Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Συμμορφωτική πρόβλεψη ακολουθιακών δεδομένων με δένδρα απόφασης

Ελευθέριος Γ. Λουβερδής
ΜΕΣ 16006

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Απρίλιος 2019

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

Συμμορφωτική πρόβλεψη ακολουθιακών δεδομένων με δένδρα απόφασης

Ελευθέριος Γ. Λουβερδής
ΜΕΣ 16006

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Απρίλιος 2019

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Επίκουρος Καθηγητής, Πελέκης Νικόλαος (Επιβλέπων)
- Καθηγητής, Θεοδωρίδης Ιωάννης
- Αναπληρωτής Καθηγητής, Κοφίδης Ελευθέριος

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

Conformal prediction of sequential data with decision trees

By

Eleftherios G. Louverdis
MES 16006

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
April 2019

This thesis was approved unanimously be the three – member committee appointed by the Department of Statistics and Actuarial Science, University of Piraeus, in accordance with the rules of the MSc program in Applied Statistics.

Committee members were:

- Assistant Professor, Pelekis Nikolaos (Supervisor)
- Professor, Theodoridis Ioannis
- Associate Professor, Kofidis Eleftherios

Approval of this thesis from the Department of Statistics and Actuarial Science, University of Piraeus, does not imply any endorsement of the opinions of the author.

Στους γονείς μου και στον αδερφό μου,

Ευχαριστίες

Ξεκινώντας, θα ήθελα να ευχαριστήσω την οικογένεια μου για όλη την στήριξη που μου δίνει όλα αυτά τα χρόνια σε ότι και αν κάνω.

Επιπρόσθετα, θα ήθελα να πω ένα μεγάλο ευχαριστώ στον καθηγητή μου και επιβλέπων κ. Νίκο Πελέκη, ο οποίος μου προσέφερε καθοριστικές συμβουλές σε όλη την περίοδο εκπόνησης της παρούσας εργασίας και ήταν πάντα διαθέσιμος να βοηθήσει σε ότι πρόβλημα και αν εμφανιζόταν.

Κλείνοντας θα ήθελα να ευχαριστήσω και συνολικά τους καθηγητές του ΠΜΣ Εφαρμοσμένης Στατιστικής, για τις γνώσεις που μου παρείχαν κατά την διάρκεια του μεταπτυχιακού προγράμματος.

Περίληψη

Σκοπός της παρούσας εργασίας είναι να ασχοληθεί με την μελέτη της συμμορφωτικής πρόβλεψης με τη χρήση δέντρων απόφασης. Πιο συγκεκριμένα, στοχεύει στην εφαρμογή της μεθόδου σε ένα πρόβλημα πρόβλεψης του τελικού αποτελέσματος αγώνων ποδοσφαίρου για δυο συγκεκριμένες ομάδες, με την χρήση των αντίστοιχων δέντρων κατηγοριοποίησης.

Στην εφαρμογή της μεθόδου χρησιμοποιήθηκαν δυο σύνολα ακολουθιακών δεδομένων (ένα για κάθε ομάδα). Το πρώτο σύνολο δεδομένων αναφέρεται στην ακολουθία αγώνων που έπαιξε η ομάδα της Άρσεναλ τη φετινή σεζόν (2018-2019) μέχρι τις 20 Μαρτίου και αποτελείται από 30 παρατηρήσεις, ενώ το δεύτερο σύνολο αναφέρεται στους αγώνες της Μπαρτσελόνα την ίδια σεζόν και αποτελείται από 27 παρατηρήσεις.

Το συμπέρασμα που προέκυψε μετά την χρήση του αντίστοιχου αλγορίθμου κατηγοριοποίησης είναι ότι το μοντέλο πρόβλεψης με την χρήση των δέντρων κατηγοριοποίησης δυσκολεύεται να προβλέψει το τελικό αποτέλεσμα των αγώνων, πράγμα που δείχνει τον βαθμό δυσκολίας για ασφαλείς προβλέψεις στο πεδίο του ποδοσφαίρου.

Abstract

The purpose of this paper is to deal with the study of conformal prediction with the use of some decision trees. More specifically, it aims at applying the method to a problem of prediction the final outcome of football matches for two specific teams, using the corresponding categorization trees.

Two sets of sequential data (one for each team) were used in the implementation of the method. The first dataset refers to Arsenal's series of matches this season (2018-2019) by March 20 and consists of 30 observations, while the second dataset refers to Barcelona's matches in the same season and consists of 27 observations.

The conclusion that emerged after the use of the corresponding categorization algorithm is that the predictive model with the use of categorization trees has difficulty in predicting the final outcome of the matches, indicating the degree of difficulty for safe predictions in the field of football.

Περιεχόμενα

ΚΕΦΑΛΑΙΟ 1

Ο ΑΙΩΝΑΣ ΤΩΝ ΔΕΔΟΜΕΝΩΝ

Η πρώτη ενότητα είναι ουσιαστικά το εισαγωγικό μέρος της συγκεκριμένης εργασίας. Στην πρώτη υποενότητα αναφέρεται ο βασικός σκοπός της παρούσας εργασίας. Στην δεύτερη υποενότητα γίνεται μια αναφορά στα διαφορετικά είδη δεδομένων που υπάρχουν, ενώ η τρίτη υποενότητα ασχολείται με τους διαφορετικούς τρόπους ανάλυσης αυτών των δεδομένων. Τέλος, η τελευταία υποενότητα του πρώτου κεφαλαίου αναφέρεται στο πόσο σημαντική είναι η ανάλυση των δεδομένων για έναν οργανισμό.

1.1. Σκοπός της εργασίας

Όπως υποδηλώνει και ο τίτλος της παρούσας εργασίας, κύριος σκοπός είναι η περιγραφή της συμμορφωτικής πρόβλεψης με δέντρα απόφασης σε ένα σύνολο ακολουθιακών δεδομένων.

Πιο συγκεκριμένα στο αρχικό μέρος της εργασίας γίνεται μια αναφορά στα διαφορετικά είδη ακολουθιακών δεδομένων που μπορεί να συναντήσουμε και τις ιδιαιτερότητες που μπορεί να έχουν τα συγκεκριμένα δεδομένα. Στην συνέχεια ακολουθεί το κύριο μέρος της εργασίας που είναι η περιγραφή του μοντέλου της συμμορφωτικής πρόβλεψης με δέντρα απόφασης και ποιες είναι οι κύριες διαφορές του με τα υπόλοιπα μοντέλα πρόβλεψης. Κλείνοντας, στο τελευταίο μέρος της εργασίας γίνεται η εφαρμογή της συμμορφωτικής πρόβλεψης με δέντρα κατηγοριοποίησης σε δυο σύνολα ακολουθιών που προέρχονται από τον χώρο του ποδοσφαίρου.

1.2. Είδη δεδομένων

Ο 21^{ος} αιώνας έχει χαρακτηριστεί πολλές φορές και ως ο αιώνας των δεδομένων. Στην καθημερινότητα μας ακούμε συχνά ότι πολλοί οργανισμοί χρησιμοποιούν δεδομένα για να κάνουν μελλοντικές προβλέψεις. Τράπεζες χρησιμοποιούν δεδομένα για να αξιολογήσουν καλυτέρα τους πελάτες τους και να ελαχιστοποιήσουν τους κινδύνους τους, μετεωρολογικές εταιρείες χρησιμοποιούν παλαιότερα δεδομένα για να κάνουν μελλοντικές προβλέψεις για τον καιρό, τα ίδια τα κράτη χρησιμοποιούν μεγάλους όγκους δεδομένων για να προγραμματίσουν όσο το δυνατόν καλύτερα τους προϋπολογισμούς τους και πολλά ακόμα παραδείγματα.

Τι είναι όμως τελικά ένα δεδομένο; Ως δεδομένο μπορεί να χαρακτηριστεί ένα διακριτό στοιχείο που είναι αποτέλεσμα παρατήρησης. Πιο συγκεκριμένα μπορεί να είναι παρατηρήσεις που έχουν γίνει για έναν άνθρωπο (για παράδειγμα το ύψος ή το βάρος του), μπορεί να είναι τιμές που αναφέρονται σε ένα γεγονός (τιμή στην κλίμακα ρίχτερ που είχε ένας σεισμός) ή ακόμα μπορεί να είναι και τιμές που αποτυπώνονται για ένα αντικείμενο (κυβικά και ίπποι ενός αυτοκίνητου).

Οι συλλογές δεδομένων, όπως γίνεται αντιληπτό, μπορεί να έχουν οποιαδήποτε μορφή και να διαφέρουν σε μεγάλο βαθμό μεταξύ τους. Πιο συγκεκριμένα μπορεί να έχουν την μορφή λέξεων, αριθμών ή συμβόλων. Επιπρόσθετα, μπορεί να έχουν και περισσότερο περίπλοκες μορφές και να αποτυπώνονται ως γραφήματα, φωτογραφίες ή βίντεο.

Γίνεται λοιπόν προφανές ότι στον αιώνα που ζούμε, στον οποίο οι νέες τεχνολογίες γεννιούνται και εξελίσσονται με τρομακτικά γρήγορους ρυθμούς, ο όγκος δεδομένων που είναι διαθέσιμος προς ανάλυση είναι τεράστιος. Ένα χαρακτηριστικό παράδειγμα τέτοιου είδους τεχνολογίας είναι τα κοινωνικά δίκτυα (social media). Μέσω των κοινωνικών δικτύων οι χρήστες ανταλλάσσουν ιδέες και απόψεις για θέματα που τους αφορούν, ανεβάζουν φωτογραφίες και βίντεο, ενημερώνουν τους φίλους τους για το μέρος που βρίσκονται ή βρέθηκαν οποιαδήποτε χρονική στιγμή και τέλος μπορούν να αντιδράσουν θετικά (Like) ή αρνητικά (Dislike) σε οποιοδήποτε θέμα τους ενδιαφέρει.

Οι επιστήμες που κατά κύριο λόγο ενδιαφέρονται και έχουν τα αντίστοιχα εργαλεία για την ανάλυση των δεδομένων είναι η Στατιστική και η Πληροφορική. Αυτές οι δυο επιστήμες προσπαθούν κάθε φορά να διακρίνουν τα χαρακτηριστικά που έχουν τα εκάστοτε δεδομένα και με την κατάλληλη επεξεργασία να επιλέξουν την μέθοδο πρόβλεψης που θα έχει τα καλύτερα αποτελέσματα. Όπως είναι προφανές, αυτό δεν είναι πάντα μια εύκολη διαδικασία καθώς τα δεδομένα γίνονται ολοένα και περισσότερο πολύπλοκα. Παρόλα αυτά υπάρχουν αρκετά μοντέλα πρόβλεψης που καταφέρνουν να προσεγγίσουν πολύ καλά τις αντίστοιχες περιπτώσεις και να έχουν μεγάλο ποσοστό επιτυχίας.

Στην παρούσα εργασία, τα δεδομένα που μας ενδιαφέρει να επεξεργαστούμε και να αναλύσουμε σε κάποιο βαθμό είναι τα ακολουθιακά δεδομένα. Τα ακολουθιακά δεδομένα διακρίνονται κυρίως σε τρεις μεγάλες κατηγορίες, οι οποίες είναι οι ακολουθίες δεδομένων, οι χρονοσειρές και τα χωροχρονικά δεδομένα. Κάθε μια από αυτές τις κατηγορίες έχει διαφορετικά χαρακτηριστικά και διαφορετικό τρόπο προσέγγισης. Πιο λεπτομερής αναφορά για τα ακολουθιακά δεδομένα γίνεται στην δεύτερη ενότητα της παρούσας εργασίας.

1.3. Μέθοδοι ανάλυσης δεδομένων

Η έννοια της ανάλυσης των δεδομένων (Data Analysis) αναφέρεται στην διαδικασία κατά την οποία εξετάζονται τα χαρακτηριστικά των υπό εξέταση δεδομένων, γίνονται οι απαραίτητες μετατροπές και καθαρισμοί των δεδομένων και τέλος επιλέγεται το καταλληλότερο μοντέλο ανάλυσης και πρόβλεψης. Αυτή η διαδικασία έχει ως κύριο στόχο την εξόρυξη όσο το δυνατόν περισσότερης πληροφόρησης για τα δεδομένα έτσι ώστε να βοηθήσει στην συνέχεια στην λήψη αποφάσεων για τους αντίστοιχους οργανισμούς.

Όπως αναφέραμε και νωρίτερα, η ανάλυση των δεδομένων είναι μια δύσκολη διαδικασία η οποία απαιτεί μεγάλη προσοχή σε όλα τα βήματα της έτσι ώστε να έχουμε τα βέλτιστα αποτελέσματα. Ανάλογα με τα χαρακτηριστικά των δεδομένων έχουν αναπτυχθεί και οι κατάλληλες μέθοδοι πρόβλεψης για κάθε περίπτωση. Παρακάτω θα αναφέρουμε κάποιες από αυτές με κάποιους μαθηματικούς ορισμούς.

Ένα πρωταρχικό μέρος της ανάλυσης των δεδομένων είναι αυτό που ονομάζεται ως περιγραφική στατιστική. Η περιγραφική στατιστική αναφέρεται στο κομμάτι της ανάλυσης που ενδιαφέρεται για την εξόρυξη πληροφορίας για ένα δείγμα ή έναν πληθυσμό. Επιπλέον, η περιγραφική στατιστική προσπαθεί με σύντομο και κατανοητό τρόπο να παρουσιάσει τα κύρια χαρακτηριστικά και ιδιομορφίες ενός δείγματος. Η συγκεκριμένη μέθοδος περιλαμβάνει και κάποια μέτρα που διαχωρίζονται σε δυο κατηγορίες. Αυτές οι κατηγορίες είναι τα μέτρα θέσης και τα μέτρα διασποράς. Στα μέτρα θέσης συγκαταλέγονται η μέση τιμή, ο σταθμικός μέσος,

η διάμεσος και η επικρατούσα τιμή. Ενώ στα μέτρα διασποράς ανήκουν το εύρος, το ενδοτεταρτημοριακό εύρος, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

Μια δεύτερη μέθοδος που χρησιμοποιείται συνέχεια στην επιστήμη της στατιστικής, είναι η ανάλυση παλινδρόμησης. Σε αυτή την μέθοδο βασικός στόχος είναι ο έλεγχος πιθανών συσχετίσεων μεταξύ μιας εξαρτημένης μεταβλητής, η οποία συμβολίζεται συνήθως με το γράμμα Υ, και μιας ή περισσότερων ανεξάρτητων μεταβλητών οι οποίες συμβολίζονται με το γράμμα Χ. Για την απλή γραμμική παλινδρόμηση ένας μαθηματικός τύπος έχει την μορφή $Y = \beta_0 + \beta_1 X + \varepsilon$. Τα β_0 και β_1 είναι κάποιες άγνωστοι παράμετροι ενώ το ε είναι το τυχαίο σφάλμα της παλινδρόμησης. Στην πολλαπλή γραμμική παλινδρόμηση το αντίστοιχο μοντέλο έχει την μορφή $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_N X_N + \varepsilon$. Στην συγκεκριμένη τεχνική μοντελοποίησης γίνονται και κάποιες υποθέσεις οι οποίες είναι απαραίτητες για την εφαρμογή του μοντέλου. Αυτές έχουν να κάνουν με την ύπαρξη γραμμικής σχέσης των μέσων τιμών της Υ με τα διαφορά επίπεδα της Χ, οι κατανομές της Υ έχουν την ίδια διασπορά για όλα τα επίπεδα της Χ (ομοσκεδαστικότητα), οι τιμές της Υ που αντιστοιχούν στα διάφορα επίπεδα της Χ είναι ανεξάρτητες μεταξύ τους και ότι η κατανομή της Υ για όλα τα επίπεδα της Χ είναι κανονική.

