιμα και κατανοητά μοτίβα που διαφορετικά θα μπορούσαν να παραμείνουν άγνωστα. Η εξόρυξη δεδομένων είναι ένας πολυεπιστημονικός τομέας που περιλαμβάνει μηχανική μάθηση, στατιστικές, οπτικοποίηση και πολλά άλλα. Μια εφαρμογή εξόρυξης δεδομένων αρχίζει συνήθως με την κατανόηση του ΄προβλήματος’ κάνοντας ανάλυση δεδομένων, όπου έπειτα προσδιορίζει τις κατάλληλες πηγές δεδομένων και τα δεδομένα στόχων. Αυτό δεν σημαίνει οτι όλες οι εργασίες εντοπισμού πληροφοριών θεωρούνται οτι αποτελούν τεχνική εξόρυξη δεδομένων.

3.2 Η ταξινόμιση (classification)

Η ταξινόμηση χρησιμοποιείται για να ταξινομεί κάθε στοιχείο σε ένα σύνολο δεδομένων σε μία από τις σχετικές κατηγορίες. Με αυτόν τον τρόπο, τα δεδομένα προστατεύονται και μπορούν να χρησιμοποιηθούν αποτελεσματικότερα. Η ταξινόμηση αποτελεί συνάρτηση εξόρυξης δεδομένων που αντιστοιχεί ένα στοιχείο σε μια συλλογή των κατηγοριών ή στις κατηγορίες στόχων. Μέσω της ταξινόμησης επιτυγχάνεται καλύτερη ασφάλεια των δεδομένων, διαχείριση κινδύνου και συμμόρφωση. Όπως αναφέρθηκε παραπάνω, η ταξινόμηση δεδομένων υποστηρίζει την ασφάλεια των δεδομένων καθώς προσφέρει ευκολία πρόσβασης, συμμόρφωση με τις κανονιστικές απαιτήσεις και ικανοποίηση πολλών επιχειρηματικών ή προσωπικών στόχων. Έτσι, τα δεδομένα πρέπει να μπορούν να αναζητηθούν και να ανακτηθούν σε μια συγκεκριμένη χρονική στιγμή. Ανάλογα με τον τύπο των δεδομένων που ανακτώνται, αντιγράφονται ή μεταδίδονται, η ταξινόμησή τους είναι μια χρήσιμη τακτική που βοηθά σε θέματα ασφάλειας. Η ταξινόμηση δεδομένων περιλαμβάνει πολλές επισημάνσεις και ετικέτες όπου ο τύπος τους καθορίζεται τόσο απο την ακεραιότητα όσο και απο την εμπιστευτικότητα τους. Τα δεδομένα ταξινομούνται ανάλογα με το επίπεδο ευαισθησίας τους.

3.2.1 Οι φάσεις της ταξινόμισης

Είναι σύνηθες για τους αλγόριθμους ταξινόμησης να έχουν δύο φάσεις:

1) Φάση δοκιμής (Testing phase): Η ετικέτα κλάσης για κάθε δοκιμαστική λειτουργία καθορίζεται από το μοντέλο εκπαίδευσης.

2) Φάση κατάρτισης (Training phase): Το μοντέλο κατάρτισης κατασκευάζεται από εκπαιδευτικές υπηρεσίες. Μπορούμε να το φανταστούμε ως ένα συνοπτικό μαθηματικό μοντέλο των κύριων ομάδων στο σύνολο δεδομένων κατάρτισης.

3.2.2 Τύποι απο τα αποτελέσματα της ταξινόμισης

Το αποτέλεσμα ταξινόμησης μπορεί να έχουν έναν από τους ακόλουθους τύπους:

1) Πρόβλεψη ετικετών(Label predictions): Για κάθε δοκιμαστική περίπτωση χωριστά, μπορούμε να προβλέψουμε την ετικέτα.

2) Αριθμητικό σκόρ(numerical score): Πιο συχνά, δίνουμε αποτελέσματα για κάθε ζεύγος περιπτώσεων που αναφέρει την τάση του δείγματος να ανήκει σε μια συγκεκριμένη κλάση.

Ωστόσο, η απόδοση της ταξινόμησης μπορεί να είναι χαμηλή στην περίπτωση που το σύνολο δεδομένων κατάρτισης είναι μικρό. Για το λόγο αυτό, το μοντέλο μπορεί να περιγράψει συγκεκριμένα χαρακτηριστικά του προηγούμενου συνόλου και δεν μπορεί να ενταχθεί στην ομάδα σε οποιαδήποτε από τις προηγούμενες περιπτώσεις. Έτσι, αυτό το μοντέλο θα ήταν σε θέση να προβλέψει με ακρίβεια τις ετικέτες των περιπτώσεων που χρησιμοποιήθηκαν για την κατασκευή τους, αλλά θα έχουν χαμηλή απόδοση. Αυτό το φαινόμενο ονομάζεται υπερφόρτωση.

3.3 Τα μοντέλα της ταξινόμισης

Το πρώτο στάδιο της διαδικασίας της ταξινόμησης είναι η επιλογή των χαρακτηριστικών. Η πλειονότητα των προβλημάτων ταξινόμησης σε πραγματικό κόσμο απαιτεί εποπτευόμενη μάθηση όπου οι υποκείμενες πιθανότητες τάξης και οι πιθανότητες ταξινόμησης είναι άγνωστες. Σε πραγματικές καταστάσεις, συχνά έχουμε λίγες γνώσεις σχετικά με τα σχετικά χαρακτηριστικά. Επομένως, για να αντιπροσωπεύεται καλύτερα ο τομέας, εισάγονται πολλά υποψήφια χαρακτηριστικά, με αποτέλεσμα την ύπαρξη άσχετων χαρακτηριστικών στην έννοια του στόχου. Ένα σχετικό χαρακτηριστικό δεν μπορεί να χαρακτηριστεί άσχετο ή περιττό για τον στόχο. Επίσης, μια άσχετη λειτουργία δεν συνδέεται άμεσα με την έννοια του στόχου, αλλά επηρεάζει τη διαδικασία εκμάθησης, με τον ίδιο τρόπο που ένα περιττό χαρακτηριστικό δεν προσθέτει τίποτα νέο στο στόχο. Σε πολλά προβλήματα ταξινόμησης, είναι δύσκολο να οδηγηθούμε σε καλούς ταξινομητές πριν αφαιρέσουμε αυτά τα ανεπιθύμητα χαρακτηριστικά λόγω του τεράστιου μεγέθους των δεδομένων. Μειώνοντας τον αριθμό των χαρακτηριστικών που δεν

συμβάλλουν στα αποτελέσματα μπορεί να μειωθεί δραστικά ο χρόνος εκτέλεσης των αλγορίθμων μάθησης και να αποδώσει έναν γενικότερο ταξινομητή. Αυτό βοηθά στην καλύτερη κατανόηση της υποκείμενης έννοιας του πραγματικού προβλήματος ταξινόμησης. Η επιλογή των χαρακτηριστικών επηρεάζει κυρίως την εκπαιδευτική φάση της ταξινόμησης. Μετά την παραγωγή χαρακτηριστικών, αντί της επεξεργασίας δεδομένων με όλα τα χαρακτηριστικά στον αλγόριθμο μάθησης απευθείας, η επιλογή χαρακτηριστικών για ταξινόμηση θα εκτελέσει πρώτα την επιλογή χαρακτηριστικών για να επιλέξει ένα υποσύνολο χαρακτηριστικών και στη συνέχεια να επεξεργαστεί τα δεδομένα με τις επιλεγμένες λειτουργίες στον αλγόριθμο εκμάθησης. Ο αλγόριθμος εκμάθησης μπορεί να είναι ανεξάρτητος από το στάδιο επιλογής χαρακτηριστικών ή μπορεί να χρησιμοποιήσει την απόδοση των αλγορίθμων μάθησης για την αξιολόγηση της ποιότητας των επιλεγμένων χαρακτηριστικών.

Τα πιο γνωστά μοντέλα για την ταξινόμηση δεδομένων είναι τα εξής:

1. Δέντρα απόφασης (Decision trees): Με αυτόν τον τρόπο, τα μοντέλα ταξινόμησης χτίζονται με βάση τη μορφή και τη δομή ενός δέντρου. Έτσι, παίρνει ένα σύνολο δεδομένων και το χωρίζει σε μικρότερα και μικρότερα υποσύνολα. Ταυτόχρονα αναπτύχθηκε ένα συσχετισμένο δέντρο αποφάσεων. Αφού ολοκληρωθεί η διαδικασία, το αποτέλεσμα είναι ένα δέντρο με κόμβους αποφάσεων (decision nodes) και κόμβους φύλλων (leaf nodes). Ένας κόμβος φύλλων έχει ρόλο απόφασης ή ταξινόμησης και ο κόμβος απόφασης έχει δύο ή περισσότερα μπράντυ. Για τον καλύτερο προγνωστικό σε ένα δένδρο αντιστοιχεί ο κόμβος rood που ονομάζεται κορυφαίος κόμβος απόφασης. Τα δέντρα αποφάσεων μπορούν να εκμεταλλευτούν τα συμπεράσματα για κατηγοριοποίηση και αριθμητικά δεδομένα.Rule Based Classifier: Είναι μια ταξινόμιση η οποία χρησιμοποιεί μια συλλογή απο:

   “if…….. then”

   κανόνες.

