

ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΤΜΗΜΑ ΔΙΟΙΚΗΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ & ΤΕΧΝΟΛΟΓΙΑΣ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

Με τίτλο

«ΘΕΩΡΗΤΙΚΗ ΤΕΚΜΗΡΙΩΣΗ ΑΛΓΟΡΙΘΜΩΝ ΜΗΧΑΝΙΚΗΣ ΜΑΘΗΣΗΣ ΓΙΑ ΤΗ ΧΡΗΣΗ ΤΟΥΣ ΣΤΗΝ ΑΥΤΌΜΑΤΗ ΑΝΑΓΝΩΡΙΣΗ ΚΑΙ ΚΑΤΑΤΑΞΗ ΠΑΡΑΣΤΑΤΙΚΩΝ»

του

ΜΙΧΑΗΛ ΝΕΣΤΕΡΟΥΛΗ (Α.Μ. MSTPT1612)

Επιβλέπουσα: Αγγελική Πουλυμενάκου, Αναπληρώτρια Καθηγήτρια στην Διοικητική των Πληροφοριακών Συστημάτων

Υποβληθείσα ως μέρος των απαιτήσεων για την απόκτηση

Μεταπτυχιακού Διπλώματος (MSc) στη Διοικητική Επιστήμη και Τεχνολογία

Αθήνα, Νοέμβριος 2018




ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα διπλωματική εργασία εκπονήθηκε στα πλαίσια του Μεταπτυχιακού Προγράμματος Σπουδών «Διοικητικής Επιστήμης και Τεχνολογίας» του ομώνυμου τμήματος, υπό την επίβλεψη της Καθηγήτριας κας Αγγελικής Πουλυμενάκου. Θα ήθελα να ευχαριστήσω θερμά την κυρία Πουλυμενάκου που μου έδωσε την ευκαιρία να ασχοληθώ με ένα τόσο ενδιαφέρον αντικείμενο, που ανταποκρίνεται σε μεγάλο βαθμό στα επιστημονικά μου ενδιαφέροντα, καθώς και για τη βοήθειά της κατά τη διάρκεια της έρευνας. Επίσης, ευχαριστώ τον κύριο Δημήτριο Παπακυριακόπουλο, καθηγητή στο Τεχνολογικό Εκπαιδευτικό Ίδρυμα Αθήνας στο τμήμα Διοίκησης Επιχειρήσεων και την κυρία Αγγελική Κωστάκη, υποψήφια διδάκτωρ στο τμήμα Διοικητικής Επιστήμης και Τεχνολογίας για την συνεχή καθοδήγηση και τις συμβουλές τους σε αυτό το εγχείρημα. Τέλος, θα ήθελα να ευχαριστήσω τον κύριο Γεώργιο Χαλκιαδάκη, Τεχνικό Διευθυντή στην εταιρία Information Systems Impact για την αμέριστη συμπαράσταση, βοήθεια και συμβολή του για την ολοκλήρωση της παρούσας εργασίας.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ΠΕΡΙΛΗΨΗ

Σε αυτή την εργασία παρουσιάζεται η προσπάθεια αυτόματης αναγνώρισης και κατηγοριοποίησης παραστατικών, με βάση τις πληροφορίες που υπάρχουν διαθέσιμες από την πολυετή δράση της εταιρίας Information Systems Impact στην ηλεκτρονική ανταλλαγή εγγράφων και δομημένης πληροφορίας (EDI), με το συνολικό αριθμό των παραστατικών να ξεπερνούν τα 40 εκατομμύρια. Οι ήδη υπάρχοντες μηχανισμοί αναγνώρισης εγγράφων που χρησιμοποιήθηκαν, αποτελούν Human – Driven διαδικασίες κατά τις οποίες τα στοιχεία ενός παραστατικού αναγνωρίζονται με τη χρήση του χώρου.

Η κάθε μία από τις παραπάνω Human – Driven διαδικασίες περιλαμβάνει τις μορφές αξιών και ημερομηνιών, το διαχωρισμό του εκάστοτε παραστατικού σε 3 περιοχές (Κεφαλίδα, Γραμμές και Συνολικές αξίες) και το σύνολο πεδίων που την αποτελούν που υπάρχουν σε αυτές τις περιοχές.