(Μάρκος Κούτρας, Χαράλαμπος Ευαγγελάρας, 2010)

Μια ακόμα τεχνική στο πεδίο της ανάλυσης των δεδομένων είναι και η κατηγοριοποίηση. Η κατηγοριοποίηση είναι η διαδικασία κατά την οποία ένα νέο στοιχείο ταξινομείται με βάση ένα προκαθορισμένο σύνολο κατηγοριών. Κύριος σκοπός της κατηγοριοποίησης είναι η δημιουργία ενός αποτελεσματικού μοντέλου που θα χρησιμοποιηθεί για την ταξινόμηση μελλοντικών δεδομένων. Η διαδικασία της κατηγοριοποίησης είναι η ακόλουθη. Αρχικά, δημιουργείται το μοντέλο με βάση ένα σύνολο παραδειγμάτων που έχουν κατηγοριοποιηθεί στο παρελθόν. Αυτά τα παραδείγματα ονομάζονται δεδομένα εκπαίδευσης (training data). Τα δεδομένα εκπαίδευσης αναλύονται από τον κατάλληλο αλγόριθμο κατηγοριοποίησης με σκοπό τον σχηματισμό του μοντέλου. Στη συνέχεια ακολουθεί το μέρος της αξιολόγησης του μοντέλου. Σε αυτό το σημείο χρησιμοποιούνται κάποια δοκιμαστικά δεδομένα (test data) για τον υπολογισμό της ακρίβειας του μοντέλου. Αν το μοντέλο κριθεί έγκυρο και αποτελεσματικό τότε μπορεί στην συνέχεια να χρησιμοποιηθεί για την πρόβλεψη μελλοντικών περιπτώσεων.

Στο πεδίο τώρα της πληροφορικής, οι αναλυτές προσπαθούν με κάποιους αλγορίθμους να ερμηνεύσουν αυτόν τον τεράστιο όγκο δεδομένων. Πιο συγκεκριμένα, οι εφαρμογές έχουν δείξει ότι για μεγάλα σύνολα δεδομένων οι καταλληλότερες μέθοδοι ανάλυσης είναι αυτές που χρησιμοποιούν κάποιους αλγορίθμους εξόρυξης γνώσης. Αυτοί οι αλγόριθμοι προσπαθούν με αυτόματο ή ημιαυτόματο τρόπο να δημιουργήσουν ένα αποτελεσματικό μοντέλο το οποίο θα εξάγει όσο το δυνατόν μεγαλύτερη πληροφορία και θα έχει κατανοητή δομή για τον αναγνώστη. Κάποιοι από τους πιο γνωστούς αλγορίθμους εξόρυξης γνώσης είναι οι C4.5, k-means, support vector machines, Apriori, EM, PageRank, AdaBoost, kNN, Naive Bayes και ο Cart.

Στην παρούσα εργασία η τεχνική που θα αναλύσουμε και θα εφαρμόσουμε σε ένα σύνολο δεδομένων είναι η συμμορφωτική πρόβλεψη (Conformal prediction). Περισσότερα για την συμμορφωτική πρόβλεψη αλλά και για κάποιους αλγορίθμους με δέντρα απόφασης αναφέρονται στην τρίτη ενότητα της εργασίας.

1.4. Σημασία της ανάλυσης των δεδομένων

Ο όρος της ανάλυσης των δεδομένων αναφέρεται σε διάφορες ποιοτικές και ποσοτικές διαδικασίες που έχουν ως βασικό σκοπό να βελτιώσουν τον τρόπο λειτουργίας ή την παραγωγή ενός οργανισμού. Πιο συγκεκριμένα, κάθε οργανισμός προσπαθεί να βρει την βέλτιστη λύση έτσι ώστε να μεγιστοποιεί το κέρδος του με το μικρότερο δυνατό κόστος. Συνεπώς, κάθε επιχείρηση προσπαθεί να διαχειριστεί με τον καλύτερο τρόπο όλα αυτά τα διαφορετικά δεδομένα που έχει στην διάθεση της με στόχο την βελτίωση της προοπτικής του κέρδους της.

Αν εμβαθύνουμε λίγο στην διαδικασία της ανάλυσης των δεδομένων μπορούμε να καταλάβουμε τον λόγο για τον οποίο είναι τόσο σημαντική. Όπως αναφέραμε και στην αρχή της ενότητας, ο αιώνας που ζούμε χαρακτηρίζεται από μια συνεχή εμφάνιση νέων τεχνολογιών που παράγουν με την σειρά τους καινούργια δεδομένα διαθέσιμα προς ανάλυση. Λαμβάνοντας λοιπόν υπόψη ότι κάθε οργανισμός λειτουργεί ανταγωνιστικά με τους υπολοίπους όμοιους του, καταλαβαίνουμε ότι προσπαθεί συνεχώς να βρει πληροφόρηση για τον τρόπο που κινείται η αγορά νωρίτερα από τους ανταγωνιστές του. Το μεγαλύτερο μέρος αυτής της πληροφορίας κρύβεται μέσα σε αυτούς τους τεράστιους όγκους δεδομένων. Είναι κατανοητό λοιπόν, ότι όλοι οι μεγάλοι οργανισμοί έχουν μια ομάδα αναλυτών που έχουν ως κύριο σκοπό την εξόρυξη γνώσης η οποία είναι άγνωστη μέχρι εκείνη την στιγμή στο σύνολο της αγοράς. Η εξόρυξη αυτής της γνώσης θα βοηθήσει την εταιρεία στην δημιουργία καινοτόμων μεθόδων που θα την οδηγήσει σε πλεονεκτική θέση έναντι των ανταγωνιστών της.

Επιπρόσθετα, ακόμα και στην περίπτωση που η εταιρεία δεν καταφέρει να εξορύξει κάποια γνώση η οποία θα της δημιουργήσει πλεονεκτήματα έναντι των ανταγωνιστικών εταιρειών, είναι πολύ σημαντική η ανάλυση αυτών των δεδομένων έτσι ώστε να βρίσκεται στο ίδιο επίπεδο γνώσης με τις υπόλοιπες.

Αν ξεφύγουμε λίγο από το κομμάτι του ανταγωνισμού των εταιρειών, θα καταλάβουμε ότι η διαδικασία της ανάλυσης των δεδομένων είναι πολύ σημαντική και σε τομείς που έχουν ως σκοπό την βελτίωση της ζωής του ανθρώπου. Σημαντικοί τομείς είναι για παράδειγμα η υγεία και η παιδεία μιας κοινωνίας. Στο κομμάτι της υγείας, για παράδειγμα, μπορεί να χρησιμοποιηθούν παλαιότερα δεδομένα για ανάλυση, τα οποία θα βοηθήσουν στο κομμάτι της πρόληψης και της θεραπείας μιας ασθένειας. Συγκεκριμένα, μπορούν να εντοπιστούν οι συνήθειες των ανθρώπων που οδηγούν στην εμφάνιση κάποιας μορφής καρκίνου και οι οποίες είναι άγνωστες μέχρι εκείνη την στιγμή. Έτσι με την κατάλληλη ενημέρωση υπάρχει η προοπτική της αναθεώρησης και διακοπής τέτοιων συνήθειων. Στο κομμάτι τώρα της παιδείας, μπορεί ένα κράτος να συγκρίνει δεδομένα άλλων κρατών με διαφορετικά συστήματα παιδείας και να επιλέξει αυτό που θεωρεί ότι είναι καλύτερο. Δηλαδή μπορεί για παράδειγμα να δει πόσες ώρες την ημέρα ένας μαθητής μπορεί να είναι αποδοτικός και να προσπαθεί να μην ξεπερνά αυτές τις ώρες διδασκαλίας. Ακόμα μπορεί να συγκρίνει αν είναι καλύτερο για τους μαθητές, ειδικότερα στις πιο μικρές ηλικίες, να λύνουν τις ασκήσεις την ώρα του μαθήματος και την υπόλοιπη ώρα της ημέρας να ασχολούνται με άλλες δραστηριότητες.

ΚΕΦΑΛΑΙΟ 2

ΑΚΟΛΟΥΘΙΑΚΑ ΔΕΔΟΜΕΝΑ

Στην ενότητα αυτή γίνεται μια σύντομη αναφορά στο είδος των δεδομένων που αφορούν την συγκεκριμένη εργασία. Το πρώτο μέρος της ενότητας αναφέρεται σε ακολουθίες δεδομένων, το δεύτερο μέρος αναφέρεται στην μεγάλη κατηγορία των χρονοσειρών και τέλος η τρίτη υποενότητα αφορά σε δεδομένα του χωροχρόνου. Για κάθε μια από τις κατηγορίες ακολουθιάκων δεδομένων δίνονται κάποιοι ορισμοί μαζί με κάποια παραδείγματα.

2.1. Ακολουθίες

Με τον όρο ακολουθία (sequence) αναφερόμαστε σε μια λίστα ή ένα σετ δεδομένων το οποίο χαρακτηρίζεται από κάποιας μορφής διάταξη. Μια ακολουθία μπορεί να αποτελείται από αριθμούς ή από σύμβολα τα οποία είναι ταξινομημένα και ακολουθούν κάποιο συγκεκριμένο μοτίβο. Πέρα από την ιδιότητα της διάταξης μια ακολουθία μπορεί να είναι άπειρη ή πεπερασμένη ανάλογα με το πλήθος των όρων της.

Παραδείγματα ακολουθιών συναντάμε συνέχεια στην καθημερινότητα μας, κάποιες από τις οποίες μπορεί να είναι απλές όπως η ακολουθία των περιττών αριθμών 1,3,5,7,9,… η οποία είναι μια άπειρη ακολουθία ή να έχουν λίγο πιο πολύπλοκη δομή όπως είναι για παράδειγμα η ακολουθία Φιμπονάτσι (Fibonacci sequence) η οποία είναι η 0,1,1,2,3,5,8,13,21,34,55,89,144,… στην οποία βλέπουμε ότι κάθε επόμενος αριθμός είναι το άθροισμα των δυο προηγούμενων. Οι τρεις τελείες που ακολουθούν τις δυο παραπάνω ακολουθίες είναι ένας όχι και τόσο επίσημος συμβολισμός για να δείξει ότι οι ακολουθίες αποτελούνται από άπειρους όρους (άπειρες ακολουθίες).

Όπως αναφέρθηκε και προηγουμένως μια ακολουθία μπορεί να αποτελείται και από ονομαστικές τιμές (σύμβολα). Παραδείγματα τέτοιων ακολουθιών είναι για παράδειγμα μια πεπερασμένη ακολουθία της μορφής * * - * * - * * - οπού παρατηρούμε ότι μετά από δυο διαδοχικές ‘νιφάδες’ (*) ακολουθεί μια παύλα (-). Άλλο παράδειγμα μπορεί να είναι μια άπειρη ακολουθία από γεωμετρικά σχήματα όπως αυτή που ακολουθεί ▲■●▲■●▲■●….

Ένας βασικός ορισμός για τις ακολουθίες είναι αυτός που αναφέρεται στην ακολουθία πραγματικών αριθμών.

Ορισμός 1

Ακολουθία πραγματικών αριθμών καλείται κάθε συνάρτηση

$$\alpha: \mathbb{N} \to \mathbb{R} : n \to a(n) := \alpha_n$$

(Αθανάσιος Σ. Κυριαζής, 2004)

Όπως αναφέρθηκε και παραπάνω, δεδομένα ακολουθιών μπορούμε να συναντήσουμε σε πολλούς τομείς ή δραστηριότητες της καθημερινότητας μας. Ένας τομέας μπορεί να είναι

ακολουθίες προϊόντων που αγοράζονται από πελάτες σε ένα super market. Μέσα από αυτές τις ακολουθίες μπορεί κάποιος να προσπαθήσει να βρει συσχετίσεις μεταξύ των προϊόντων έτσι ώστε να βοηθήσει στο κομμάτι του marketing και της προώθησης προϊόντων της αντίστοιχης εταιρείας. Άλλο πεδίο έρευνας των ακολουθιών είναι οι ακολουθίες επισκέψεων των χρηστών ενός Η/Υ σε συγκεκριμένες ιστοσελίδες (views). Με βάση αυτή τη δραστηριότητα του χρήστη, οι διαφημιστικές εταιρείες μπορούν να εξάγουν πολύ σημαντικά συμπεράσματα έτσι ώστε να βελτιώσουν την προωθητική πολιτική της επιχείρησης τους. Για παράδειγμα ένας χρήστης μπορεί να κάνει μια έρευνα αγοράς κάποιου ρούχου μέσα από τον Η/Υ του. Με βάση τις επισκέψεις που έχει κάνει σε συγκεκριμένες εικόνες ρούχων, η διαφημιστική εταιρεία μπορεί να κάνει τις απαραίτητες συγκρίσεις μεταξύ των εικόνων (π.χ. το χρώμα ή το στυλ των ρούχων), και να προωθήσει μια διαφημιστική εικόνα του προϊόντος στον συγκεκριμένο χρήστη που να προσεγγίζει την έρευνα αγοράς του. Ακόμη ένα πεδίο έρευνας που εμφανίζονται ακολουθίες δεδομένων είναι η βιοπληροφορική. Η βιοπληροφορική συνδυάζει διαφορετικές επιστήμες όπως η πληροφορική, τα μαθηματικά, η στατιστική και η μηχανική με βασικό σκοπό την εξόρυξη συσχετίσεων μέσα από ακολουθίες βιολογικών δεδομένων. Πιο συγκεκριμένα χρησιμοποιούνται διάφορες μέθοδοι εξόρυξης δεδομένων όπως είναι τα νευρωτικά και μπεϋζιανά δίκτυα ή άλλοι γενετικοί αλγόριθμοι.

(Βικιπαιδεια Βιοπληροφορικη ,2019)

Ένας δεύτερος ορισμός για τις ακολουθίες δεδομένων ο οποίος ταιριάζει καλύτερα στα παραπάνω παραδείγματα, είναι ο ακόλουθος.

Ορισμός 2

Ακολουθία είναι μια ταξινομημένη λίστα από σετ αντικειμένων (itemsets)

$$S = (I_1, I_2, \dots, I_n) \; με \; I_\kappa \subseteq I \; (1 \leq \kappa \leq n)$$

όπου S : Ακολουθία και I : Σετ αντικειμένων

με $I = (i_1, i_2, \dots, i_m)$

(Philippe Fournier-Viger, Jerry Chun-Wei Lin, Rage Uday Kiran, Yun Sing Koh, Rincy Thomas, 2017)

Η ανάλυση ακολουθιακών δεδομένων γίνεται με διάφορες τεχνικές εξόρυξης γνώσης. Αρχικά αυτά τα δεδομένα αποθηκεύονται σε μια βάση δεδομένων και εν συνέχεια μπορεί να γίνει η ανάλυση τους με σκοπό την εξόρυξη σημαντικών πληροφοριών που θα βοηθήσουν στην μελλοντική λήψη αποφάσεων του εκάστοτε οργανισμού.