   Ένας κανόνας εκφράζεται ως:

   IF condition THEN conclusion

2. Πιθανοί ταξινομητές: Κατασκευάζουν ένα μοντέλο που έχει ως στόχο να ποσοτικοποιεί τη σχέση μεταξύ των μεταβλητών χαρακτηριστικών και της μεταβλητής στόχου ως πιθανότητα. Αν και υπάρχουν πολλά μοντέλα που το έκαναν αυτό, δύο από τα πιο δημοφιλή είναι:

- Bayes ταξινομητές: Χρησιμοποιούνται συχνά οι ταξινομητές Naïve Bayes οι οποίοι είναι μια οικογένεια απλών πιθανοτικών ταξινομητών που βασίζονται στην εφαρμογή του θεωρήματος Bayes με ισχυρές υποθέσεις ανεξαρτησίας μεταξύ των χαρακτηριστικών.
- Λογιστική παλινδρόμηση: Η μεταβλητή στόχος που προκύπτει από μια κατανομή Bernoulli όπου η μέση τιμή καθορίζεται από μια παραμετρική λογική συνάρτηση. Είναι αντίθετη με το μοντέλο Bayes που ασχολείται με ένα συγκεκριμένο μοντέλο.

3. Μάθηση βασισμένη σε περιστατικά: Η εκπαίδευση καθυστερεί μέχρι το τελευταίο βήμα της ταξινόμησης. Το πιο απλό παράδειγμα που μπορεί να δοθεί για να περιγραφεί αυτό είναι:

- "Παρόμοιες περιπτώσεις έχουν παρόμοιες ετικέτες κατηγορίας".

4. Αξιολόγηση ταξινομητή: Γνωρίζοντας ένα μοντέλο ταξινόμησης και ποσοτικοποιώντας την ακρίβειά του σε συγκεκριμένο σύνολο δεδομένων μπορούμε να αξιολογήσουμε την αποτελεσματικότητα του ταξινομητή συγκρίνοντας διαφορετικά μοντέλα ώστε να διατηρήσουμε τη καλύτερη παράμετρο για συγκεκριμένο πλήθος δεδομένων καθώς και πολλούς μετα-αλγορίθμους.

5. Ένας φορέας υποστήριξης: είναι ένας διακριτικός ταξινομητής που ορίζεται τυπικά από ένα διαχωριστικό υπερ-επίπεδο. Με άλλα λόγια, δίνοντας επισημασμένα training data (επίβλεψη μάθησης), ο αλγόριθμος εξάγει ένα βέλτιστο υπερ-επίπεδο το οποίο κατηγοριοποιεί νέα παραδείγματα. Σε διδιάστατους χώρους αυτό το υπερ-επίπεδο είναι μια γραμμή που διαιρεί ένα αντικείμενο σε δύο μέρη όπου κάθε τάξη βρίσκεται σε κάθε πλευρά.

6. Νευρωνικά δίκτυα(neural networks) : Είναι ένα παράδειγμα επεξεργασίας πληροφοριών που εμπνέεται από τον τρόπο με τον οποίο τα βιολογικά νευρικά συστήματα, όπως ο εγκέφαλος, επεξεργάζονται πληροφορίες. Το βασικό στοιχείο αυτού του παραδείγματος είναι οτι η σύνθεσή του χαρακτηρίζεται από μεγάλο αριθμό διασυνδεδεμένων στοιχείων επεξεργασίας (νευρώνων) που λειτουργούν στο σύνολό τους για την επίλυση συγκεκριμένων προβλημάτων. Οι ANNs, όπως και οι άνθρωποι, μαθαίνουν μέσω παραδείγματος. Ένα ANN έχει ρυθμιστεί για μια συγκεκριμένη εφαρμογή, όπως αναγνώριση προτύπων ή ταξινόμηση δεδομένων, μέσω μιας διαδικασίας μάθησης. Η εξειδίκευση σε βιολογικά συστήματα περιλαμβάνει προσαρμογές στις συνδέσεις που υπάρχουν μεταξύ των νευρώνων.

3.4 Αλγόριθμοι εξόριξης δεδομένων

Ένας αλγόριθμος στην εξόρυξη δεδομένων ή στη μηχανική μάθηση περιλαμβάνει ένα σύνολο θεωρίας και υπολογισμών που συνθέτουν ένα μοντέλο από δεδομένα. Ο αλγόριθμος ξεκινάει πρώτα να αναλύει τα δεδομένα που παρέχετε, αναζητώντας συγκεκριμένους τύπους μοτίβων που στοχεύουν στη δημιουργία ενός μοντέλου. Με τα αποτελέσματα που θα εξάγουμε απο αυτήν την ανάλυση θα βρούμε τις καλύτερες παραμέτρους για τη δημιουργία του μοντέλου εξόρυξης. Αυτές οι παράμετροι εφαρμόζονται στη συνέχεια σε ολόκληρο το σύνολο δεδομένων για την εξαγωγή μοτίβων και λεπτομερών στατιστικών στοιχείων.

Οι μορφές που μπορεί να πάρει ένα μοντέλο εξόρυξης που δημιουργήθηκε από έναν αλγόριθμο με τα δεδομένα σας είναι:

• Ένα σύνολο ομάδων που περιγράφουν τον τρόπο με τον οποίο σχετίζονται οι περιπτώσεις σε ένα σύνολο δεδομένων.

• Ένα δέντρο απόφασης που προβλέπει ένα αποτέλεσμα και περιγράφει πώς επηρεάζουν τα αποτελέσματα τα διαφορετικά κριτήρια.

• Ένα μαθηματικό μοντέλο που προβλέπει τις πωλήσεις.

• Ένα σύνολο κανόνων που περιγράφουν τον τρόπο ομαδοποίησης των προϊόντων σε μια συναλλαγή.

3.4.1 Η επιλογή του καλύτερου αλγορίθμου

Πρόκληση αποτελεί η επιλογή του αλγορίθμου. Μπορούμε να επιλέξουμε απο ένα πλήθος αλγορίθμων για την ίδια εργασία, ωστόσο, κάθε αλγόριθμος δίνει ένα διαφορετικό αποτέλεσμα και ορισμένοι αλγόριθμοι μπορούν να παράγουν περισσότερους από έναν τύπους αποτελεσμάτων. Οι αλγόριθμοι ταξινόμησης προβλέπουν μία ή περισσότερες διακριτές μεταβλητές, με βάση τα χαρακτηριστικά του συνόλου δεδομένων.

- Οι αλγόριθμοι παλινδρόμησης προβλέπουν μία ή περισσότερες συνεχείς αριθμητικές μεταβλητές, όπως το κέρδος ή η ζημία, βάσει άλλων χαρακτηριστικών στο σύνολο δεδομένων.
- Οι αλγόριθμοι τμηματοποίησης διαιρούν τα δεδομένα σε ομάδες ή συμπλέγματα αντικειμένων που έχουν παρόμοιες ιδιότητες.
- Οι αλγόριθμοι σύνδεσης βρίσκουν συσχετισμούς μεταξύ διαφορετικών χαρακτηριστικών σε ένα σύνολο δεδομένων. Η πιο κοινή εφαρμογή αυτού του είδους του αλγορίθμου είναι η δημιουργία κανόνων σύνδεσης, οι οποίοι μπορούν να χρησιμοποιηθούν σε μια ανάλυση market basket.
- Οι αλγόριθμοι ανάλυσης ακολουθίας συνοψίζουν συχνές ακολουθίες ή τμήματα δεδομένων, όπως μια σειρά κλικ σε μια τοποθεσία Web ή μια σειρά γεγονότων καταγραφής που προηγούνται της συντήρησης της μηχανής.

3.4.2 Οι 10 κάλυτεροι αλγόριθμοι

1) Ο αλγόριθμος C4.5:

   Είναι ένας αλγόριθμος που οδηγεί σε έναν ταξινομητή με τη μορφή ενός δέντρου αποφάσεων και ο εμπνευστής του είναι ο Ross Quinlan. Για να γίνει το ίδιο, το C4.5 χρησιμοποιεί ένα σύνολο δεδομένων που αντιπροσωπεύουν πράγματα που έχουν ήδη ταξινομηθεί. Λέγεται ως στατιστικός ταξινομητής και αποτελεί επέκταση του αλγορίθμου ID3 του Quinlan. Τα δέντρα που παράγονται χρησιμοποιούνται για ταξινόμηση. Συχνά αναφέρεται ως "ένα πρόγραμμα δέντρων αποφάσεων ορόσημο που είναι ίσως η μηχανή μάθησης που χρησιμοποιείται περισσότερο στην πράξη μέχρι σήμερα".