Κάποιες από τις σημαντικότερες τεχνικές εξόρυξης γνώσης δίνονται παρακάτω.

- Συσταδοποίηση (Clustering)
- Κατηγοριοποίηση (Classification)
- Ανάλυση ακραίων περιπτώσεων (Outlier Analysis)
- Εξόρυξη προτύπων από ακολουθίες (Sequential Pattern Mining)

Μια περισσότερο λεπτομερής αναφορά για τις παραπάνω τεχνικές, δίνεται συνολικά στο τέλος της παρούσας ενότητας.

2.2. Χρονοσειρές

Ένα δεύτερο είδος ακολουθιακών δεδομένων που χρησιμοποιείται ευρέως στην εξόρυξη γνώσης είναι οι χρονοσειρές (Time series). Με τον όρο χρονοσειρά αναφερόμαστε σε μια σειρά συνεχών δεδομένων που έχουν ως κύρια χαρακτηριστικά την καθορισμένη διάταξη των παρατηρήσεων διαχρονικά και την καθορισμένη εξάρτηση μεταξύ των διαδοχικών παρατηρήσεων της σειράς (Λευτέρης Ι. Θαλασσινός, 1991).

Μια βασική διαφορά των χρονοσειρών από τις ακολουθίες, που αναφερθήκαμε στην προηγούμενη ενότητα, είναι ότι οι τιμές μιας χρονοσειράς παίρνουν αποκλειστικά αριθμητικές τιμές σε αντίθεση με τις ακολουθίες που μπορεί να έχουν και ονομαστικές τιμές (σύμβολα).

Παραδείγματα χρονοσειρών μπορεί να είναι δεδομένα που προέρχονται από τιμές μετοχών (π.χ. η τιμή κλεισίματος της μετοχής σε κάποιο διεθνές χρηματιστήριο), τιμές θερμοκρασίας (π.χ. οι τιμές μέσης θερμοκρασίας για κάποιο συγκεκριμένο χρονικό διάστημα) και διάφορες ακόμα παρόμοιες περιπτώσεις. Οι τιμές αυτές μπορεί να αναφέρονται σε ημερήσια, εβδομαδιαία, μηνιαία ή ετήσια δεδομένα.

Στο πεδίο έρευνας των χρονοσειρών έχουν αναπτυχθεί διάφορες μέθοδοι πρόβλεψης οι οποίες έχουν πλεονεκτήματα και μειονεκτήματα. Κάποιες από τις πιο σημαντικές είναι ο Απλός Κινητός Μέσος, η Απλή Εκθετική Εξομάλυνση, η Εκθετική Εξομάλυνση με προσαρμογή στην Τάση (μέθοδος Holt), καθώς επίσης και η ανάλυση που γίνεται σε μοντέλα ARIMA.

Στη συνέχεια ακολουθεί μια περιληπτική αναφορά για τις μεθόδους που αναφέρθηκαν παραπάνω.

Απλός Κινητός Μέσος (Simple Moving Average)

Ο απλός κινητός μέσος είναι μια μέθοδος πρόβλεψης χρονολογικών σειρών η οποία βασίζεται στον αριθμητικό μέσο όρο των m πιο πρόσφατων παρατηρήσεων μιας χρονοσειράς (m περίοδοι). Ο λόγος για τον οποίο συμβαίνει αυτό είναι ότι οι πιο πρόσφατες παρατηρήσεις μιας χρονοσειράς είναι καταλληλότερες για την διαδικασία της πρόβλεψης σε σχέση με τις παλαιότερες παρατηρήσεις.

Μια μαθηματική εξίσωση για τον υπολογισμό του απλού κινητού μέσου είναι η παρακάτω:

$$\hat{Y}_{t+1} = \frac{1}{m}\sum_{j=1}^{m} Y_{t-j+1} = \frac{1}{m}(Y_t + Y_{t-1} + \cdots + Y_{t-m+1})$$

όπου

$\hat{Y}_{t+1}$: η πρόβλεψη για την περίοδο (t + 1)

m : ο αριθμός των περιόδων που χρησιμοποιούνται για τον υπολογισμό της τιμής του μέσου όρου

Απαραίτητη προϋπόθεση για την εφαρμογή ενός μοντέλου απλού κινητού μέσου είναι να γνωρίζουμε την τιμή του m.

(Χρήστος Ν. Αγιακλόγλου, Γεώργιος Σ. Οικονόμου, 2004).

Απλή Εκθετική Εξομάλυνση (Simple Exponential Smoothing)

Στην απλή εκθετική εξομάλυνση η πρόβλεψη μιας χρονοσειράς βασίζεται σε κάποιο σταθμικό μέσο όρο, ο οποίος δίνει πολύ μεγαλύτερη βαρύτητα στις πιο πρόσφατες παρατηρήσεις σε σχέση με την βαρύτητα που δίνεται στις πιο απομακρυσμένες στο παρελθόν.

Η μαθηματική εξίσωση της συγκεκριμένης μεθόδου είναι :

$$\hat{Y}_{t+1} = aY_t + a(1-a)Y_{t-1} + a(1-a)^2Y_{t-2} + \cdots$$

όπου

α : σταθερά εξομάλυνσης, $0 \leq \alpha \leq 1$

Από την παραπάνω σχέση γίνεται φανερό ότι όσο πιο μεγάλη είναι η τιμή που παίρνει η σταθερά εξομάλυνσης α, τόσο μεγαλύτερη βαρύτητα δίνεται στις πιο πρόσφατες παρατηρήσεις και αντίστοιχα τόσο μικρότερη στις πιο απομακρυσμένες (Χρήστος Ν. Αγιακλόγλου, Γεώργιος Σ. Οικονόμου, 2004).

Εκθετική Εξομάλυνση με προσαρμογή στην Τάση : Μέθοδος Holt (Exponential Smoothing adjusted for Trend)

Η συγκεκριμένη μέθοδος πρόβλεψης χρησιμοποιείται όταν εμφανίζεται τάση στις παρατηρήσεις μιας χρονοσειράς. Βασικά στοιχεία ενός τέτοιου μοντέλου πρόβλεψης είναι οι δύο παράμετροι εξομάλυνσης. Η πρώτη παράμετρος αναφέρεται στην εξομάλυνση των τιμών της χρονοσειράς και συμβολίζεται με το γράμμα α, ενώ η δεύτερη παράμετρος χρησιμοποιείται για την εξομάλυνση της τάσης και συμβολίζεται με το γράμμα β.

Οι μαθηματικές σχέσεις που χρησιμοποιούνται για την εξομάλυνση των τιμών της χρονοσειράς και της τάσης αντίστοιχα είναι οι ακόλουθες:

Για τις τιμές της χρονοσειράς:

$$A_t = aY_t + (1-a)(A_{t-1} + T_{t-1})$$

με το α να παίρνει τιμές μεταξύ του 0 και του 1 ($0 \leq \alpha \leq 1$).

Όπου A_t : οι εξομαλυνθείσες τιμές της χρονοσειράς για t = 2,3,…,n. Για t = 1 έχουμε την αρχική συνθήκη $A_1 = Y_1$.

Για την τάση:

$$T_t = \beta(A_t - A_{t-1}) + (1-\beta)T_{t-1}$$

με $0 \leq \beta \leq 1$.

Όπου T_t: οι εξομαλυνθείσες τιμές της τάσης για t = 2,3,…,n. Για t = 1 έχουμε την αρχική συνθήκη $T_1 = 0$.

Αν θέλουμε να προβλέψουμε την $\hat{Y}_{t+h}$ για την h μελλοντική περίοδο χρησιμοποιούμε τον τύπο

$$\hat{Y}_{t+h} = A_t + hT_t \ , \qquad \text{όπου h} = 1,2,3,\ldots$$

(Χρήστος Ν. Αγιακλόγλου, Γεώργιος Σ. Οικονόμου, 2004)

Ανάλυση μοντέλων ARIMA

Μια ακόμη πολύ σημαντική μέθοδος πρόβλεψης στο πεδίο έρευνας των χρονοσειρών είναι τα μοντέλα ARIMA (Autoregressive integrated moving average). Τα μοντέλα ARIMA εφαρμόζονται κυρίως σε περιπτώσεις όπου στα δεδομένα της χρονοσειράς εμφανίζονται ενδείξεις μη-στασιμότητας. Μια χρονοσειρά είναι μη-στάσιμη όταν παρατηρούνται σημαντικές διαφοροποιήσεις στις διακυμάνσεις των τιμών της με το πέρασμα του χρόνου.

Ένας μαθηματικός ορισμός για τα μοντέλα ARIMA είναι ο ακόλουθος:

Μια διαδικασία x_t λέγεται ότι είναι ένα μοντέλο ARIMA(p,d,q) αν

$$\nabla^d x_t = (1 - B)^d x_t$$

είναι ένα μοντέλο ARMA(p,q). Γενικά το μοντέλο γράφεται ως

$$\varphi(B)(1 - B)^d x_t = \theta(B) w_t$$

όπου

φ(B) : the autoregressive operator

θ(B) : the moving average operator

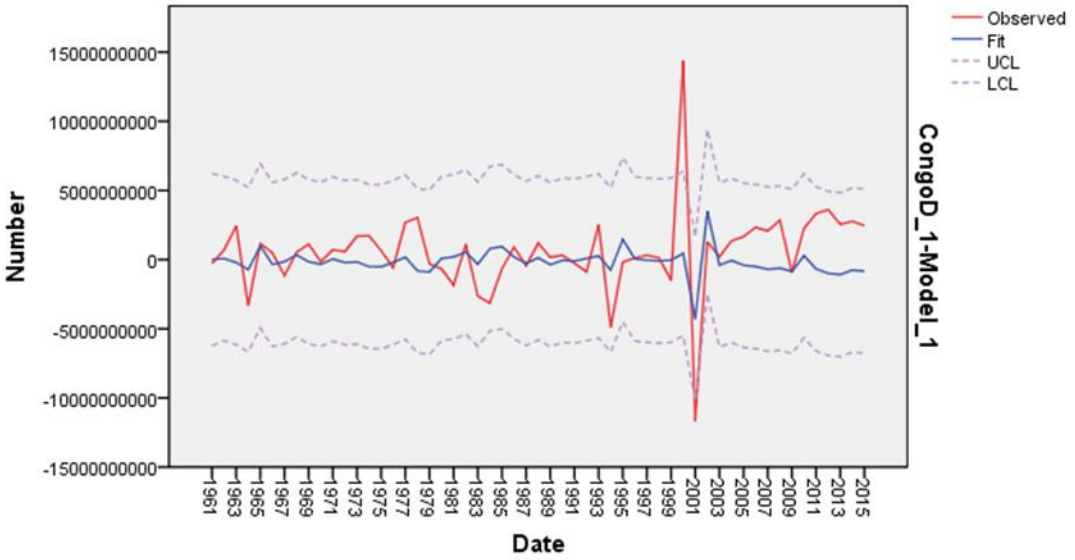
B : the backshift operator

(Robert H. Shumway, David S. Stoffer, 2006)

Παρακάτω ακολουθεί ένα παράδειγμα εφαρμογής της ανάλυσης ARIMA στα δεδομένα μιας χρονοσειράς. Τα δεδομένα που χρησιμοποιήθηκαν στην ανάλυση μας αναφέρονται στο ακαθάριστο εγχώριο προϊόν (gross domestic product) της Λαϊκής Δημοκρατίας του Κονγκό την περίοδο από το 1960 έως και το 2015 (σε δολάριο Αμερικής).

Αρχικά ελέγχθηκε αν η χρονοσειρά είναι στάσιμη ή όχι. Εφόσον υπήρχαν ενδείξεις μη στασιμότητας, το επόμενο βήμα ηταν να μετασχηματίσουμε τη χρονοσειρά έτσι ώστε να μην παραβιάζεται η υπόθεση της στασιμότητας.

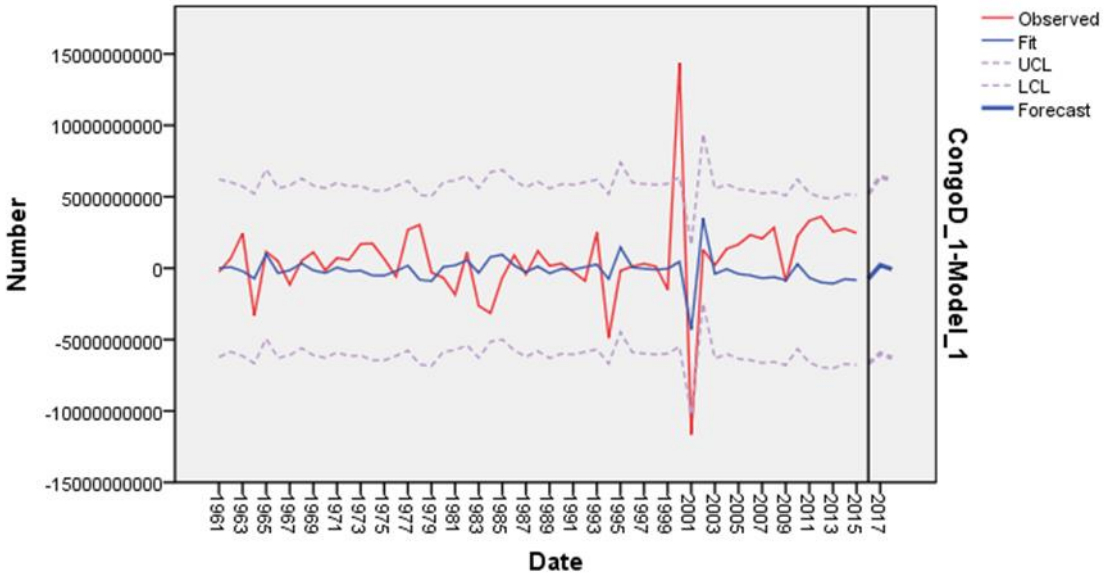
Μετά την σχετική διαδικασία για την επιλογή του βέλτιστου μοντέλου, προέκυψε το ακόλουθο γράφημα.



Όσον αφορά το διάγραμμα βλέπουμε τις τιμές της χρονοσειράς μαζί με τις εκτιμήσεις τους και τα κάτω και άνω όρια του διαστήματος εμπιστοσύνης των εκτιμήσεων.

Παρακάτω φαίνονται τα αποτελέσματα των προβλέψεων μέχρι το 2018 για τις μελλοντικές τιμές της χρονοσειράς καθώς επίσης και ένα διάγραμμα στο οποίο περιλαμβάνονται και οι προβλέψεις των τιμών της χρονοσειρας.

Predicted values	Year
-734342832	2016
219616396	2017
-65679624	2018



2.3. Χωροχρονικά δεδομένα

Τα χωροχρονικά δεδομένα (Spatio -Temporal Data) είναι μια ειδική περίπτωση δεδομένων που συνδυάζουν τις τρεις διαστάσεις του ευκλείδειου χώρου (μήκος, πλάτος, ύψος), με μια τέταρτη διάσταση η οποία είναι ο χρόνος.

Η προσέγγιση των γνωρίσματων των χωροχρονικών δεδομένων διαφέρει από την αντίστοιχη προσέγγιση που γίνεται σε μη-χωροχρονικά δεδομένα. Μια ποιοτική διαφορά των χωροχρονικών δεδομένων από τα υπόλοιπα δεδομένα είναι η παρουσία κάποιας μορφής εξάρτησης μεταξύ των μετρήσεων που οφείλεται σε μεγάλο βαθμό στις τέσσερις διαστάσεις των χωροχρονικών δεδομένων.

Για παράδειγμα, πολλές από τις μεθοδολογίες εξόρυξης γνώσης από ΄΄κλασσικά΄΄ δεδομένα (μη-χωροχρονικά), βασίζονται στις δυο πολύ σημαντικές υποθέσεις οι οποίες είναι ότι οι περιπτώσεις των δεδομένων είναι ανεξάρτητες καθώς επίσης και ότι ακολουθούν την ίδια κατανομή (συνήθως κανονική). Ωστόσο, οι παραπάνω υποθέσεις πολλές φορές παραβιάζονται όταν έχουμε να διαχειριστούμε χωροχρονικά δεδομένα. Αυτό συμβαίνει γιατί οι περιπτώσεις τέτοιου είδους δεδομένων είναι δομικά συσχετισμένες μεταξύ τους στο πλαίσιο των διαστάσεων του χώρου και του χρόνου, και παρουσιάζουν ποικίλες ιδιότητες σε διαφορετικές χωρικές περιοχές και χρονικές περιόδους. Συνεπώς, πρέπει να δίνεται ιδιαίτερη προσοχή στην διαχείριση τέτοιων δεδομένων, γιατί αν παραβλέψουμε τις συσχετίσεις που υπάρχουν, μπορούμε να οδηγηθούμε σε κακή ακρίβεια πρόβλεψης και σε λάθος ερμηνεία των αποτελεσμάτων. Είναι φανερό λοιπόν, ότι υπάρχει η ανάγκη δημιουργίας μιας γενικής (συνολικής) μεθοδολογίας που να μπορεί να εφαρμοστεί σε διαφορετικά προβλήματα μεταξύ τους. (Gowtham Atluri, Anuj Karpatne, Vipin Kumar, 2017)

Υπάρχουν πολλά πεδία στα οποία υπάρχει πληθώρα χωροχρονικών δεδομένων. Παρακάτω γίνεται μια σύντομη περιγραφή τέτοιων περιπτώσεων.

Κλιματική επιστήμη

Στην συγκεκριμένη επιστήμη γίνεται συλλογή ιστορικών ή πιο πρόσφατων δεδομένων που σχετίζονται με τις συνθήκες που επικρατούν στην ατμόσφαιρα αλλά και σε θαλάσσιες περιοχές. Τέτοια παραδείγματα μπορεί να είναι η θερμοκρασία ή η υγρασία που επικρατεί σε μια περιοχή σε ένα συγκεκριμένο χρονικό διάστημα. Βασικός σκοπός στην μελέτη τέτοιων δεδομένων είναι, μέσα από συσχετίσεις και ομοιότητες, να καταλάβουμε το φυσικό περιβάλλον στο οποίο ζούμε και, μέσω των προβλέψεων, να προετοιμαστούμε για μελλοντικές δυσμενείς συνθήκες που είναι πιθανό να εμφανιστούν.

Νευροεπιστήμη

Η νευροεπιστήμη είναι η επιστημονική μελέτη του νευρικού συστήματος. Κάποια παραδείγματα νευροεπιστήμης είναι το ηλεκτροεγκεφαλογράφημα και η μαγνητο-εγκεφαλογραφία. Σκοπός στη μελέτη τέτοιων δεδομένων είναι η κατανόηση βασικών αρχών του εγκεφάλου και εν συνεχεία ο καθορισμός πιθανών διαταραχών κάτω από κανονικές συνθήκες (ψυχικές διαταραχές). Η ανίχνευση τέτοιων διαταραχών είναι χρήσιμη στον σχεδιασμό διαγνωστικών διαδικασιών και στην ανάπτυξη της κατάλληλης θεραπείας για τους ασθενείς.

Περιβαλλοντική επιστήμη

Στο πεδίο αυτής της επιστήμης γίνεται μελέτη της ποιότητας του αέρα και πιο συγκεκριμένα γίνονται μετρήσεις για ρύπους όπως, το μονοξείδιο του άνθρακα, το διοξείδιο του αζώτου, το διοξείδιο του θείου ή άλλων βλαβερών για την υγειά μικροσωματιδίων. Ταυτόχρονα γίνεται μελέτη και της ποιότητας του νερού με μετρήσεις που σχετίζονται με παράγοντες το pH και η θολότητα. Ακόμα υπάρχει η δυνατότητα να γίνουν μετρήσεις που είναι σχετικές με την ηχορύπανση. Σκοπός της συγκεκριμένης επιστήμης είναι η ανίχνευση πιθανών αλλαγών στα επίπεδα της μόλυνσης, ο εντοπισμός των παραγόντων που είναι υπεύθυνοι για την ρύπανση και ο σχεδιασμός αποτελεσματικών πολιτικών που θα έχουν ως στόχο την μείωση των διαφορετικών τύπων μόλυνσης.

Επιδημιολογία

Στα νοσοκομεία υπάρχουν αποθηκευμένα ηλεκτρονικά αρχεία με έναν τεράστιο όγκο δεδομένων υγείας. Αυτά τα δεδομένα είναι κυρίως διαγνώσεις που έχουν γίνει σε ασθενείς όπου κάθε διάγνωση έχει το δικό της χωρικό και χρονικό σημείο τα οποία συσχετίζονται μεταξύ τους. Σκοπός στη μελέτη δεδομένων υγείας είναι η ανίχνευση χωροχρονικών μοτίβων σε διαφορετικές ασθένειες και η μελέτη της εξάπλωσης μιας επιδημίας. Τέτοιου είδους δεδομένα μπορούν να συνδυαστούν με κλιματικά δεδομένα με σκοπό την ανίχνευση συσχετίσεων μεταξύ παραγόντων του περιβάλλοντος και της δημοσίας υγείας. Η ανακάλυψη τέτοιων συσχετίσεων θα βοηθήσει στην ανάπτυξη πολιτικών για την βελτίωση της ζωής του πληθυσμού.

Μέσα κοινωνικής δικτύωσης

Οι χρήστες των μέσων κοινωνικής δικτύωσης, όπως είναι για παράδειγμα το Facebook και το Instagram, μοιράζονται τις εμπειρίες τους σε συγκεκριμένο χώρο και χρόνο. Αυτό έχει ως συνέπεια την μεταφορά μεγάλου όγκου ειδήσεων (δεδομένων) σε ελάχιστο χρονικό διάστημα. Αυτή η πολύ γρήγορη μεταφορά δεδομένων μπορεί να βοηθήσει τους αρμοδίους οργανισμούς

στην δημιουργία κατάλληλων ενεργειών έτσι ώστε να βοηθήσουν το κοινωνικό σύνολο. Για παράδειγμα μέσω τέτοιων δικτύων μπορεί να διαδοθεί πολύ γρηγορά η εξάπλωση ενός φυσικού φαινομένου όπως είναι μια πυρκαγιά και να γίνουν οι κατάλληλες ενέργειες αντιμετώπισης της.

Δυναμική κυκλοφορίας

Δεδομένα τέτοιου τύπου αναφέρονται κυρίως σε διαδρομές που γίνονται από πελάτες με την υπηρεσία του ταξί. Σε αυτά τα δεδομένα υπάρχει συνεχής πληροφόρηση σχετικά με τα χωροχρονικά στοιχεία μιας διαδρομής. Είναι γνωστός δηλαδή, ο χώρος και ο χρόνος που ένας πελάτης επιβιβάστηκε ή αποβιβάστηκε από ένα ταξί καθώς επίσης και όλα τα ενδιάμεσα σημεία. Με αυτόν τον τρόπο μπορούμε να δούμε την χωρική μετακίνηση του πληθυσμού σε συνάρτηση με τον χρόνο αλλά και την επιρροή εξωτερικών παραγόντων όπως είναι η κυκλοφοριακή συμφόρηση και ο καιρός. Αυτό θα βοηθήσει στον σχεδιασμό ενεργειών που θα έχουν ως στόχο την μείωση της κυκλοφοριακής συμφόρησης. Ακόμα δίνεται η δυνατότητα να σχεδιαστούν τεχνικές που θα ανιχνεύουν την βέλτιστη διαδρομή. (Gowtham Atluri, Anuj Karpatne, Vipin Kumar, 2017)

Τα χωροχρονικά δεδομένα, όπως αναφέρθηκε και νωρίτερα, διαφέρουν αρκετά από τα συνηθισμένα δεδομένα. Υπάρχουν δυο γενικές ιδιότητες που εμφανίζονται στα χωροχρονικά δεδομένα. Η πρώτη είναι η αυτοσυσχέτιση ενώ η δεύτερη είναι η ετερογένεια.

Αυτοσυσχέτιση (Auto-correlation) : Σε τομείς που περιλαμβάνονται χωροχρονικά δεδομένα, οι παρατηρήσεις που γίνονται σε κοντινές τοποθεσίες και κοντινές χρονικές στιγμές παρουσιάζουν κάποιας μορφής εξάρτηση μεταξύ τους (π.χ. ίδια θερμοκρασία σε κοντινά μέρη την ίδια χρονική περίοδο). Αυτό έχει ως αποτέλεσμα οι κλασσικοί αλγόριθμοι εξόρυξης γνώσης, που υποθέτουν ανεξαρτησία μεταξύ των παρατηρήσεων, να μην είναι κατάλληλοι σε τέτοιες εφαρμογές.

Ετερογένεια (Heterogeneity) : Μια άλλη υπόθεση που γίνεται στους κλασσικούς αλγορίθμους εξόρυξης γνώσης είναι η ομοιογένεια (στασιμότητα) των παρατηρήσεων, γεγονός που σημαίνει ότι κάθε παρατήρηση ανήκει στον ίδιο πληθυσμό και συνεπώς ακολουθεί την ίδια κατανομή. Ωστόσο, στα χωροχρονικά δεδομένα εμφανίζεται συχνά ετερογένεια (μη-στασιμότητα) μεταξύ των παρατηρήσεων τόσο με βάση τον χώρο όσο και με βάση τον χρόνο σε πολλά επίπεδα.

Τα χωροχρονικά δεδομένα ανάλογα με κάποιες ιδιαιτερότητες που μπορεί να παρουσιάζουν, ταξινομούνται και στην αντίστοιχη κατηγορία. Κάποιες από τις βασικότερες κατηγορίες που κατανέμονται τα συγκεκριμένα δεδομένα, ανάλογα με τον τύπο τους, είναι οι παρακάτω:

- Δεδομένα συμβάντων (Event data)
- Δεδομένα τροχιάς (Trajectory data)
- Δεδομένα σημείου αναφοράς (Point reference data)
- Δεδομένα raster (Raster data)

Στα δεδομένα συμβάντων και τροχιάς καταγράφονται παρατηρήσεις από διακριτά γεγονότα ενώ στα δεδομένα σημείου αναφοράς και στα raster παίρνουμε πληροφορία από συνεχή ή διακριτά χωροχρονικά πεδία.

Παρακάτω ακολουθεί μια αναφορά για τις κατηγορίες των χωροχρονικών δεδομένων.

Δεδομένα συμβάντων

Κάθε χωροχρονικό δεδομένο χαρακτηρίζεται από την τοποθεσία (σημείο) που έλαβε χώρα αλλά και την αντίστοιχη χρονική στιγμή που συνέβη. Η συλλογή χωροχρονικών σημείων που αναφέρονται σε ένα συμβάν ονομάζεται χωρικό μοτίβο σημείων (spatial point pattern) [Gatrell et al. 1996]. Ένα χωρικό μοτίβο σημείων απεικονίζεται σε ένα δισδιάστατο σύστημα συντεταγμένων όπου το l_i αναφέρεται στην τοποθεσία και το t_i στον χρόνο.

Δεδομένα τροχιάς

Τροχιά ονομάζεται ένα μονοπάτι στο οποίο μετακινούνται τα κινητά αντικείμενα κατά τη διάρκεια του χρόνου. Τα συγκεκριμένα δεδομένα συλλέγονται με την βοήθεια κάποιων αισθητήρων στο κινούμενο αντικείμενο που μεταδίδουν συνεχώς πληροφορία σχετικά με την τοποθεσία του αντικειμένου με το πέρασμα του χρόνου.

Δεδομένα σημείου αναφοράς

Ένα δεδομένο σημείου αναφοράς αποτελείται από μετρήσεις ενός συνεχούς χωροχρονικού πεδίου όπως είναι για παράδειγμα η θερμοκρασία ή ένας πληθυσμός που μεταβάλλονται με το πέρασμα του χρόνου. Τα δεδομένα σημείου αναφοράς λέγονται και γεωστατιστικά δεδομένα.

Δεδομένα raster

Στα δεδομένα raster, οι μετρήσεις ενός συνεχούς ή διακριτού χωροχρονικού πεδίου καταγράφονται σε σταθερές θέσεις στον χώρο και σε σταθερές χρονικές στιγμές. Αυτό έρχεται σε αντίθεση με τα δεδομένα σημείου αναφοράς οπού οι χωροχρονικές περιοχές αναφοράς μπορεί να αλλάζουν συνεχώς την τοποθεσία τους με το πέρασμα του χρόνου και να συλλέγουν καταγραφές σε διαφορετικές χρονικές στιγμές.

Όσον αφορά τον τρόπο που ταξινομείται η εξόρυξη γνώσης από χωροχρονικά δεδομένα, έχουμε τις ακόλουθες κατηγορίες:

- Συσταδοποίηση (Clustering)
- Προγνωστική μάθηση (Predictive learning)
- Ανίχνευση αλλαγής (Change detection)
- Εξόρυξη συχνών προτύπων (Frequent pattern mining)
- Ανίχνευση ανωμαλιών (Anomaly detection)
- Εξόρυξη συσχέτισης (Relationship mining)

(Gowtham Atluri, Anuj Karpatne, Vipin Kumar, 2017)

Παρακάτω δίνεται μια περιληπτική αναφορά των τεχνικών εξόρυξης γνώσης, οι οποίες μπορούν να χρησιμοποιηθούν και στην περίπτωση των ακολουθιακών δεδομένων.

Συσταδοποίηση (Clustering)

Συσταδοποίηση είναι η διαδικασία κατά την οποία ομαδοποιείται ένα σύνολο δεδομένων με τέτοιο τρόπο έτσι ώστε τα αντικείμενα που ανήκουν στην ίδια ομάδα (στην ίδια συστάδα) να έχουν μεγαλύτερη ομοιότητα μεταξύ τους σε σχέση με τα αντικείμενα των άλλων ομάδων (L.V. Bijuraj, 2013). Στην βιβλιογραφία έχουν προταθεί πολλές τεχνικές συσταδοποίησης κάποιες από τις οποίες έχουν να κάνουν με ιεραρχικούς αλγόριθμους, αλγόριθμους ανταγωνιστικής μάθησης, διαμεριστικούς αλγόριθμους και κάποιους centroid-based αλγόριθμους.

Κατηγοριοποίηση (Classification)

Η κατηγοριοποίηση είναι μια ακόμα τεχνική εξόρυξης γνώσης η οποία χρησιμοποιείται για την πρόβλεψη της ομάδας που ανήκει ένα αντικείμενο μέσα από ένα σύνολο δεδομένων. Η πρόβλεψη της ομάδας κάθε νέας παρατήρησης βασίζεται σε προηγούμενες παρατηρήσεις των οποίων η ομάδα είναι ήδη γνωστή. Κάποιες πολύ γνωστές μεθοδολογίες κατηγοριοποίησης είναι τα δέντρα απόφασης, τα νευρωνικά δίκτυα, η Bayesian κατηγοριοποίηση και διάφορες ακόμα τεχνικές.