2) Με τη σειρά του ο k-means:

   k-means αποτελεί είδος ομαδοποίησης που είναι επίσης γνωστή ως πλησιέστερος ταξινομητής βαρύτητας ή ο αλγόριθμος The Rocchio που είναι μια μέθοδος κβαντισμού διανυσμάτων. Το k-means βοηθά στη δημιουργία ομάδων k από ένα σύνολο αντικειμένων έτσι ώστε τα μέλη του να μοιάζουν. Είναι μια πολύ γνωστή τεχνική ανάλυσης που χρησιμοποιείται για την εξερεύνηση ενός συνόλου δεδομένων.

3) Ακόμα ο Support vector machines:

Όταν πρόκειται για μηχανήματα φορέων υποστήριξης μηχανικής μάθησης, τα μοντέλα μάθησης με επίβλεψη είναι εφοδιασμένα με συναφείς αλγόριθμους μάθησης, οι οποίοι ως αποτέλεσμα αναλύουν τα δεδομένα που χρησιμοποιούνται για την ανάλυση της παλινδρόμησης και της ταξινόμησης. Δημιουργώντας έτσι ένα μοντέλο αυτού του τύπου που είναι μια αναπαράσταση των παραδειγμάτων ως σημεία στο διάστημα, τα οποία χαρτογραφούνται περαιτέρω έτσι ώστε τα παραδείγματα των ξεχωριστών κατηγοριών διαιρούνται έπειτα από ένα σαφές χάσμα το οποίο θα έπρεπε να είναι όσο το δυνατόν ευρύτερο.

4) Ο αλγόριθμος Apriori:

Το Apriori είναι ένας αλγόριθμος που χρησιμοποιείται για τη συχνή στοιχειοθετημένη εξόρυξη και τη συσχέτιση με τη μάθηση γενικών βάσεων συναλλαγών. Ο αλγόριθμος προχωράει από την ταυτοποίηση των στοιχείων που είναι συχνές στη βάση δεδομένων και στη συνέχεια την επέκταση σε μεγαλύτερα στοιχεία, εφόσον αυτά τα σύνολα στοιχείων εμφανίζονται αρκετά συχνά στη βάση δεδομένων. Αυτά τα συνηθισμένα σύνολα στοιχείων που καθορίζονται από το Apriori μπορούν να χρησιμοποιηθούν για τον καθορισμό των κανόνων σύνδεσης που στη συνέχεια τονίζουν τις γενικές τάσεις.

5) Έπειτα ο EM(Expectation-Maximization):

Στην περίπτωση στατιστικών στοιχείων, χρησιμοποιείται μια μέθοδος επανάληψης που χρησιμοποιείται για την εύρεση των μέγιστων πιθανοτήτων ή εκτιμήσεων των παραμέτρων στα στατιστικά μοντέλα, που βασικά εξαρτάται από τις μη παρατηρούμενες λανθάνοντες μεταβλητές.

6) Ο PageRank (PR):

Το PageRank (PR) που ονομάστηκε από τον συνεργάτη Larry Page, ένας από τους ιδρυτές της Google, είναι ένας αλγόριθμος που χρησιμοποιείται από την Αναζήτηση Google για την κατάταξη των ιστότοπων στα αποτελέσματα των μηχανών αναζήτησης. Ο PageRank, αυτός είναι ο πρώτος αλγόριθμος που χρησιμοποίησε η εταιρεία, δεν είναι ο μόνος αλγόριθμος που χρησιμοποιείται από την Google για την αναζήτηση αποτελεσμάτων, αλλά είναι ο πιο γνωστός τρόπος μέτρησης της σημασίας των σελίδων του ιστότοπου.

7) AdaBoost:

Με τους δημιουργούς Yoav Freund και Robert Schapire είναι ένας μηχανογραφικός μετα-αλγόριθμος. Η απόδοσή του μπορεί να βελτιωθεί

μέσω της χρήσης του με άλλους τύπους αλγορίθμων. Το AdaBoost είναι ευαίσθητο σε ακραίες τιμές.

8) Εξίσου σημαντικός ο kNN:
Ο k-NN αλγόριθμος είναι ένας τύπος αργής μάθησης που θεωρείται ως μια μη παραμετρική μέθοδος η οποία χρησιμοποιείται για ταξινόμηση και παλινδρόμηση. Η είσοδος αποτελείται από τα k πλησιέστερα παραδείγματα εκπαίδευσης στο χώρο των χαρακτηριστικών και η έξοδος εξαρτάται από το αν ο αλγόριθμος χρησιμοποιείται για ταξινόμηση ή παλινδρόμηση. Αυτός ο αλγόριθμος θεωρείται και είναι επίσης μεταξύ των απλούστερων αλγορίθμων μηχανικής μάθησης.

9) Αλγόριθμος Naive Bayes:
Όταν πρόκειται για μηχανική μάθηση, οι ταξινομητές Naive Bayes που θεωρούνται εξαιρετικά κλιμακωτοί είναι γνωστοί ως μια οικογένεια απλών πιθανοτικών ταξινομητών που βασίζονται στην εφαρμογή του θεωρήματος του Bayes με τη βοήθεια ισχυρών ανεξάρτητων υποθέσεων μεταξύ των χαρακτηριστικών.

10) Ακόμα ο CART:
Είναι ένας αλγόριθμος που χρησιμεύει για ταξινόμηση και δέντρα παλινδρόμησης. Είναι μια τεχνική μάθησης δέντρων αποφάσεων που είτε εξάγει ταξινομήσεις ή παλινδρομικά δέντρα όπως και το C4.5. Οι λόγοι που θα οδηγήσουν τον χρήστη στη χρήση του C4.5 ισχύουν και για τον CART, αφού και οι δύο είναι τεχνικές μάθησης δέντρων αποφάσεων και χαρακτηριστικών όπως και η ευκολία ερμηνείας και εξήγησης

3.5 Τι είναι οι χρονοσειρές και η χρήση τους

Τα τελευταία χρόνια,σε πολλές εφαρμογές στη ζωή μας, όπως στα οικονομικά, στην ιατρική, την πρόβλεψη καιρού και αλλες, οι χρονοσειρές είναι το αντικείμενο ενδιαφέροντος. Οι χρονοσειρές είναι μια ακολουθία καλά καθορισμένων αριθμητικών σημείων που λαμβάνονται σε διαδοχικά ισαπέχουσα σημεία κατά το χρόνο. Μια χρονοσειρά μπορεί να ληφθεί σε οποιαδήποτε μεταβλητή που αλλάζει με την πάροδο του χρόνου. Πολύ συχνά σχεδιάζονται μέσω γραφημάτων. Επιτρέπει τη συγκέντρωση των δεδομένων κατά τρόπο που να παρέχει τις ζητούμενες πληροφορίες, επειδή δεν υπάρχει μέγιστος ή ελάχιστος εσωτερικός χρόνος που πρέπει να συμπεριληφθεί. Η ανάλυση χρονοσειρών μας βοηθά να κατανοήσουμε ποιές είναι οι εικονικές τιμές που οδηγούν σε τάση στα σημεία χρονοσειρών με τα κατάλληλα μοντέλα. Επίσης, η ανάλυση χρονοσειρών, μας βοηθά να συλλέγουμε

σημαντικά χαρακτηριστικά των δεδομένων. Υπάρχουν πολλά οφέλη και εφαρμογές της ανάλυσης χρονοσειρών που περιλαμβάνουν διαφορετικούς τύπους:

- Πρόβλεψη, που χρησιμοποιείται στις επιχειρηματικές προβλέψεις για την άντληση δεδομένων από τις προηγούμενες τάσεις.

- Ανάλυση παρεμβολής, χρησιμοποιείται για να αποφασίσει αν μια διέγερση της εργασίας μπορεί να αλλάξει τις χρονοσειρές. Για παράδειγμα, αν η αποτελεσματικότητα βελτιώθηκε μετά από αύξηση μισθού.

-Αναπτυξιακή ανάλυση, αποσκοπεί στο διαχωρισμό των κυκλικών συνιστωσών σε μια χρονοσειρά. Για παράδειγμα, κυκλικές αλλαγές στην αποδοτικότητα των εργαζομένων.