Ανάλυση ακραίων περιπτώσεων (Outlier Analysis)

Η ανάλυση ακραίων περιπτώσεων είναι μια διαδικασία η οποία προσπαθεί να ανακαλύψει αντικείμενα από ένα σύνολο δεδομένων τα οποία έχουν πολύ διαφορετική συμπεριφορά από την αναμενομένη (Krishna Modi, Prof Bhavesh Oza, 2017). Οι τεχνικές που χρησιμοποιούνται για την ανίχνευση ακραίων παρατηρήσεων δεν διαφέρουν σε μεγάλο βαθμό από αυτές που έχουν αναφερθεί ήδη και έχουν να κάνουν με τεχνικές ανίχνευσης συστάδων, μπεϋζιανά δίκτυα, μοντέλα Markov (HMM) και διάφορες άλλες μεθόδους.

Εξόρυξη προτύπων από ακολουθίες (Sequential Pattern Mining)

Η ανάλυση προτύπων από ακολουθιακά δεδομένα είναι μια τεχνική εξόρυξης γνώσης η οποία έχει ως στόχο την ανίχνευση παρόμοιων (πανομοιότυπων) προτύπων μέσα στα δεδομένα έτσι ώστε να αναγνωρίσει πιθανές σχέσεις μεταξύ τους (Kapil Sharma, Ashok, Dr. Harish Rohil, 2014). Κάποιες τεχνικές που χρησιμοποιούνται στον συγκεκριμένο τομέα εξόρυξης γνώσης είναι οι αλγόριθμοι GSP, FreeSpan, PrefixSpan και διάφοροι άλλοι αλγόριθμοι.

ΚΕΦΑΛΑΙΟ 3

ΣΥΜΜΟΡΦΩΤΙΚΗ ΠΡΟΒΛΕΨΗ

ΜΕ ΔΕΝΤΡΑ ΑΠΟΦΑΣΗΣ

Στην ενότητα 3 της εργασίας γίνεται μια εισαγωγή στο πεδίο της συμμορφωτικής πρόβλεψης με την χρήση κάποιων δέντρων απόφασης. Πιο συγκεκριμένα στην πρώτη υποενότητα παρουσιάζονται κάποιες έννοιες και ορισμοί που αναφέρονται στην συμμορφωτική πρόβλεψη. Στην συνέχεια γίνεται μια σύντομη αναφορά στα δέντρα απόφασης. Ενώ στο τελευταίο μέρος της ενότητας γίνεται μια σύνδεση της συμμορφωτικής πρόβλεψης με κάποιους αλγόριθμους των δέντρων απόφασης.

3.1. Συμμορφωτική πρόβλεψη

Η συμμορφωτική πρόβλεψη (Conformal Prediction) είναι μια ακόμη μέθοδος στο πεδίο της εξόρυξης γνώσης. Βασικό χαρακτηριστικό της συγκεκριμένης μεθόδου είναι ότι χρησιμοποιεί προηγούμενη εμπειρία που υπάρχει από παλαιοτέρα δεδομένα για να καθορίσει ακριβή επίπεδα εμπιστοσύνης σε μελλοντικές προβλέψεις.

Η συμμορφωτική πρόβλεψη έχει σχεδιαστεί για μια απευθείας σύνδεσης (on-line) τοποθέτηση στην οποία τα y (labels) προβλέπονται με διαδοχικό τρόπο. Δηλαδή, κάθε label αποκαλύπτεται πριν την πρόβλεψη του επόμενου. Ακόμη, μπορεί να εφαρμοστεί σε μια πληθώρα από μεθόδους όπως για παράδειγμα στη μέθοδο του κοντινότερου γείτονα (nearest – neighbor method), σε μοντέλα ανάλυσης παλινδρόμησης και φυσικά σε αλγορίθμους δέντρων απόφασης (decision trees algorithms).

Η συγκεκριμένη μέθοδος πρόβλεψης διαφέρει αρκετά από μεθόδους που έχουν ως βασική υπόθεση την ανεξαρτησία των περιπτώσεων. Η συμμορφωτική πρόβλεψη είναι μια συσσωρευτική μέθοδος πράγμα που σημαίνει ότι οι διαδοχικές προβλέψεις που γίνονται βασίζονται στην συσσωρευτική πληροφόρηση. Επομένως, γίνεται αντιληπτό ότι στα αντίστοιχα σετ δεδομένων υπάρχει μεγάλος βαθμός συσχετίσεων μεταξύ των διαδοχικών περιπτώσεων.

Για παράδειγμα αν θέλουμε μέσω του x_n να προβλέψουμε το y_n ,τότε είναι απαραίτητο να χρησιμοποιήσουμε και όλα τα προηγούμενα ζεύγη $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$.

Για να είναι έγκυρη μια 90% on-line μέθοδος, είναι αναγκαία προϋπόθεση το 90% όλων αυτών των διαδοχικών προβλέψεων να είναι σωστές.

Γενικά, ένας 90% εκτιμητής συμμόρφωσης είναι αυτός που χρησιμοποιεί τις προηγούμενες περιπτώσεις $(x_1, y_1), \dots, (x_{n-1}, y_{n-1})$ και το x_n για να δώσει ένα σετ όπως το $\Gamma^{0.1}((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), x_n)$, όπου η πρόβλεψη θα περιέχει το y_n. Αν ο εκτιμητής είναι έγκυρος, τότε η πρόβλεψη $y_n \in \Gamma^{0.1}((x_1, y_1), \dots, (x_{n-1}, y_{n-1}), x_n)$ θα έχει 90% πιθανότητα πριν την παρατήρηση των προηγούμενων περιπτώσεων. (Glenn Shafer, Vladimir Vovk, 2008)

Όπως συμβαίνει σε όλα τα μοντέλα πρόβλεψης, αυτό που μας ενδιαφέρει κατά κύριο λόγο εκτός από την εγκυρότητα τους, είναι και ο βαθμός της αποτελεσματικότητας τους. Ένα μοντέλο είναι αποτελεσματικό αν δίνει σχετικά μικρά διαστήματα πρόβλεψης και συνεπώς περισσότερη πληροφόρηση για την υπό μελέτη περίπτωση.

Επιπλέον είναι σημαντικό να αναφέρουμε ότι όταν χρησιμοποιούμε την συμμορφωτική πρόβλεψη δεν μας ενδιαφέρει η κατανομή που μπορεί να έχουν τα δεδομένα (π.χ. κανονική κατανομή) έτσι ώστε να έχουμε μια έγκυρη συμμορφωτική περιοχή πρόβλεψης. Αντίθετα η αποτελεσματικότητα της συμμορφωτικής πρόβλεψης εξαρτάται από την κατανομή των δεδομένων αλλά και από το μέτρο μη συμμόρφωσης στο οποίο θα αναφερθούμε στην συνέχεια.

Δυο βασικά στοιχεία της συμμορφωτικής πρόβλεψης είναι το μέτρο μη συμμόρφωσης (nonconformity measure) και ο συμμορφωτικός αλγόριθμος (conformal algorithm). Παρακάτω ακολουθούν κάποιοι ορισμοί για τα δυο αυτά μέτρα της συμμορφωτικής πρόβλεψης.

Μέτρο μη συμμόρφωσης (Nonconformity measure)

Το μέτρο μη συμμόρφωσης είναι μια μέθοδος η οποία μετράει πόσο ασυνήθιστη φαίνεται μια νέα περίπτωση σε σχέση με παλαιότερες περιπτώσεις. Πιο συγκεκριμένα πρόκειται για το πρωταρχικό μέρος της συμμορφωτικής πρόβλεψης και είναι μια συνάρτηση A(B,z) που μετράει πόσο διαφορετικό είναι ένα παράδειγμα z από τα παραδείγματα μιας τσάντας (bag) B.

Σε αυτό το σημείο είναι σημαντικό να δοθούν κάποιοι ορισμοί που θα διευκολύνουν την κατανόηση της παραπάνω πρότασης.

Τσάντα (Bag) ή αλλιώς multiset είναι μια συλλογή στοιχείων στην οποία επιτρέπεται η επανάληψη. Δηλαδή, περιέχει στοιχεία τα οποία δεν παρουσιάζουν κάποιας μορφής διάταξη και συνεπώς κάθε στοιχείο μπορεί να εμφανιστεί παραπάνω από μια φορά.

Ένας συμβολισμός μιας τσάντας είναι ο ακόλουθος.

Έστω μια λίστα $a_1, \dots, a_N$.

Για μια τσάντα που περιέχει τα στοιχεία της παραπάνω λίστας ο συμβολισμός της είναι ο εξής:

$$\{\!|a_1, \dots, a_N|\!\}$$

στον οποίο δεν υπάρχει κάποια πληροφόρηση σχετικά με την διάταξη της λίστας.

Συνεπώς ένας μαθηματικός ορισμός για το μέτρο μη συμμόρφωσης είναι ο παρακάτω:

$$A(B,z) \coloneqq d(\hat{z}(\mathrm{B}),\mathrm{z})$$

όπου

$\hat{z}$(B) : η μέθοδος εκτίμησης

d : η απόσταση

B : η τσάντα

A(B,z) : το μέτρο μη συμμόρφωσης

Ο καθορισμός του μέτρου μη συμμόρφωσης είναι ένα από τα σημαντικότερα σημεία στην συμμορφωτική πρόβλεψη. Η σωστή επιλογή του μέτρου μη συμμόρφωσης βοηθάει στο να έχουμε αποτελεσματικά διαστήματα πρόβλεψης. Όσο πιο μικρά είναι τα διαστήματα πρόβλεψης τόσο πιο αποτελεσματική θα είναι και η μέθοδος πρόβλεψης. Επιπλέον, πρέπει να γνωρίζουμε ότι για οποιοδήποτε μέτρο μη συμμόρφωσης και να επιλέξουμε, ο συμμορφωτικός αλγόριθμος θα δώσει έγκυρα αποτελέσματα τα οποία όμως μπορεί να μην είναι αποδοτικά. Το γεγονός αυτό δείχνει πόσο αναγκαία είναι η σωστή επιλογή του μέτρου μη συμμόρφωσης. (Glenn Shafer, Vladimir Vovk, 2008)

Συμμορφωτικός αλγόριθμος (Conformal algorithm)

Ένα ακόμη πολύ σημαντικό μέρος της συμμορφωτικής πρόβλεψης είναι ο συμμορφωτικός αλγόριθμος. Σκοπός του συμμορφωτικού αλγορίθμου είναι η μετατροπή ενός μέτρου μη συμμόρφωσης σε περιοχές πρόβλεψης.

Παρακάτω ακολουθεί ένας πιο αυστηρός ορισμός για τον συμμορφωτικό αλγόριθμο.

Δοθέντος ενός μέτρου μη συμμόρφωσης, ο συμμορφωτικός αλγόριθμος παράγει μια περιοχή πρόβλεψης Γ^{ε} για κάθε πιθανότητα σφάλματος ε. Η περιοχή Γ^{ε} είναι μια (1-ε) – περιοχή πρόβλεψης. Δηλαδή περιέχει το y με πιθανότητα τουλάχιστον ίση με 1-ε.

Για παράδειγμα ο συμβολισμός $\Gamma^{0.1}$ αναφέρεται για μια περιοχή πρόβλεψης σε 90% διάστημα εμπιστοσύνης.

Στην περίπτωση της ανάλυσης παλινδρόμησης όπου το y είναι ένας αριθμός, το $\Gamma^{0.1}$ είναι ένα διάστημα που περικλείει το $\hat{y}$. Ενώ σε μια περίπτωση κατηγοριοποίησης, όπου το y έχει έναν περιορισμένο αριθμό πιθανών τιμών, το $\Gamma^{0.1}$ μπορεί να αποτελείται από κάποιες από αυτές ή στην ιδανική περίπτωση μόνο από μια.

Όπως είναι κατανοητό, ο συμμορφωτικός αλγόριθμος υποθέτει ότι έχει γίνει νωρίτερα η επιλογή ενός μέτρου μη συμμόρφωσης για να ξεκινήσει την λειτουργία του. Όσο καταλληλότερο είναι το μέτρο μη συμμόρφωσης που θα έχει επιλεγεί τόσο καλύτερα αποτελέσματα θα δώσει και ο συμμορφωτικός αλγόριθμος.

Ακόμη είναι πολύ σημαντικό το γεγονός ότι οι περιοχές πρόβλεψης που δημιουργούνται από τον συμμορφωτικό αλγόριθμο δεν μεταβάλλονται καθώς το μέτρο μη συμμόρφωσης μετατρέπεται μονοτονικά. (Glenn Shafer, Vladimir Vovk, 2008)

Ανταλλαξιμότητα (Exchangeability)

Μια ακόμη πολύ σημαντική έννοια στο πεδίο της συμμορφωτικής πρόβλεψης είναι η ανταλλαξιμότητα. Η υπόθεση της ανταλλαξιμότητας είναι ελαφρώς ασθενέστερη (όχι τόσο αυστηρή) από την αντίστοιχη υπόθεση της ανεξαρτησίας των μεταβλητών που εξάγονται από μια κατανομή πιθανότητας.

Ένας απλοϊκός ορισμός της ανταλλαξιμότητας είναι ο εξής:

Έστω οι μεταβλητές $Z_1, \dots, Z_N$. Υποθέτουμε ότι για οποιαδήποτε συλλογή των N τιμών, οι N! διαφορετικές διατάξεις είναι το ίδιο πιθανόν να συμβούν. Σε αυτή την περίπτωση λέμε ότι οι μεταβλητές $Z_1, \dots, Z_N$ είναι ανταλλάξιμες.

Ένας ακόμη ορισμός για την ανταλλαξιμότητα ο οποίος είναι λίγο περισσότερο μαθηματικός και χρησιμοποιείται για την αποφυγή κάποιων τεχνικών προβλημάτων που εμφανίζονται στις κατανομές είναι ο παρακάτω που αναφέρεται στην ανταλλαξιμότητα με την χρήση μεταθέσεων:

Οι μεταβλητές $Z_1, \dots, Z_N$ είναι ανταλλάξιμες αν για κάθε μετάθεση τ των ακέραιων αριθμών 1,…,N, οι μεταβλητές $W_1, \dots, W_N$ όπου $W_i = Z_{\tau(i)}$, έχουν την ίδια κατανομή πιθανότητας με τις $Z_1, \dots, Z_N$.

Από τον παραπάνω ορισμό γίνεται φανερό ότι οι ανεξάρτητες τυχαίες μεταβλητές που ακολουθούν την ίδια κατανομή είναι ανταλλάξιμες.

Σε αυτό το σημείο είναι πολύ σημαντικό να γίνει κατανοητό ότι η έννοια της ανταλλαξιμότητας υποδηλώνει ότι οι μεταβλητές ακολουθούν την ίδια κατανομή, αλλά από την άλλη μεριά οι ανταλλάξιμες μεταβλητές δεν είναι απαραίτητα και ανεξάρτητες.

Επιπρόσθετα με τους παραπάνω ορισμούς, υπάρχει και ένας λίγο διαφορετικός ορισμός της ανταλλαξιμότητας. Ο συγκεκριμένος ορισμός ΄΄κοιτάζει΄΄ οπισθοδρομικά από μια κατάσταση κατά την οποία γνωρίζουμε τις μη-διατεταγμένες τιμές των $Z_1, \dots, Z_N$.