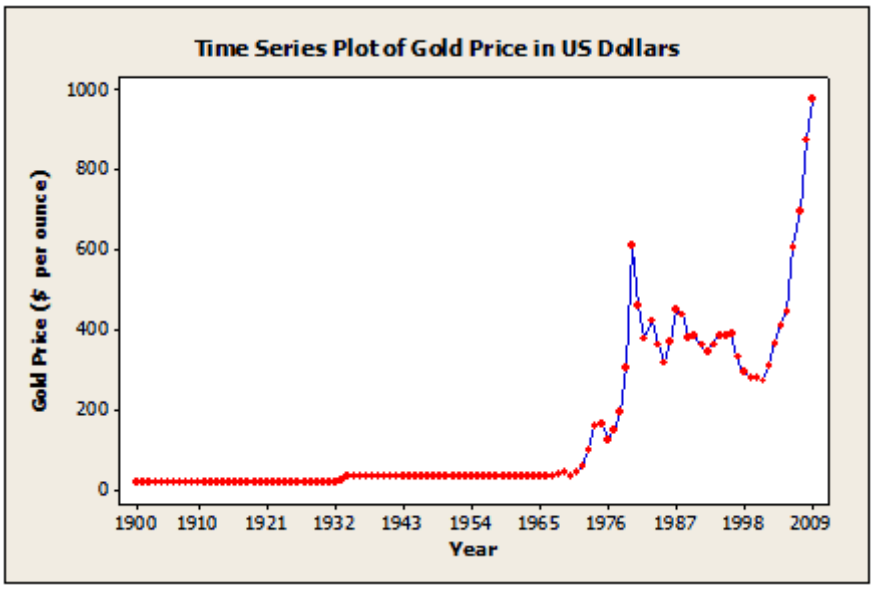
- Εκτενής ανάλυση, ασχολείται με τη σχέση δύο χρονοσειρών και μελετά την εξάρτηση και των δύο.

-Επιστημονική ανάλυση, για τον προσδιορισμό της τάσης σε μια χρονοσειρά χρησιμοποιώντας γραφήματα.

3.6 Παραγείγματα γραφημάτων

Ακολουθούν ορισμένα παραδείγματα χρονοσειρών μέσω γραφημάτων για καλύτερη κατανόηση:

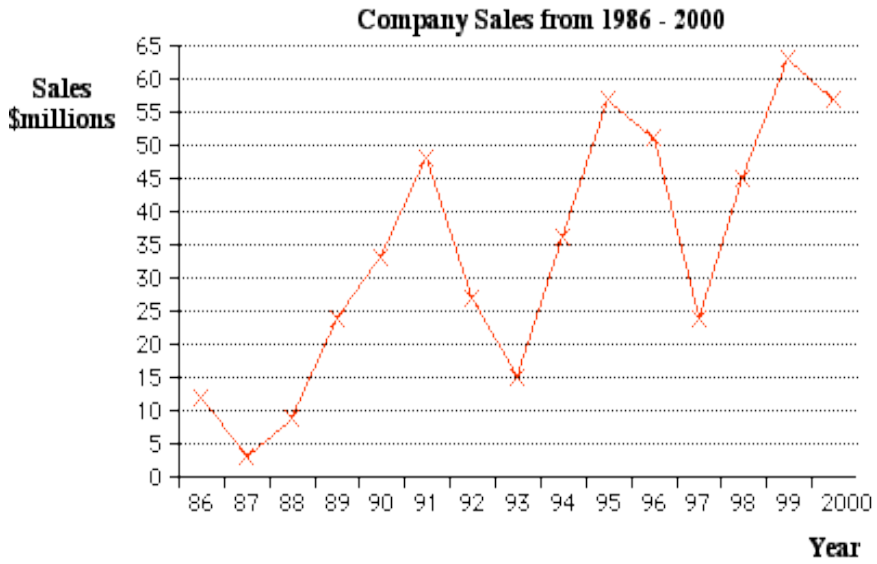
Στο ακόλουθο διάγραμμα παρατηρούμε τη μεταβολή της τιμής του χρυσού σε σχέση με το δολαρίο σε βάθος χρόνου:



ΕΙΚΌΝΑ 3.1 Χρυσό σε δολάρια

Η εικόνα ανακτήθηκε απο την παρακάτω διεύθυνση https://www.quora.com/What-is-the-time-series-graph .

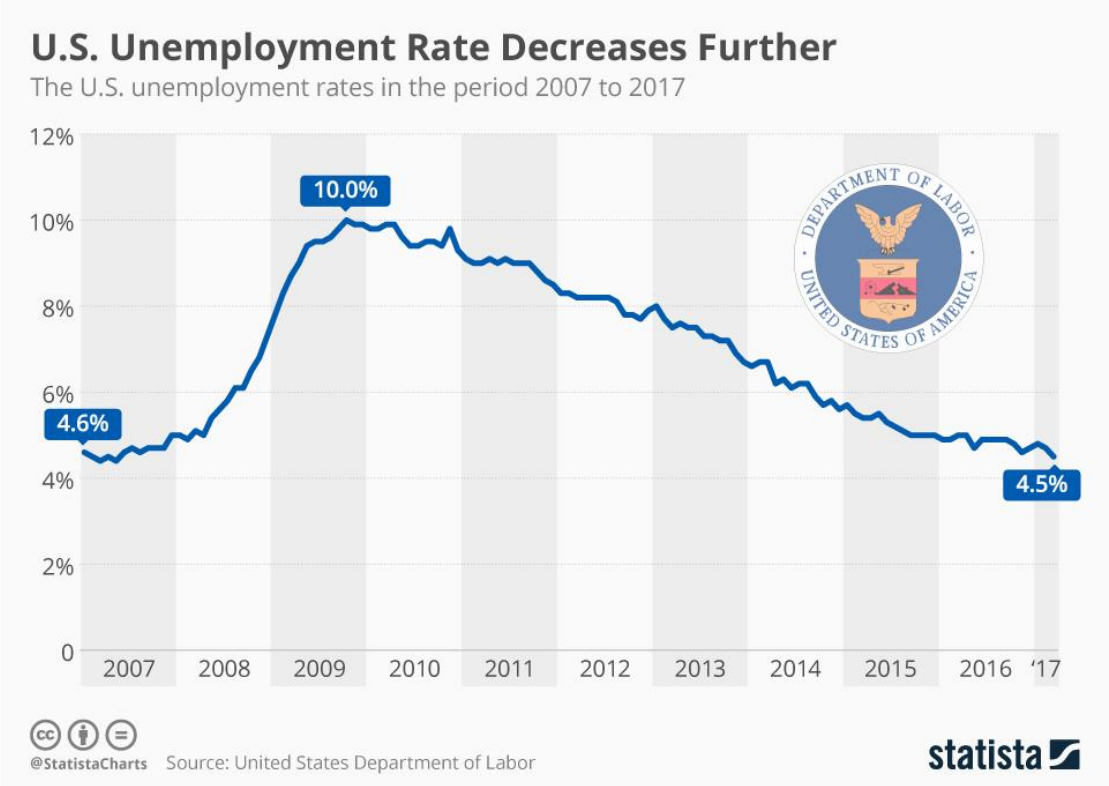
Στο παρακάτω διάγραμμα μπορούμε να διακρίνουμε τις αλλαγές στις πωλήσεις μιας εταιρείας σε μια περίοδο αρκετών ετών:



ΕΙΚΌΝΑ 3.2 : Πωλήσεις υπολογιστών σε εκατομμύρια δολάρια για τα έτη 1986-2000

Πηγή της παραπάνω εικόνας αποτελεί η ηλεκτρονική διεύθυνση: http://bestmaths.net/online/index.php/year-levels/year-12/year-12-topic-list/time-series/ .

Ακολουθεί ένα γράφημα που δείχνει το ποσοστό ανεργίας από το 2007. Αλλαγές που υπάρχουν με το πέρασμα του χρόνου:



ΕΙΚΌΝΑ 3.3: Ο ρυθμός ανεργίας των ΗΠΑ

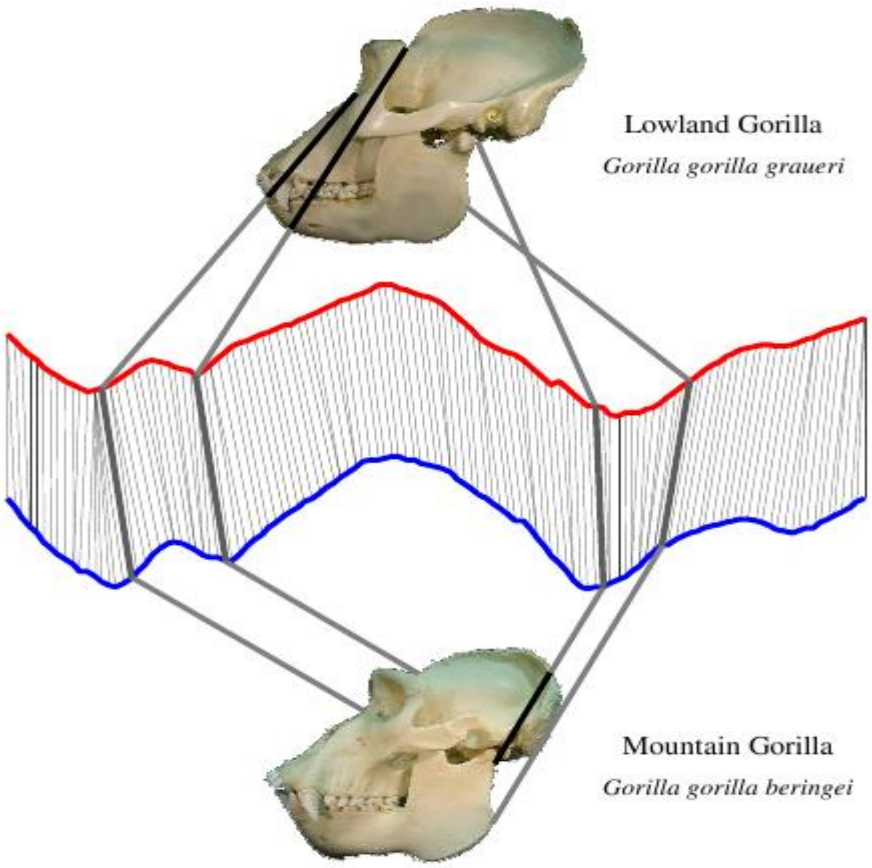
Ανάκτηση εικόνας απο: https://www.statista.com/chart/8974/us-unemployment-rate .