Με βάση την παραπάνω οπτική ακολουθούν τρεις ισοδύναμοι ορισμοί της κοινής κατανομής μιας ακολουθίας τυχαίων μεταβλητών $Z_1, \dots, Z_N$ κάθε ένας από τους οποίους μπορεί να χρησιμοποιηθεί ως ορισμός της ανταλλαξιμότητας.

1) Για κάθε τσάντα B μεγέθους N και για κάθε περίπτωση $\alpha_1, \dots, \alpha_N$, έχουμε ότι η

$$P_r(z_1 = \alpha_1 \& \dots \& z_N = \alpha_N \mid \{z_1, \dots, z_N\} = B)$$

είναι ίση με την πιθανότητα ότι τα διαδοχικά τυχαία σχέδια από την τσάντα B χωρίς αντικατάσταση παράγουν αρχικά το α_N, εν συνεχεία το α_{N-1}, και ούτω καθεξής, μέχρι το τελευταίο στοιχείο που θα έχει απομείνει στην τσάντα B να είναι το α_1.

2) Για κάθε n, $1 \le n \le N$, το Z_N είναι ανεξάρτητο από τα $Z_{N+1}, \dots, Z_N$ δοθέντος της τσάντας $\{z_1, \dots, z_n\}$ και για κάθε τσάντα B μεγέθους n, έχουμε

$$P_r\left(z_n = \alpha \mid \{z_1, \dots, z_N\} = B\right) = \frac{k}{n} ,$$

όπου το k είναι ο αριθμός των φορών όπου το α εμφανίζεται στην τσάντα B.

3) Για κάθε τσάντα B μεγέθους N και για κάθε περίπτωση $\alpha_1, \dots, \alpha_N$ έχουμε

$$P_r(z_1 = \alpha_1 \& \dots \& z_N = \alpha_N \,|\, \{z_1, \dots, z_N\} = B) = \begin{cases} \frac{n_1! \dots n_k!}{N!} & \text{αν } B = \{\alpha_1, \dots, \alpha_N\} \\ 0 & \text{αν } B \neq \{\alpha_1, \dots, \alpha_N\} \end{cases}$$

όπου k είναι ο αριθμός των διακριτών τιμών μεταξύ των α_i και $n_1, \dots, n_k$ είναι οι σχετικοί αριθμοί των φορών που συνέβησαν.

Αν όλα τα α_i είναι διακριτά τότε η έκφραση $(n_1! \dots n_k!)/(N!)$ γίνεται $1/(N!)$.

(Glenn Shafer, Vladimir Vovk, 2008)

Το πεδίο της συμμορφωτικής πρόβλεψης, κάτω από την υπόθεση της ανταλλαξιμότητας, διαχωρίζεται σε δυο περιπτώσεις απευθείας σύνδεσης (on-line) πρόβλεψης. Η πρώτη περίπτωση βασίζεται σε πρόβλεψη από παλιά παραδείγματα και μόνο, ενώ η δεύτερη περίπτωση αναφέρεται σε πρόβλεψη που γίνεται χρησιμοποιώντας τα χαρακτηριστικά του νέου αντικειμένου.

Παρακάτω δίνονται κάποιοι ορισμοί για τις δυο αυτές περιπτώσεις μαζί με τους αντίστοιχους συμμορφωτικούς αλγορίθμους.

Πρόβλεψη από παλιά παραδείγματα και μόνο (Prediction from old examples alone)

Ακριβώς πριν την παρατήρηση της περίπτωσης z_n, γίνεται η πρόβλεψη της βασισμένοι στα προηγούμενα παραδείγματα $z_1, \dots, z_{n-1}$.

Ο αντίστοιχος αλγόριθμος για την συγκεκριμένη περίπτωση λειτουργεί ως εξής:

Εισαγωγή: Μέτρο μη συμμόρφωσης A, επίπεδο σημαντικότητας ε, περιπτώσεις $z_1, \dots, z_{n-1}$, περίπτωση z.

Λειτουργία: Αποφάσισε αν θα πρέπει να συμπεριληφθεί η περίπτωση z στο διάστημα $\gamma^{\varepsilon}(z_1, \dots, z_{n-1})$.

Αλγόριθμος:

1. Προσωρινά θέτουμε $z_n := z$.
2. Για i = 1,…,n, θέτουμε $\alpha_i := A(\{z_1, \dots, z_n\} \setminus \{z_i\}, z_i)$.
3. Θέτουμε $p_z := \frac{\text{αριθμος των περιπτωσεων } i \text{ τετοιες ωστε } 1 \leq i \leq n \text{ και } \alpha_i \geq \alpha_n}{n}$.
4. Συμπεριλαμβάνουμε την περίπτωση z στο διάστημα $\gamma^{\varepsilon}(z_1, \dots, z_{n-1})$ αν και μονο αν το $p_z > \varepsilon$.

Πρόβλεψη χρησιμοποιώντας τα χαρακτηριστικά του νέου αντικειμένου (Prediction using features of the new object)

Κάθε περίπτωση z_i αποτελείται από ένα αντικείμενο x_i και μια περίπτωση (label) y_i. Ο συμβολισμός είναι $z_i = (x_i, y_i)$. Παρατηρούμε σε ακολουθία τα $x_1, y_1, \dots, x_N, y_N$. Ακριβώς πριν την παρατήρηση του y_n, γίνεται η πρόβλεψη του με βάση ότι έχουμε παρατηρήσει μέχρι την δεδομένη χρονική στιγμή, το x_n, αλλά και με βάση τις προηγούμενες περιπτώσεις $z_1, \dots, z_{n-1}$.

Ο αλγόριθμος για την παραπάνω περίπτωση πρόβλεψης έχει την ακόλουθη λειτουργία:

Εισαγωγή: Μέτρο μη συμμόρφωσης A, επίπεδο σημαντικότητας ε, περιπτώσεις $z_1, \dots, z_{n-1}$, αντικείμενο x_n, περίπτωση (label) y.

Λειτουργία: Αποφάσισε αν θα πρέπει να συμπεριληφθεί η περίπτωση y στο διάστημα $\Gamma^{\varepsilon}(z_1, \dots, z_{n-1}, x_n)$.

Αλγόριθμος:

1. Προσωρινά θέτουμε το $z_n := (x_n, y)$.
2. Για i = 1,…,n, θέτουμε $\alpha_i := A(\{z_1, \dots, z_n\} \setminus \{z_i\}, z_i)$.
3. Θέτουμε $p_y := \frac{\#\{i=1,\dots,n \mid \alpha_i \geq \alpha_n\}}{n}$.
4. Συμπεριλαμβάνουμε την περίπτωση y στο διάστημα $\Gamma^{\varepsilon}(z_1, \dots, z_{n-1}, x_n)$ αν και μονο αν το $p_y > \varepsilon$.

Οι προβλέψεις που γίνονται με κάποιον από τους συμμορφωτικούς αλγορίθμους έχουν τις εξής ιδιότητες:

1) Είναι αμετάβλητες σε σχέση με την διάταξη των παλαιοτέρων περιπτώσεων. Τυπικά, αυτό σημαίνει ότι ο εκτιμητής γ είναι μια συνάρτηση δυο μεταβλητών, και πιο συγκεκριμένα του επιπέδου σημαντικότητας ε και της τσάντας B των παλαιότερων περιπτώσεων. Γράφουμε $\gamma^{\varepsilon}(B)$ για την πρόβλεψη, η οποία είναι ένα υποσύνολο του υποθετικού διαστήματος Z.

2) Η πιθανότητα ενός διαδοχικού γεγονότος (hit) είναι τουλάχιστον ίση με το διαφημιζόμενο (advertised) επίπεδο εμπιστοσύνης. Δηλαδή, για κάθε θετικό ακέραιο n και για κάθε κατανομή πιθανότητας κάτω από την οποία οι $z_1, \dots, z_n$ είναι ανταλλάξιμες ισχύει ότι

$$P_r\{z_n \in \gamma^{\varepsilon}(\{z_1, \dots, z_{n-1}\})\} \geq 1 - \varepsilon.$$

3) Τα διαστήματα πρόβλεψης είναι εμφωλευμένα. Δηλαδή, αν $\varepsilon_1 \geq \varepsilon_2$, τότε $\gamma^{\varepsilon_1}(B) \subseteq \gamma^{\varepsilon_2}(B)$.

Οι συμμορφωτικοί εκτιμητές ικανοποιούν πάντα τις τρεις παραπάνω ιδιότητες.

(Glenn Shafer, Vladimir Vovk, 2008)

3.2. Δέντρα απόφασης

Ένα δέντρο απόφασης είναι ακόμα ένα εργαλείο που χρησιμοποιείται ευρέως στον τομέα της εξόρυξης γνώσης. Το δέντρο απόφασης έχει την μορφή δενδροδιαγράμματος και είναι πολύ χρήσιμο στην απεικόνιση κάποιων αλγορίθμων απόφασης.

Τα δέντρα απόφασης είναι η βασική τεχνολογία που χρησιμοποιείται στο κομμάτι της κατηγοριοποίησης αλλά και γενικότερα ως μέθοδος πρόβλεψης και σε άλλες περιοχές της εξόρυξης γνώσης.

Το πιο σύνηθες σύστημα δέντρου απόφασης είναι το ID3 που δημιουργήθηκε στο σύστημα γνώσης CLS και στη συνέχεια εξελίχθηκε στον αλγόριθμο C4.5 (C5.0) που ασχολείται κυρίως με περιπτώσεις που έχουν συνεχή χαρακτηριστικά.

Ο αλγόριθμος που χρησιμοποιεί δέντρα απόφασης είναι ένα από τα σημαντικότερα μέτρα κατηγοριοποίησης στο πεδίο της εξόρυξης γνώσης. Στα δέντρα απόφασης κάθε εσωτερικός κόμβος υποδηλώνει έναν έλεγχο που γίνεται σε κάποιο συγκεκριμένο χαρακτηριστικό, κάθε κλαδί αντιπροσωπεύει ένα αποτέλεσμα του ελέγχου και κάθε κόμβος φύλλων αντιπροσωπεύει μια κλάση (κατηγορία).

Ο αλγόριθμος κατηγοριοποίησης με δέντρα απόφασης χωρίζεται κυρίως σε δυο βήματα. Το πρώτο βήμα έχει να κάνει με την διαδικασία της κατασκευής (construction) του δέντρου, ενώ το δεύτερο αφορά περιπτώσεις που θα χρειαστεί να ΄΄κουρευτεί΄΄ (pruning) ένα δέντρο.

Παρακάτω ακολουθεί μια σύντομη περιγραφή των δυο αυτών λειτουργιών.

Κατασκευή ενός δέντρου απόφασης

Η διαδικασία της κατασκευής ενός δέντρου απόφασης βασίζεται στην εισαγωγή ενός σετ δεδομένων από διάφορες περιπτώσεις και έχει ως αποτέλεσμα ένα δυαδικό ή τριαδικό δέντρο.

Πιο συγκεκριμένα, στο πρωταρχικό μέρος της διαδικασίας γίνεται μια προσπάθεια να χωριστεί το αρχικό σύνολο δεδομένων σε μικρότερα υποσύνολα που πιθανώς να αποτελούν κάποιες ανεξάρτητες μεταβλητές. Στη συνέχεια, χρησιμοποιείται το κέρδος καθαρότητας (purity gain) με σκοπό την αξιολόγηση του συγκεκριμένου μοντέλου. Τέλος, επιλέγεται το μοντέλο που μεγιστοποιεί τη συνάρτηση του κέρδους καθαρότητας.

Παρακάτω δίνεται η συνάρτηση του κέρδους καθαρότητας.

$$\Delta = I(D) - \sum_{i=1}^{n} P(D_i) * I(D_i)$$

όπου

Δ: το κέρδος καθαρότητας

I(*): το μέτρο ακαθαρσίας του αντίστοιχου κόμβου

D: το αρχικό σύνολο δεδομένων

D_i: το i υποσύνολο

P(D_i): το ποσοστό του αρχικού συνόλου δεδομένων (D) που τοποθετείτε στο D_i

Όσον αφορά το μέτρο ακαθαρσίας I(*), κάθε αλγόριθμος δέντρων απόφασης χρησιμοποιεί και κάποιο διαφορετικό μέτρο. Δυο από τα πιο γνωστά μέτρα ακαθαρσίας είναι η εντροπία (entropy) και ο δείκτης Gini.

Ακολουθούν οι δυο μαθηματικές συναρτήσεις των παραπάνω μέτρων.

$$Entropy(t) = -\sum_{i=1}^{C} p(c_i|t) log_2 p(c_i|t)$$

$$Gini(t) = 1 - \sum_{i=1}^{C} [p(c_i|t)]^2$$

όπου

c : ο αριθμός των κλάσεων

p($c_i|t$) : το κλάσμα των περιπτώσεων που ανήκουν στην κλάση c_i του αντίστοιχου κόμβου t.

(Ulf Johansson, Henrik Boström, Tuve Löfström, 2013)

Το κλειδί στην κατασκευή ενός δέντρου απόφασης είναι ο τρόπος που θα γίνει η καλύτερη επιλογή κάποιου χαρακτηριστικού. Οι έρευνες έχουν δείξει ότι όσο πιο μικρό είναι ένα δέντρο απόφασης τόσο βελτιώνεται και το επίπεδο πρόβλεψης. Συνεπώς, για να έχουμε όσο το δυνατόν μικρότερο δέντρο, πρέπει να γίνει η καλύτερη επιλογή του χαρακτηριστικού που θα ελεγχθεί.

΄΄Κούρεμα΄΄ ενός δέντρου απόφασης

Όπως είναι γνωστό, το πεδίο της εξόρυξης γνώσης ασχολείται κυρίως με σύνολα δεδομένων που προέρχονται από τον πραγματικό κόσμο. Συνεπώς, γίνεται κατανοητό ότι τέτοιου είδους δεδομένα δεν είναι πάντα καταλληλά για να εφαρμοστούν σε ένα μοντέλο πρόβλεψης. Αυτό συμβαίνει γιατί μπορεί κάποιες περιπτώσεις να εμφανίζουν ελλειπούσες τιμές (missing values) ή να παρουσιάζουν κάποιου είδους ΄΄θόρυβο΄΄ (noise).

Τέτοια προβλήματα μπορούν να αντιμετωπιστούν από έναν αλγόριθμο των δέντρων απόφασης μέσω της διαδικασίας του κλαδέματος (pruning). Το κλάδεμα ή κούρεμα του δέντρου απόφασης αναφέρεται στην διαδικασία κατά την οποία απομακρύνουμε κάποια κομμάτια του δέντρου, κάτι το οποίο έχει ως αποτέλεσμα την μείωση του μεγέθους του δέντρου συνολικά.

Η τεχνική του κλαδέματος, πέρα από την αντιμετώπιση των παραπάνω προβλημάτων, βοηθάει ακόμα και στην καλύτερη κατανόηση του δέντρου.

Οι δυο βασικές στρατηγικές στην τεχνική του κλαδέματος ενός δέντρου είναι το Forward-Pruning και το Post-Pruning. Το Forward-Pruning είναι το κλάδεμα που γίνεται πριν την

ολοκλήρωση της διαδικασίας ανάπτυξης του δέντρου, ενώ το Post-Pruning είναι το κλάδεμα που συμβαίνει ενώ έχει ολοκληρωθεί η διαδικασία ανάπτυξης ενός δέντρου απόφασης.