Παρατηρούμε ότι όλα τα παραδείγματα είναι παροδικά με το χρόνο.

4 Μετατροπή εικόνων σε χρονοσειρές για εξόριξη δεδομένων

4.1 Μετατροπή εικόνων σε χρονοσειρές για εξόρυξη δεδομένων

Σύμφωνα με έρευνα του Mike Izbicki που δημοσιεύτηκε στις 28-10-2011 (https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html), μας βοηθά να καταλάβουμε πώς να συγκρίνουμε δύο αντικείμενα χρησιμοποιώντας ακτινική σάρωση(radial scanning) και δυναμική χρονική στρέβλωση (dynamic time warping). Καταρχήν, δημιούργησε ένα μέτρο απόστασης για δύο εικόνες. Αυτή η αναφορά θα διδάξει τον τρόπο δημιουργίας μέτρου απόστασης χρησιμοποιώντας ανάλυση χρονοσειρών. Για αντικείμενα με σταθερό, άκαμπτο σχήμα αυτή η τεχνική είναι η καλύτερη. Για παράδειγμα, θα λειτουργεί καλά στην ταξινόμηση εικόνων κρανίων, αλλά όχι σε εικόνες ανθρώπων. Το ανθρώπινο σώμα είναι λυγισμένο, αλλά τα κρανία έχουν πάντα το ίδιο σχήμα. Κάνουμε μια σύγκριση μεταξύ κρανίων.



ΕΙΚΌΝΑ 0.1: Τα κρανία ως χρονοσειρά

Πηγή της εικόνας αποτελεί το άρθρο : https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Όπως μάθαμε στο προηγούμενο κεφάλαιο, η χρονοσειρά είναι οτιδήποτε μπορεί να γραφεί ως ενα γράφημα. Η μελέτη των χρονοσειρών είναι αρκετά εκτεταμένη και έχει αναπτύξει μεταβλητές τεχνικές για την ανάλυσή τους. Μετατρέποντας την εικόνα σε χρονοσειρά όλα τα στοιχεία που μας βοηθούν να την επεξεργαστούμε είναι διαθέσιμα σε εμάς.

Για να μετρήσετε την απόσταση της χρονοσειράς, πρέπει να εκτελέσετε τα παρακάτω βήματα:

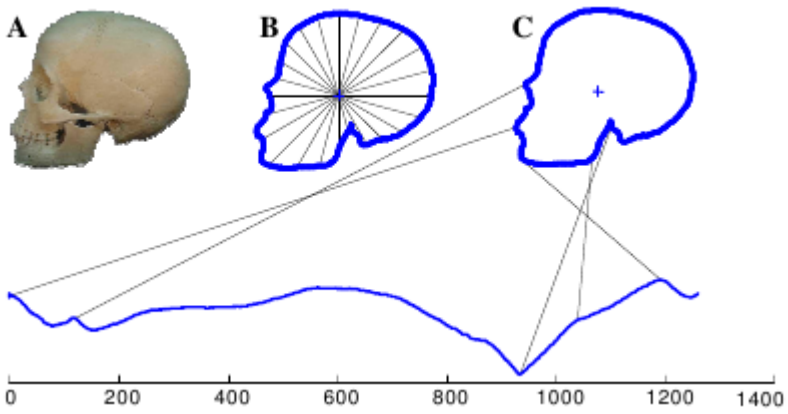
1. Μετατρέψτε τις εικόνες σε μια χρονοσειρά.

2. Βρείτε την απόσταση μεταξύ δύο εικόνων βρίσκοντας την απόσταση μεταξύ των χρονικών σειρών τους.

Για κάθε ένα από τα παραπάνω βήματα, υπάρχει μια επιλογή μεταξύ πολλών αλγορίθμων. Παρακάτω θα παρουσιάσουμε δύο εικόνες μετατροπής αλγορίθμων σε χρονολογικές σειρές, γραμμική και ακτινική σάρωση. Στη συνέχεια θα εξετάσουμε δύο αλγόριθμους που μας βοηθούν στη μέτρηση της χρονοσειράς: Ευκλείδεια απόσταση και δυναμική χρονική στρέβλωση. Τέλος, θα μάθουμε ποια είδη προβλημάτων χειρίζεται καλά ή όχι αυτή τη θεωρία.

ΒΗΜΑ 1Α: Δημιουργια της χρονοσειράς μέσω ακτινικής σάρωσης

Βλέπουμε ένα παράδειγμα ανθρώπινου κρανίου:



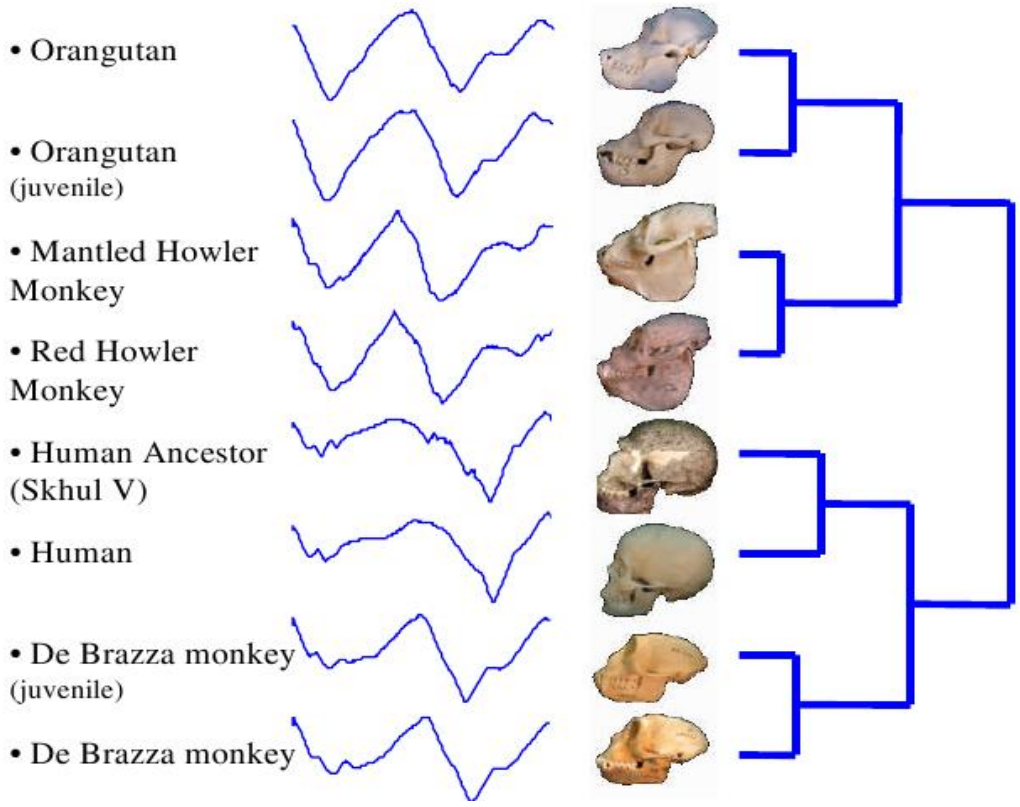
ΕΙΚΌΝΑ 0.2: Η ανάλυση του κρανίου

Εικόνα απο τη διεύθυνση : https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Ξεκινάμε με το περίγραμμα του κρανίου. Στη συνέχεια, προσπαθούμε να βρούμε την απόσταση του κρανίου από τη μέση σε κάθε σημείο της γραμμής (Β). Μετά από

αυτό είμαστε έτοιμοι να το σχεδιάσουμε ως μια χρονοσειρά αυτών των αποστάσεων (C). Οι γραμμές του γράφου C που συνδέουν το κρανίο με τη χρονοσειρά δείχνουν πού κάθε σημείο του κρανίου βρίσκεται στο γράφημα. Στη συγκεκριμένη περίπτωση, ξεκινήσαμε από το στόμα του κρανίου και μετακινηθήκαμε προς το πίσω μέρος του κεφαλιού.