(Qing-yun Dai, Chun-ping Zhang and Hao Wu, 2016)

Παρακάτω δίνεται ένα παράδειγμα προβλήματος με την χρήση ενός δέντρου απόφασης.

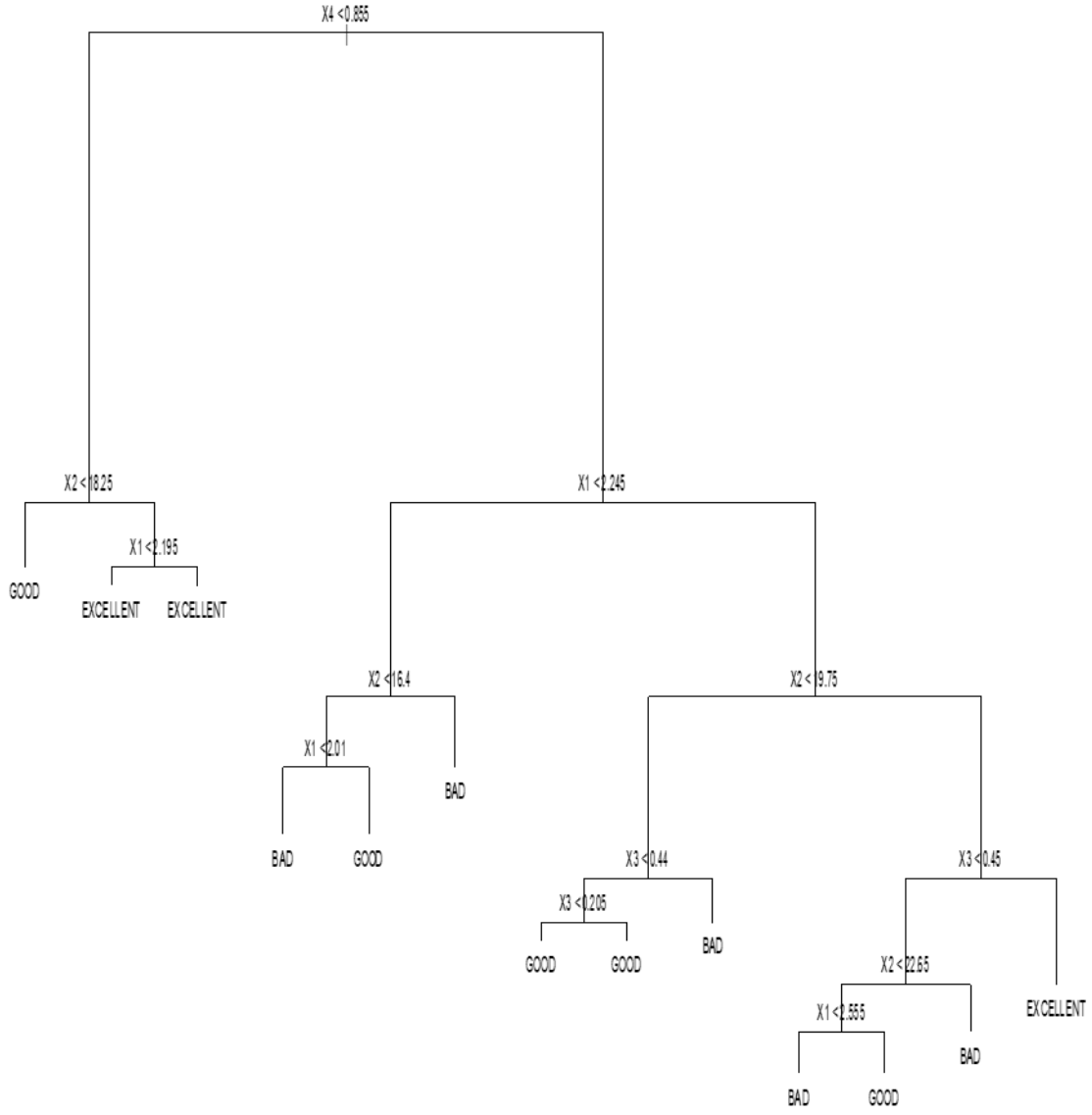
Το πρόβλημα αναφέρεται σε μια ομάδα της Euroleague στην οποία έχει προταθεί ένας παίκτης μεγάλης αξίας που αγωνιζόταν την προηγουμένη χρονιά στο NBA. Η ομάδα δεν μπορεί να αποφασίσει για το αν αξίζει να ρισκάρει τα χρήματα της για τον συγκεκριμένο αθλητή. Πριν πάρει την τελική της απόφαση, υποβάλει τον αθλητή σε κάποιες συγκεκριμένες εξετάσεις. Τα αποτελέσματα είναι τα εξής:

$$X1 = 2.26 \qquad X2 = 16.2 \qquad X3 = 0.31 \qquad X4 = 1.020$$

Οι παραπάνω ποσοτικές μεταβλητές (X1, X2, X3, X4) αναφέρονται σε κάποιες μετρήσεις, από τις οποίες οι δυο πρώτες προέκυψαν από ένα συγκεκριμένο τεστ αντοχής και οι άλλες δυο από αιματολογικές εξετάσεις.

Ακόμη, δίνεται και μια κατηγορική μεταβλητή τριών επιπέδων σύμφωνα με την οποία αξιολογείται η συνολική απόδοση των παικτών στο τέλος του πρωταθλήματος. Η απόδοση κάθε αθλητή κατατάσσεται σε ένα από τα τρία επίπεδα λαμβάνοντας υπόψη τα συνολικά ποσοστά ευστοχίας, τον χρόνο συμμέτοχης και τις αμυντικές του επιδόσεις. Η κατηγορική μεταβλητή ονομάζεται PERF και έχει τις ακόλουθες τιμές: Excellent, Good, Bad.

Με την βοήθεια του στατιστικού πακέτου Minitab 14 έχουμε το ακόλουθο δέντρο παλινδρόμησης.



Κοιτάζοντας τα αποτελέσματα του δέντρου απόφασης, βλέπουμε ότι ο συγκεκριμένος αθλητής κατατάσσεται στην κατηγορία κάλος (GOOD). Είναι λογικό ένας καλαθοσφαιριστής μεγάλης αξίας που προέρχεται από υψηλότερο επίπεδο, όπως είναι το NBA σε σύγκριση με την Euroleague, να κατατάσσεται τουλάχιστον στην κατηγορία κάλος (GOOD). Το γεγονός όμως ότι δεν κατατάσσεται στην κατηγορία εξαιρετικός (EXCELLENT), είναι λογικό να κάνει την ομάδα να ψάξει για κάποιον άλλο καλαθοσφαιριστή που είναι πιο οικονομικός και κατατάσσεται στην ιδιά κατηγορία (GOOD).

3.3. Συμμορφωτική πρόβλεψη με δέντρα απόφασης

Η συμμορφωτική πρόβλεψη είναι ένα σχετικά νέο πεδίο στον τομέα της εξόρυξης γνώσης. Αρχικά αναπτύχθηκε ως μια μέθοδος πρόβλεψης στον τομέα της ανάλυσης παλινδρόμησης και στη συνέχεια προσαρμόστηκε και σε μεγάλο μέρος προβλημάτων κατηγοριοποίησης.

Στην συμμορφωτική πρόβλεψη, τα προβλεπτικά μοντέλα παράγουν σετ προβλέψεων με έναν περιορισμό (φράγμα) στο ποσοστό σφάλματος. Εφόσον το ποσοστό σφάλματος είναι προκαθορισμένο από την δομή της μεθόδου, γίνεται εύκολα αντιληπτό ότι η πιο σημαντική ιδιότητα των συμμορφωτικών εκτιμητών είναι η αποτελεσματικότητα τους. Δηλαδή, μας ενδιαφέρει η ελαχιστοποίηση του αριθμού των στοιχείων στα σετ πρόβλεψης (μικρά διαστήματα πρόβλεψης).

Τα αποτελέσματα έρευνας έχουν δείξει ότι για να βελτιωθεί η αποτελεσματικότητα των συμμορφωτικών εκτιμητών είναι απαραίτητη η χρήση των δέντρων απόφασης χωρίς να χρησιμοποιηθεί η τεχνική του ΄΄κλαδέματος΄΄ στα δέντρα. Αντίθετα, είναι προτιμότερο να χρησιμοποιηθούν κάποιες εξομαλυμένες εκτιμήσεις πιθανότητας.

Όπως αναφέρθηκε και σε προηγούμενες ενότητες, η συμμορφωτική πρόβλεψη αντιμετωπίζει το πρόβλημα της εκτίμησης της αβεβαιότητας κάθε μεμονωμένης εκτίμησης με λίγο διαφορετικό τρόπο από τις συνηθισμένες μεθόδους οι οποίες βαθμονομούν τις προβλεπόμενες πιθανές κλάσεις. Δηλαδή, αντί να κάνει μια μεμονωμένη εκτίμηση μιας πιθανής κλάσης ή να κάνει μια εκτίμηση για την κατανομή των περιπτώσεων y, ένας συμμορφωτικός εκτιμητής παράγει ένα σετ περιπτώσεων y με ένα οριοθετημένο από πριν ποσοστό σφάλματος.

Στην συγκεκριμένη μέθοδο χρησιμοποιούνται διάφοροι γνωστοί αλγόριθμοι πρόβλεψης όπως είναι ο ANNs, ο kNN, ο SVMs, τα τυχαία δάση και φυσικά τα δέντρα απόφασης. Για να εφαρμοστεί ένας αλγόριθμος όπως αυτοί που προαναφέρθηκαν είναι απαραίτητο να γίνει και μια συγκεκριμένη προσαρμογή του.

Τα δέντρα απόφασης είναι μια πολύ διαδεδομένη τεχνική machine learning καθώς παράγει αρκετά ακριβείς προβλέψεις (υψηλή αποτελεσματικότητα) και ακόμα παράγει μοντέλα που είναι εύκολο να ερμηνευθούν. Επιπλέον, είναι μια σχετικά γρήγορη μέθοδος στην οποία απαιτείται ελάχιστη ρύθμιση των παραμέτρων. Οι δυο πιο γνωστοί αλγόριθμοι δέντρων απόφασης είναι ο C4.5/C5.0 και ο αλγόριθμος CART.

Όσον αφορά τους συμμορφωτικούς εκτιμητές, που μπορούν να χρησιμοποιηθούν και στην τεχνική των δέντρων απόφασης, υπάρχουν αρκετά είδη. Οι δυο πιο γνωστοί είναι μεταγωγικοί και οι επαγωγικοί. Οι πρώτοι απαιτούν για κάθε νέα περίπτωση την επαναδιαμόρφωση του μοντέλου που χρησιμοποιείται (Μεταγωγική συμμορφωτική πρόβλεψη), ενώ οι δεύτεροι διαμορφώνουν το μοντέλο μόνο μια φορά και δεν χρειάζεται η τροποποίηση του για κάθε νέα περίπτωση (Επαγωγική συμμορφωτική πρόβλεψη).

(Henrik Linusson, 2017)

Βασικό χαρακτηριστικό των πιο προσφάτων συμμορφωτικών εκτιμητών είναι ότι χρησιμοποιούν δυο ξεχωριστά μοντέλα πρόβλεψης. Το πρώτο μοντέλο, το οποίο είναι και το βασικό, παρέχει πληροφόρηση σχετικά με το κεντρικό σημείο κάθε διαστήματος πρόβλεψης.

Το δεύτερο μοντέλο πρόβλεψης ονομάζεται μοντέλο κανονικοποίησης και χρησιμοποιείται ως κλίμακα για κάθε διάστημα πρόβλεψης σύμφωνα με το εκτιμώμενο επίπεδο δυσκολίας για κάθε περίπτωση ελέγχου.

Το μοντέλο κανονικοποίησης έχει ιδιαίτερη σημασία στην αντιμετώπιση των προβλημάτων με την μέθοδο της συμμορφωτικής πρόβλεψης. Πιο συγκεκριμένα τα οφέλη που παρέχει το συγκεκριμένο μοντέλο είναι τα εξής:

1. Τα παραγόμενα διαστήματα πρόβλεψης διαφέρουν σε μέγεθος, έτσι ώστε οι περιπτώσεις που είναι ευκολότερο να προβλεφθούν θα λάβουν μικρότερα διαστήματα σε σχέση με τις περιπτώσεις όπου η πρόβλεψη τους είναι δυσκολότερη. Δηλαδή, παρέχεται επιπρόσθετη πληροφόρηση ανά περίπτωση.

2. Τα παραγόμενα διαστήματα πρόβλεψης είναι γενικά μικρότερα, και συνεπώς πιο αποτελεσματικά, κατά μέσο όρο.

(Ulf Johansson, Henrik Boström, Tuve Löfström, 2013)

ΚΕΦΑΛΑΙΟ 4

ΕΦΑΡΜΟΓΗ

Στο κεφάλαιο 4 της παρούσας εργασίας θα εφαρμόσουμε την μέθοδο της συμμορφωτικής πρόβλεψης με δέντρα κατηγοριοποίησης σε ένα σύνολο ακολουθιακών δεδομένων που προέρχονται από τον χώρο του ποδοσφαίρου.

4.1. Εφαρμογή

Ξεκινώντας είναι απαραίτητο να αναφέρουμε ότι ο αλγόριθμος που χρησιμοποιήθηκε για την πρόβλεψη των δεδομένων προέρχεται από την ιστοσελίδα GitHub (https://github.com/donlnz/nonconformist/blob/master/examples/icp_classification_tree.py)

Επιπλέον, τις τιμές για την δημιουργία του σετ δεδομένων τις συλλέξαμε από την ιστοσελίδα FlashScore.com (https://www.flashscore.com/).

Ο αλγόριθμος χρησιμοποιήθηκε στην συγκεκριμένη ιστοσελίδα (GitHub) για να προβλέψει την κατηγορία του λουλουδιού από την οποία προέρχονται κάποιες παρατηρήσεις. Το σύνολο δεδομένων που χρησιμοποιείται είναι το iris flower dataset του διάσημου Βρετανού στατιστικού Ρόναλντ Φίσερ. Από τα αντίστοιχα αποτελέσματα είδαμε ότι το μοντέλο πρόβλεψης είναι αρκετά αποτελεσματικό και καταφέρνει σε ποσοστό 90% να βρει την πραγματική κατηγορία από την οποία προέρχονται οι αντίστοιχες παρατηρήσεις.

Επειδή το συγκεκριμένο σετ δεδομένων δεν είναι ακολουθία, σκεφτήκαμε να χρησιμοποιήσουμε τον κώδικα με κάποιες μετατροπές έτσι ώστε να δούμε αν καταφέρνει το ίδιο καλά να προβλέψει το τελικό αποτέλεσμα των παιχνιδιών κάποιας ομάδας ποδοσφαίρου.

Η ομάδα που επιλέχτηκε αρχικά είναι η Άρσεναλ η οποία αγωνίζεται στην κορυφαία κατηγορία του Αγγλικού ποδοσφαίρου. Για να έχει η εφαρμογή μας και κάποιο ενδιαφέρον πρόβλεψης σκεφτήκαμε αντί να πάρουμε τα στατιστικά των αγώνων αφού έχει τελειώσει το παιχνίδι και γνωρίζουμε το τελικό αποτέλεσμα, να πάρουμε τα αντίστοιχα στατιστικά του 1^{ου} ημίχρονου και να δούμε αν το μοντέλο θα καταφέρει να προβλέψει σωστά το τελικό αποτέλεσμα.