Κάθε τύπος κρανίου παράγει μια διαφορετική χρονοσειρά:

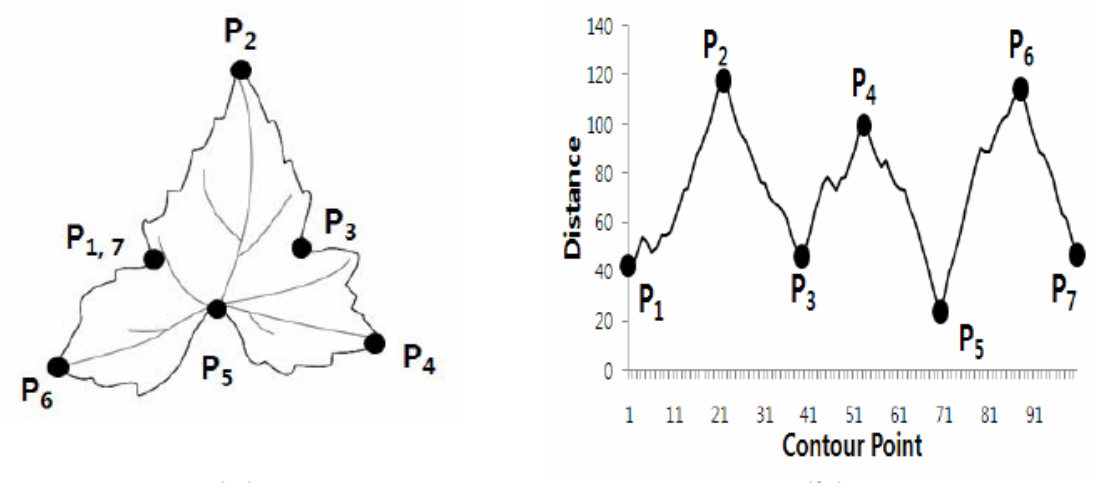


ΕΙΚΌΝΑ 0.3: Η χρονοσειρά του κάθε κρανίου

Η εικόνα παραχωρήθηκε απο : https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Παρατηρούμε ότι μπορούμε να εντοπίσουμε διαφορές στις χρονοσειρές μεταξύ κάθε ομαδοποίησης.

Ένα άλλο παράδειγμα που διερευνά το πανεπιστήμιο της Κορέας είναι αυτό που προσπαθεί να εξάγει ένα αποτέλεσμα για το είδος του δέντρου από το σχήμα των φύλλων του:



ΕΙΚΌΝΑ 0.4: Η ανάλυση του φύλλου

Πηγή της εικόνας : https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Τα επισημασμένα εικονίδια στην αριστερή πλευρά της εικόνας αντιστοιχούν στις σημειωμένες θέσεις των χρονοσειρών που δημιουργήθηκαν στη δεξιά εικόνα. Κάθε φυτό έχει ξεχωριστό σχήμα φύλλου, έτσι ώστε η ακτινική σάρωση να την βοηθήσει να το ταξινομήσει. Κάθε φύλλο θα είναι μοναδικό, αλλά το πρότυπο κορυφών και κοιλάδων στις προκύπτουσες χρονολογικές σειρές θα πρέπει να είναι παρόμοιο αν το είδος φυτού είναι το ίδιο. Παρατηρούμε ότι τα γραφήματα που δημιουργήθηκαν στις δύο παραπάνω περιπτώσεις είναι πολύ διαφορετικά. Έτσι, καταγράφονται σημαντικές πληροφορίες σχετικά με το σχήμα των αντικειμένων που χρησιμοποιούνται στο στάδιο σύγκρισης.

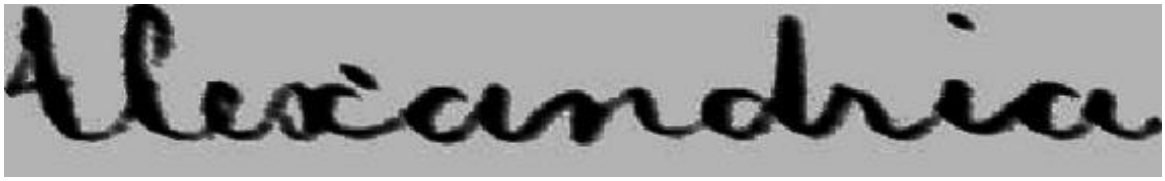
• ΒΗΜΑ 1Β: Δημιουργία μιας χρονοσειράς με γραμμική σάρωση

Στα αντικείμενα που δεν είναι κυκλικά, η ακτινική σάρωση δεν έχει νόημα. Ένα παράδειγμα είναι οι χειρόγραφες λέξεις. Ένας μεγάλος αριθμός των γραπτών του George Washington έχει επεξεργαστεί και αναλυθεί από το Πανεπιστήμιο της Μασαχουσέτης χρησιμοποιώντας τη μέθοδο γραμμικής σάρωσης. Στην εικόνα που ακολουθεί υπάρχει η λέξη "Αλεξάνδρεια" όπως την έγραψε ο Ουάσινγκτον:

ΕΙΚΌΝΑ 0.5: Αλεξάνδρεια

Εικόνα απο την ηλεκτρονική διεύθυνση https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

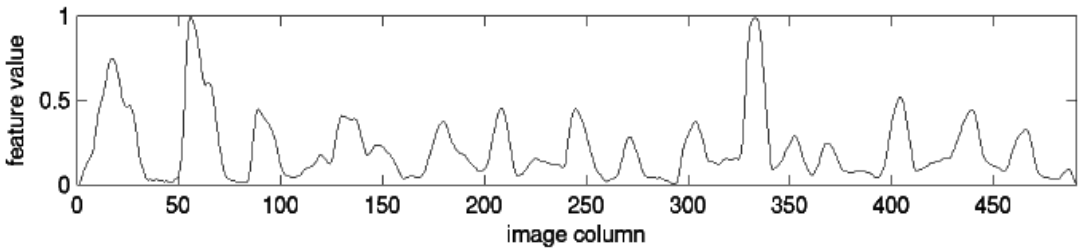
Τώρα, αφαιρούμε την κλίση από την εικόνα.



ΕΙΚΌΝΑ 0.6: Χωρίς κλίση

Η εικόνα αποτελεί μέρος άρθρου που βρίσκεται στη διεύθυνση : https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Με αυτό τον τρόπο θέλουμε να δημιουργήσουμε μια χρονοσειρά από τη λέξη:



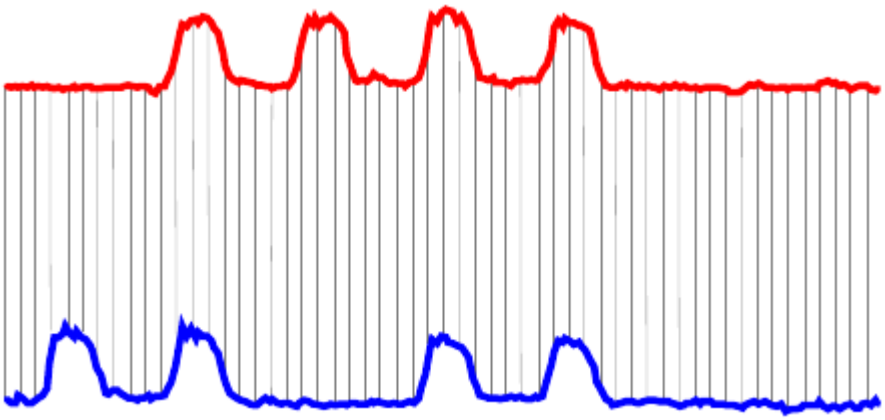
ΕΙΚΌΝΑ 0.7: Η χρονοσειρά απο τη λέξη Αλεξάνδρεια

Πηγή ανάκτησης της εικόνας: https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Οι χρονοσειρές δημιουργούνται όπως παρακάτω, ξεκινάμε στα αριστερά της εικόνας και εξετάζουμε κάθε στήλη εικονοστοιχείων με τη σειρά τους. Η τιμή σε κάθε "χρόνο" είναι ακριβώς ο αριθμός των σκοτεινών εικονοστοιχείων στη συγκεκριμένη στήλη. Αν κοιτάξετε προσεκτικά τις χρονολογικές σειρές, θα πρέπει να είστε σε θέση να προσδιορίσετε κάθε χτύπημα που αντιστοιχεί σε ένα συγκεκριμένο γράμμα. Μερικά γράμματα, όπως και το "d", παίρνουν δύο χτυπήματα στις χρονοσειρές επειδή έχουν δύο περιοχές με υψηλή συγκέντρωση σκοτεινών εικονοστοιχείων. Τώρα που μπορούμε να δημιουργήσουμε τις χρονολογικές σειρές, ας υπολογίσουμε πώς να τις συγκρίνουμε.

ΒΗΜΑ 3: Συγκρίνοντας τις αποστάσεις

Η ευκλείδεια απόσταση είναι ο ευκολότερος τρόπος για τη χρονοσειρά να δημιουργεί ένα μέτρο απόστασης που θα χρησιμοποιηθεί. Αυτός ήταν ουσιαστικά ο σκοπός της δημιουργίας τους. Υπάρχουν δύο χρονολογικές σειρές:



ΕΙΚΌΝΑ 0.8: Ευκλείδεια απόσταση

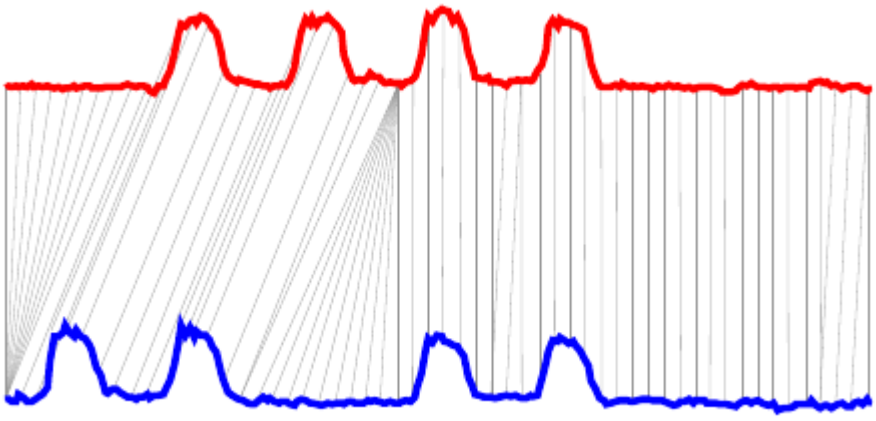
Η εικόνα προέρχεται απο: https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Ο τύπος της Ευκλείδειας απόστασης:

$$\text{distance} = \sqrt{\sum_{i=1}^{N} (\text{red}_i - \text{blue}_i)^2}$$

όπου το red-i είναι το ύψος της κόκκινης σειράς στο "χρόνο" i, το blue-i είναι το ύψος της μπλε σειράς στο "χρόνο" i, και το N είναι το μήκος της χρονοσειράς. Αυτός ο υπολογισμός που τρέχει στο χρόνο O(N) είναι όσο γρήγορος και απλός γίνεται. Η Dynamic Time Warping (DTW) αποτελεί έναν πιο εξειδικευμένο τρόπο σύγκρισης των χρονοσειρών που προσπαθεί να συγκρίνει παρόμοιες περιοχές σε κάθε χρονοσειρά μεταξύ τους.

Βρίσκουμε τις συγκεκριμένες χρονοσειρές μέσω του DTW:



ΕΙΚΌΝΑ 0.9: DTW για χρονοσειρές

Η εικόνα ανακτήθηκε απο: https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html .

Στο παραπάνω σχήμα, κάθε καμπύλη της κυανής γραμμής αντιστοιχεί στις καμπύλες της κόκκινης γραμμής και στις επίπεδες περιοχές. Το γεγονός είναι ότι μόνο ένα στοιχείο μπορεί να ανατεθεί σε πολλά σημεία της χρονοσειράς. Όταν το φαινόμενο αυτό συνέβη, το DTW δίνει μια απόσταση σχεδόν μηδενική, είναι σχεδόν τέλειο. Με άλλο τρόπο, η ευκλείδεια απόσταση θα ήταν δύσκολο να αποκομίσει αποτελέσματα και οι αποστάσεις θα ήταν τεράστιες.

Τέλος, υπάρχουν διάφοροι τρόποι με τους οποίους μπορούμε να συγκρίνουμε τις χρονοσειρές. Ο αλγόριθμος που χρησιμοποιείται περισσότερο ονομάζεται Μακρύτερη κοινή δευτερεύουσα ακολουθία (Longest Common Sub-Sequence (LCSS)). Είναι χρήσιμο για αντιστοίχιση εικόνων που υποφέρουν από απόφραξη.

4.2 Πότε χρησιμοποιείται η ανάλυση απο χρονοσειρές

Η ανάλυση χρονοσειρών μπορεί να βοηθήσει μόνο στη διαμόρφωση αντικειμένων. Είναι προϋπόθεση να μην αλλάζει το χρώμα και η σύνθεσή τους. Σύμφωνα με αυτούς τους δύο δείκτες, η ανάλυση των κρανίων, των φύλλων, των γραμμάτων και άλλων άκαμπτων αντικειμένων κάνει καλά την ανάλυση σε χρονοσειράς. Επειδή τα παραπάνω στοιχεία δεν αλλάζουν σημαντικά με το πέρασμα του χρόνου, μπορούν να δώσουν τα δικά τους αποτελέσματα οπουδήποτε και αν μετρηθούν. Ίσως η ανάλυση της χρονοσειράς δεν έχει τα ίδια αποτελέσματα σε αντικείμενα που είναι ευέλικτα και μπορούν συχνά να αλλάξουν τη θέση τους. Οι άνθρωποι είναι πιο δύσκολο να εξερευνηθούν και να εκπροσωπούνται ως χρονολογικές σειρές καθώς μπορούν να κινηθούν, να τρέξουν, να πηδήξουν και οτιδήποτε άλλο.

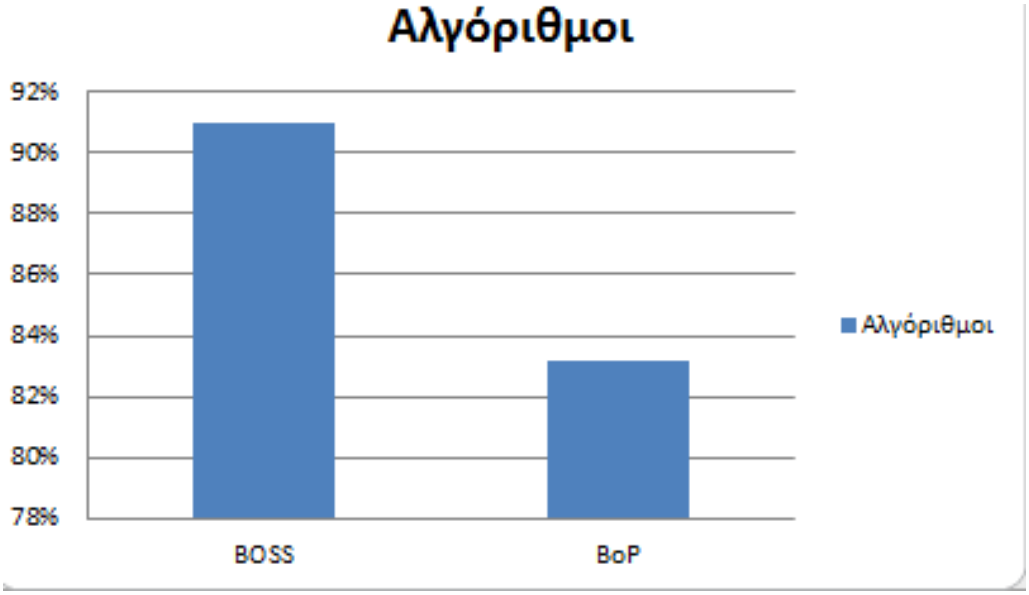
5 Η περιγραφή του συνόλου δεδομένων και η αξιολόγηση της επιτυχίας

5.1 Περιγραφή του συνόλου δεδομένων

Οι πληροφορίες για τα δεδομένα της έρευνας αντλήθηκαν απο την ηλεκτρονική διευθυνση www.timeseriesclassification.gr απο την οποία ελήφθησε και ο κώδικας μέσω του πακέτου Yoga το οποίο παρατηρεί δύο αθλητές που εξασκούνται στη γιόγκα μπροστά σε μια πράσινη οθόνη. Θέλουμε να διακρίνουμε ποιοι από αυτούς είναι άνδρες ή γυναίκες. Κάθε εικόνα μετατράπηκε σε μια μονοδιάστατη σειρά με την εύρεση του περιγράμματος και τη μέτρηση της απόστασης του περιγράμματος απο το κέντρο. Έχουμε 300 άτομα από δύο κατηγορίες ατόμων (άνδρες, γυναίκες) και 3000 ήταν το μεγεθος της δοκιμής αφου το κάθε ατομο άλλαξε 10 στάσεις.Ο αλγόριθμος αναγνώριζε το κέντρο του σώματος του και έφερνε 426 ευθείες προς κάθε κατεύθυνση στα πλαίσια του σώματος προκειμένου να δημιουργηθεί η χρονοσειρά. Ο τύπος πληροφοριών που παρέχει το πακέτο είναι εικόνες. Γνωρίζουμε ήδη απο το πακέτο δεδομένων πως ο πιο αποτελεσματικός κώδικας είναι ο BOSS με ποσοστό επιτυχίας 90,99%.