Ας δούμε σε αυτό το σημείο πιο αναλυτικά το σετ δεδομένων που χρησιμοποιήσαμε. Αρχικά, πρέπει να πούμε ότι τα δεδομένα που συλλέξαμε αναφέρονται στην τρέχουσα αγωνιστική περίοδο (2018-2019) και προέρχονται αποκλειστικά από τους 30 πρώτους αγώνες της Άρσεναλ στο Αγγλικό πρωτάθλημα, χωρίς να συνυπολογίζονται τα παιχνίδια των Αγγλικών κυπέλλων ή των Ευρωπαϊκών κυπέλλων.

Οι τέσσερις στατιστικές κατηγορίες που χρησιμοποιήσαμε για να προσπαθήσουμε να προβλέψουμε το τελικό αποτέλεσμα είναι η κατοχή μπάλας, οι συνολικές πάσες και τα κόρνερ που είχε η ομάδα στο πρώτο ημίχρονο, μαζί με το αποτέλεσμα του πρώτου ημιχρόνου. Επειδή το αποτέλεσμα του ημιχρόνου καθώς επίσης και το τελικό αποτέλεσμα είναι κατηγορικές μεταβλητές τους εκχωρήσαμε τις ακόλουθες αριθμητικές τιμές:

1-> Νίκη

2-> Ήττα

0-> Ισοπαλία

Για να μην υπάρχει σύγχυση αυτές οι κατηγορίες παρέμεναν αμετάβλητες είτε πρόκειται για εντός έδρας παιχνίδι είτε για εκτός έδρας.

Το συγκεκριμένο σετ δεδομένων είναι μια ακολουθία αγώνων καθώς η ομάδα κάθε εβδομάδα παίζει κάποιο παιχνίδι και συλλέγονται τα αντίστοιχα στατιστικά στοιχεία. Πρέπει ακόμα να γίνει ξεκάθαρο ότι είναι ακολουθία και όχι χρονοσειρά, γιατί σε περίπτωση χρονοσειράς ο χρόνος συλλογής των στοιχείων είναι σταθερός (π.χ. ημερησίως, εβδομαδιαίως κ.ο.κ.). Ενώ αντίθετα ένας αγώνας πρωταθλήματος ποδοσφαίρου μπορεί να γίνεται κάθε εβδομάδα αλλά διαφορετική ημέρα. Για παράδειγμα κάποιοι αγώνες γίνονται Σάββατο ενώ κάποιοι άλλοι την Κυριακή.

Εν συνεχεία, θα δώσουμε μια σύντομη περιγραφή του κώδικα που χρησιμοποιήσαμε.

Αλγόριθμος

Βήμα 1 : Εισαγωγή των δεδομένων

Βήμα 2 : Χρησιμοποιείται μια εντολή μετάθεσης των δεδομένων

Βήμα 3 : Ταξινομούνται τα δεδομένα κατά αύξουσα σειρά

Βημα 4 : Χωρίζεται το αρχικό σύνολο δεδομένων σε τρία υποσύνολα (train set, calibrate set, test set)

Βήμα 5 : Χρησιμοποιούνται οι συναρτήσεις ClassifierNc, ClassifierAdapter, IcpClassifier, DecisionTreeClassifier και MarginErrFunc για την επεξεργασία των δεδομένων των train set και calibrate set

Βήμα 6 : Δημιουργία ενός πίνακα με 4 στήλες

Βήμα 7 : Εξαγωγή των δεδομένων

Σχόλιο: Το Βήμα 3 ήταν μια αναγκαία μετατροπή που κάναμε στον κώδικα έτσι ώστε να μπορέσουμε να τον χρησιμοποιήσουμε σε ακολουθία δεδομένων και να μην μπερδεύει τη σειρά των παρατηρήσεων.

Το πρόγραμμα που χρησιμοποιήθηκε για την εφαρμογή του μοντέλου είναι η γλώσσα προγραμματισμού Python. Επιπρόσθετα πρέπει να αναφέρουμε ότι το προκαθορισμένο ποσοστό σφάλματος που χρησιμοποιήσαμε είναι 10%. Παρακάτω ακολουθούν τα αποτελέσματα.

	0	1	2	3
0	c0	c1	c2	Truth
1	1	1	1	1
2	0	1	1	2
3	1	0	1	1
4	1	1	1	1
5	1	0	1	2
6	1	1	1	1
7	1	1	0	1
8	1	1	1	1
9	1	1	1	0
10	1	1	0	1

Ξεκινώντας την περιγραφή των αποτελεσμάτων πρέπει να αναφέρουμε τι σημαίνουν τα αποτελέσματα του πίνακα. Οι 3 πρώτες στήλες (c0,c1,c2) αναφέρονται στην κατηγορία που προβλέπει ο κώδικας για το τελικό αποτέλεσμα, ενώ η τέταρτη στήλη (Truth) αναφέρεται στην πραγματική κατηγορία του τελικού αποτελέσματος.

Κοιτάζοντας τα αποτελέσματα παρατηρούμε ότι ο κώδικας δυσκολεύεται σε μεγάλο βαθμό να προβλέψει την πραγματική κατηγορία από την οποία προέρχεται κάθε παρατήρηση. Το γεγονός αυτό μπορεί να οφείλεται στο ότι η Άρσεναλ στις πρώτες 20 αγωνιστικές ήταν ισόπαλη στο ημίχρονο και τις περισσότερες φορές κατάφερνε να κερδίσει στο τέλος. Ενώ αντίθετα στις τελευταίες 10 αγωνιστικές κατάφερνε στις περισσότερες των περιπτώσεων να κερδίζει από το ημίχρονο.

Για να εξετάσουμε λίγο περισσότερο το συγκεκριμένο κομμάτι αποφασίσαμε να χρησιμοποιήσουμε τον ίδιο κώδικα και για τους 27 πρώτους αγώνες πρωταθλήματος της Μπαρτσελόνα (2018-2019) η οποία αγωνίζεται στο ισπανικό πρωτάθλημα. Παρατηρήσαμε ότι η Μπαρτσελόνα κατάφερνε τις περισσότερες φορές να κερδίζει το παιχνίδι από το πρώτο ημίχρονο. Η μόνη αλλαγή που έγινε είναι ότι πήραμε σαν στατιστικό τις συνολικές ευκαιρίες

που είχε στο πρώτο ημίχρονο και αφαιρέσαμε από το σύνολο δεδομένων την κατηγορία με τις τιμές των κόρνερ.

Τα αποτελέσματα δίνονται παρακάτω.

	0	1	2	3
0	c0	c1	c2	Truth
1	1	1	1	1
2	0	1	1	1
3	1	1	1	1
4	1	1	1	0
5	1	1	1	0
6	1	1	0	1
7	1	1	1	1
8	0	1	1	1
9	1	1	1	1

Κοιτάζοντας τα παραπάνω αποτελέσματα βλέπουμε ότι και πάλι ο κώδικας δυσκολεύεται να εντοπίσει την πραγματική κατηγορία από την οποία προέρχονται οι παρατηρήσεις. Ενώ δηλαδή τα στατιστικά στοιχεία της Μπαρτσελόνα είναι περισσότερο ξεκάθαρα και η συγκεκριμένη ομάδα καταφέρνει τις περισσότερες φορές να κερδίζει από το ημίχρονο, και πάλι ο κώδικας δεν μπορεί να προβλέψει το τελικό αποτέλεσμα.

Για να γίνουμε περισσότερο συγκεκριμένοι, ο κώδικας εμφανίζει διαστήματα πρόβλεψης που είναι εξαιρετικά πλατιά. Δηλαδή πολλές φορές προβλέπει ότι κάθε τελικό αποτέλεσμα είναι το ίδιο πιθανό να συμβεί. Για αυτό τον λόγο δίνει την τιμή 1 (τιμή πρόβλεψης κατηγορίας τελικού αποτελέσματος) σε πάνω από μια κατηγορία.

Για παράδειγμα στην πρώτη γραμμή των αποτελεσμάτων για τα παιχνίδια της Μπαρτσελόνα ο κώδικας προβλέπει ότι μπορεί να έρθει νίκη, ισοπαλία αλλά και ήττα σύμφωνα με τα στατιστικά που έχει. Δηλαδή τα διαστήματα πρόβλεψης είναι έγκυρα αλλά δεν δίνουν μεγάλη πληροφόρηση για το τελικό αποτέλεσμα (εξαιρετικά πλατιά).

ΚΕΦΑΛΑΙΟ 5

ΣΥΜΠΕΡΑΣΜΑΤΑ

Σε αυτή την ενότητα παρουσιάζονται κάποια συμπεράσματα που εξήχθησαν μετά την εφαρμογή.

5.1. Συμπεράσματα

Ξεκινώντας, είναι πολύ σημαντικό να αναφέρουμε ότι η κύρια σκέψη, από την αρχή της συγγραφής της παρούσας εργασίας, ήταν να χρησιμοποιήσουμε την συμμορφωτική πρόβλεψη με δέντρα απόφασης σε ένα σύνολο ακολουθιακών δεδομένων που προέρχονται από τον χώρο του ποδοσφαίρου. Αυτό που μας φάνηκε περισσότερο ενδιαφέρον ήταν, με βάση κάποια στατιστικά του πρώτου ημιχρόνου, να προσπαθήσουμε να προβλέψουμε το τελικό αποτέλεσμα.

Αρχικά σκεφτήκαμε να εφαρμόσουμε το μοντέλο της συμμορφωτικής πρόβλεψης στην ακολουθία αγώνων πρωταθλήματος της Άρσεναλ για την σεζόν 2018-2019. Όπως είδαμε και παραπάνω ο κώδικας δυσκολευόταν αρκετά να προβλέψει την πραγματική κατηγορία. Έτσι, στην συνέχεια σκεφτήκαμε να δοκιμάσουμε να χρησιμοποιήσουμε το ίδιο μοντέλο για την ακολουθία αγώνων πρωταθλήματος της Μπαρτσελόνα την ίδια αγωνιστική περίοδο. Η επιλογή της Μπαρτσελόνα δεν έγινε τυχαία. Παρατηρήσαμε ότι η συγκεκριμένη ομάδα είχε πολύ υψηλό ποσοστό νικών καθώς επίσης και το γεγονός ότι η Μπαρτσελόνα κέρδιζε τις περισσότερες φορές από το πρώτο ημίχρονο. Και πάλι όμως ο αλγόριθμος δεν κατάφερνε ξεκάθαρα να εντοπίσει την πραγματική κατηγορία του τελικού αποτελέσματος.

Συμπεραίνουμε λοιπόν ότι είναι πολύ πιθανό αυτές οι ομάδες να έχουν παρόμοιο τρόπο παιχνιδιού είτε κερδίζουν είτε χάνουν. Πιο συγκεκριμένα παρατηρήσαμε ότι αυτές οι ομάδες έχουν υψηλά ποσοστά κατοχής μπάλας και συνολικών μεταβιβάσεων της μπάλας, όχι μόνο στους αγώνες που κέρδιζαν αλλά και στους αγώνες που έχαναν. Είναι λογικό λοιπόν ο αλγόριθμος να δυσκολεύεται να καταλάβει σε ποια κατηγορία ανήκουν τα αντίστοιχα στατιστικά αποτελέσματα.

Κλείνοντας πρέπει να αναφέρουμε ότι παρόλο που ο αλγόριθμος δεν κατάφερε να προβλέψει το τελικό αποτέλεσμα των αγώνων δεν σημαίνει ότι δεν μπορεί να το πέτυχει σε κάποιο άλλο σύνολο ακολουθιακών δεδομένων. Όπως αναφέραμε και στην αρχή της εφαρμογής, ο συγκεκριμένος αλγόριθμος πετυχαίνει σε ποσοστό 90% να προβλέψει την σωστή κατηγορία στο iris flower dataset.

ΠΑΡΑΡΤΗΜΑ

(Ο ΑΛΓΟΡΙΘΜΟΣ ΤΗΣ ΕΦΑΡΜΟΓΗΣ)

```
data = load_footbal()
a = np.random.permutation(data.target.size)
idx = np.sort(a)
train = idx[:int(idx.size / 3)]
calibrate = idx[int(idx.size / 3):int(2 * idx.size / 3)]
test = idx[int(2 * idx.size / 3):]
icp = IcpClassifier(ClassifierNc(ClassifierAdapter(DecisionTreeClassifier()),
                                 MarginErrFunc()))
icp.fit(data.data[train, :], data.target[train])
icp.calibrate(data.data[calibrate, :], data.target[calibrate])
prediction = icp.predict(data.data[test, :], significance=0.1)
header = np.array(['c0','c1','c2','Truth'])
table = np.vstack([prediction.T, data.target[test]]).T
df = pd.DataFrame(np.vstack([header, table]))
print(df)
```

Βιβλιογραφία

Κυριαζής, Α. (2004). Στοιχεία Απειροστικού Λογισμού, Εκδοτικός οίκος INTERBOOKS (2004)

Κούτρας, Μ. , Ευαγγελάρας, Χ. (2010). Ανάλυση Παλινδρόμησης: Θεωρία και Εφαρμογές, Εκδόσεις Σταμούλη Α.Ε. (2010)

Θαλασσινός, Λ. (1991). Ανάλυση Χρονολογικών Σειρών, Εκδόσεις Α. Σταμούλης (Πειραιάς 1991)

Αγιακλόγλου, Χ. , & Οικονόμου, Γ. (2002). Μέθοδοι προβλέψεων και ανάλυσης αποφάσεων, Εκδόσεις Ε. Μπένου (Αθήνα 2002).

Βικιπαιδεια

Fournier-Viger, P., Chun-Wei Lin, J., Kiran, R.U., Koh, Y.S., & Thomas, R. (2017). A Survey of Sequential Pattern Mining, Volume 1, Number 1, February 2017.

Bijuraj, L.V. (2013). Clustering and its Applications, NCNHIT 2013.

Modi, K. & Oza, B. (2016). Outlier Analysis Approaches in Data Mining, Volume 3 Issue 7, December 2016.

Sharma, K., Ashok., & Rohil, H. (2014) A Study of Sequential Pattern Mining Techniques, Volume-4, Issue-1, February 2014, Page number: 241-248.

Shumway, R.H., & Stofer, D.H. (2006). Time Series Analysis and Its Applications with R Examples, Second Edition, New York 2006.

Atluri, G., Karpatne, A., & Kumar, V. (2017). Spatio -Temporal Data Mining : A Survey of Problems and Methods, November 2017, 37 pages.

Shafer, G., & Vovk, V. (2008). A Tutorial on Conformal Prediction, Journal of Machine Learning Research 9 (2008) 371-421.

Dai, Q.Y., Zhang, C.P., & Wu, H. (2006). Research of Decision Tree Classification Algorithm in Data Mining.

Johansson, U., Bostrom, H., & Lofstrom, T. (2013). Conformal Prediction Using Decision Trees, 13th International Conference on Data Mining 2013 IEEE.

Linusson, H. (2017). An introduction to conformal prediction, June 13,2017.

https://github.com/donlnz/nonconformist/blob/master/examples/icp_classification_tree.py

(https://www.flashscore.com/