5.2 Σύνολο Μοτίβων (Bag of Patterns)

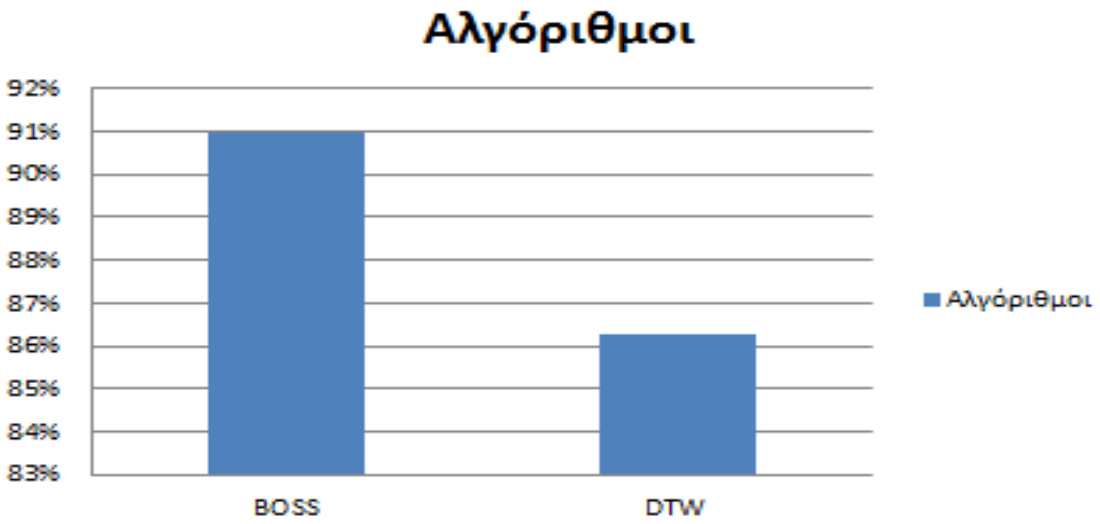
Αφού επεξεργαστήκαμε και τρέξαμε τον κώδικα BoP ο οποίος παρέχεται απο την σελίδα http://timeseriesclassification.com/dataset.php και χρειάζεται 3 τιμές ως αρχικοποίηση των παραμέτρων του οι οποίες ειναι : μετατροπή των τοπικών τμημάτων σε μια αναπαράσταση SAX που χρησιμοποιεί ένα μέγεθος παραθύρου w για να εξαγάγει σχέδια σε τοπικές υπακολουθίες, το μέγεθος του αλφάβητου και η διάσταση της σειράς των λέξεων.Δώσαμε 3 τιμές για την κάθε παράμετρο. Θέσαμε για διάσταση της σειράς με μήκος 32, το μέγεθος του αλφαβήτου 7 και το μέγεθος του παραθύρου 72. Οι ίδιες τιμές είχαν δωθεί και στον BOSS αφού θέλουμε τα αποτελέσματα να είναι συγκρίσιμα. Τρέχοντας λοιπόν τον κώδικα διαπιστώσαμε πως το ποσοστό επιτυχίας του είναι 83,2% επιβεβαιώνοντας πως ο BOSS ανταποκρίνεται καλύτερα σε αυτη τη διαδικασία. Στο γράφημα που ακολουθεί μπορούμε να διαπιστωσουμε την υπεροχή του πρώτου.



Εικόνα 5.1: Γράφημα αλγορίθμων

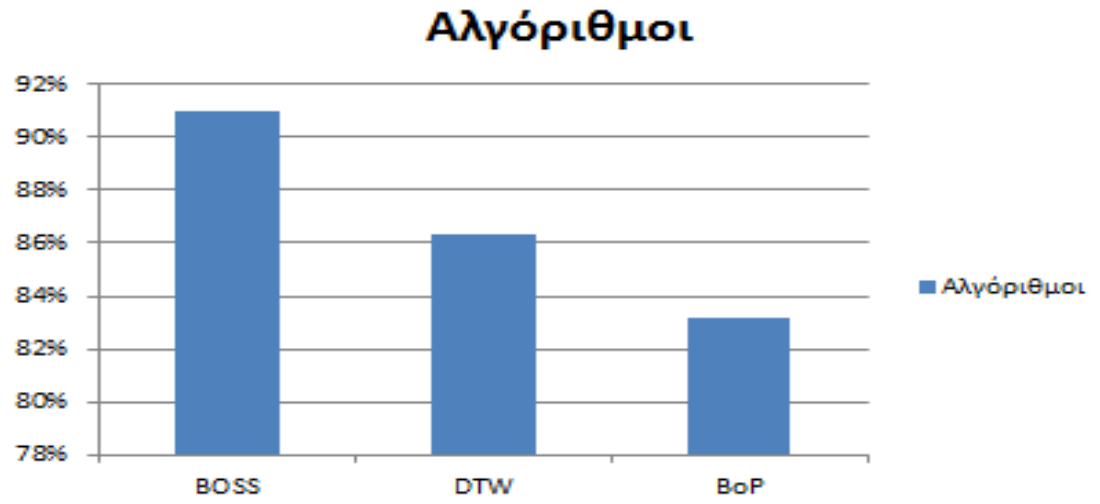
5.3 Δυναμική χρονική στρέβλωση (Dynamic Time Warping)

Παρακάτω επεξεργαζόμαστε και δουλεύουμε με τον κώδικα που χρησιμοποιείται για την πρόβλεψη του συνόλου δεδομένων Γιόγκα χρησιμοποιώντας τον αλγόριθμο δυναμικής χρονικής στρέβλωσης. Ο κώδικας παραχωρήθηκε απο τη διαδικτυακή σελίδα http://timeseriesclassification.com/dataset.php . Προκειμένου τα αποτελέσματα να είναι συγκρίσιμα δώσαμε στις παραμέτρους τις ίδιες αρχικές τιμές με τον BOSS και τον BoP. Το μήκος της λέξης πήρε την τιμή 32, το μέγεθος του αλφαβήτου ηταν 7, το μήκος του παραθύρου 72 και το πλάτος του 23. Εκτελώντας τα δεδομένα μέσω της μεθόδου DTW, έχουμε 86,3 % ποσοστό επιτυχίας κάτι που επιβεβαίωνει και πάλι πως ο BOSS αποτελεί την καλύτερη επιλογή αλγορίθμου για το συγκεκριμένο σύνολο δεδομένων. Παραθέτουμε λοιπόν το ακόλουθο γράφημα για να γίνει αντιληπτή η διαφορά αυτή:



Εικόνα 5.2: Γραφήματα BOSS-BoP

Ακολουθεί ένα συγκεντρωτικό διάγραμμα το οποίο παίρνει τα ποσοστά επιτυχίας και των τριών αλγορίθμων που περιγράψαμε και το οποίο μας δείχνει και πάλι πως ο BOSS αποτελεί τον πιο αποτελεσματικό αλγόριθμο ο οποίος αναγνωρίζει με μεγαλύτερη επιτυχία το φύλο του εξασκούμενου.



Εικόνα 5.3:Συγκεντρωτικά αποτελέσματα

Βιβλιογραφία

1. Mike Izbicki," Converting images into time series for data mining", October 2011(https://izbicki.me/blog/converting-images-into-time-series-for-data-mining.html)
2. Anthony Bagnall, Jason Lines, William Vickers και Eamonn Keogh,"Το κατάστημα ταξινόμησης UEA & UCR" ,2015(http://timeseriesclassification.com/index.php , http://timeseriesclassification.com/algorithm.php ,http://timeseriesclassification.com/dataset.php)
3. Abdullah Mueen," Time series motif discovery: dimensions and applications", February 2014 (https://onlinelibrary.wiley.com/doi/pdf/10.1002/widm.1119)
4. Page Research Optimus, "What is time series analysis?" (https://www.researchoptimus.com/article/what-is-time-series-analysis.php)
5. Ralitsa Golemanova,"7 image recognition uses of the future", Tech insider, July 2017 (https://imagga.com/blog/7-image-recognition-uses-of-the-future/)
6. Shan Zhao ,"Real-Time Hand Gesture Recognition Using Finger Segmentation", The Scientific World Journal, June 2014, (https://www.hindawi.com/journals/tswj/2014/267872/)
7. Khosrow Hassibi and Jeff Johnstone, "Image Recognition Revolution & Applications", THOUGHT LEADERSHIP W H I T E P A P E R, 2016 (http://analytics.rsystems.com/wp-content/uploads/2017/09/Image-Recognition-WhitePaper-Rsystems.pdf)
8. Zhiguang Wang and Tim Oates, "Imaging Time-Series to Improve Classification and Imputation", June 2015 (https://arxiv.org/pdf/1506.00327.pdf)
9. Hosoda Hiroto, "Automotive Image Recognition Processor "IMAPCAR" ", SoC Products, 2016 (https://www.nec.com/en/global/techrep/journal/g06/n05/pdf/t060507.pdf)