Στο παραπάνω σύνολο των πληροφοριών, δημιουργήθηκε ένας μηχανισμός για τη συλλογή των δεδομένων, χρησιμοποιήθηκε Computer Vision για το διαχωρισμό των περιοχών των παραστατικών και text classification για την κατηγοριοποίηση των λέξεων που υπάρχουν σε κάθε περιοχή, πάνω στις οποίες εφαρμόστηκαν τεχνικές επεξεργασίας κειμένων. Τέλος, χρησιμοποιήθηκαν και δένδρα απόφασης για την κατηγοριοποίηση του παραστατικού. Το σύνολο των χαρακτηριστικών που εξάχθηκαν, χωρίστηκε σε ένα σύνολο εκπαίδευσης και ένα σύνολο ελέγχου. Το σύνολο εκπαίδευσης χρησιμοποιήθηκε για την εκπαίδευση του δένδρου απόφασης και το σύνολο ελέγχου για την τελική αναγνώριση και κατηγοριοποίηση του παραστατικού.

Τα αποτελέσματα θα ελεγχθούν με βάση την εμπειρία των χρηστών πάνω στις παραπάνω διαδικασίες.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ABSTRACT

This paper presents the attempt to automatically identify and categorize documents based on information available from the multi-year action of Information Systems Impact in electronic document exchange and structured information (EDI), with a total number of documents exceeding 40 million. The existing document recognition mechanisms that were used, are Human - Driven processes in which the elements of a document are identified using coordinates.

Each of the above-mentioned Human-Driven processes includes the formats of values and dates, the separation of each document into 3 areas (Header, Lines and Footer) and the set of fields that are present in these areas.

In the above set of information, a mechanism was created for data collection, Computer Vision was used to separate the document areas and text classification to categorize the words in each area on which text editing techniques were applied. Finally, decision trees were used to categorize the documents. The set of exported features was divided into training set and a test set. The training set was used to train the decision tree and the test set for the final identification and categorization of the document.

The results were tested based on user experience on the above procedures.

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Περιεχόμενα

Κεφάλαιο 1: Στόχος εργασίας - Ορισμός προβλήματος 7

1.1 Στόχος εργασίας 7

1.2 Ορισμός προβλήματος 7

Κεφάλαιο 2: Βασικές έννοιες και μέθοδος 9

2.1 Πλαίσιο - Θέματα 9

2.2 Βασικές Τεχνικές 10

2.2.1 Μηχανική Μάθηση 10

2.2.2 Feature selection 11

2.2.3 Mahalanobis Distance 13

2.2.4 Επίδοση αλγορίθμων Μηχανικής Μάθησης 14

2.2.5 Computer Vision 15

Κεφάλαιο 3: Δεδομένα 16

3.1 Τύποι, κατηγορίες και process επεξεργασίας 16

3.2 Σύνοψη και περιγραφή μηχανισμών αναγνώρισης 16

3.3 Επιλογή μεταβλητών 19

3.4 Association Rules 20

Κεφάλαιο 4: Αλγόριθμοι Classification 23

4.1 Τύποι αλγορίθμων και χρησιμοποίηση του καθενός 23

4.2 Naïve Bayes 23

4.3 Decision Trees 24

4.4 Support Vector Machines 27

4.5 Εφαρμογή Mahalanobis Distance ως κριτήριο κατηγοριοποίησης 27

4.6 Text Classification 30

Κεφάλαιο 5: Project Methodology 33

5.1 Μεθοδολογία Knowledge Discovery in Databases 33

5.2 Μοντέλο CRISP-DM (CRoss-Industry Standard Process for Data Mining) 34

Κεφάλαιο 6: Προτεινόμενη Λύση 38

6.1 Προαπαιτούμενα - Παραδοχές 38

6.2 Overview of the technical solution - Tech Stack 39

6.3 Συνολική λειτουργία αλγορίθμου 40

Κεφάλαιο 7: Συμπεράσματα και Μελλοντική Δράση 43

Βιβλιογραφία 44

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

Κεφάλαιο 1: Στόχος εργασίας - Ορισμός προβλήματος

1.1 Στόχος εργασίας

Η παρούσα διπλωματική εργασία έχει ως στόχο την έρευνα και αναζήτηση μεθόδων για την αξιοποίηση της μηχανικής μάθησης στην αυτόματη κατηγοριοποίηση και αναγνώριση παραστατικών.

Το παραπάνω προκύπτει από τον συνεχώς αυξανόμενο όγκο τιμολογίων που πρέπει να διαχειριστεί ένας οργανισμός, καθώς η διαχείριση και καταχώρισή τους αποτελεί μία χρονοβόρα και κοστοβόρα διαδικασία. Όλες οι εταιρείες σε όλο τον κόσμο πρέπει να επεξεργάζονται τιμολόγια. Τα τιμολόγια είναι αυτά που κάνουν τον επιχειρηματικό κόσμο να περιστρέφεται και κάθε χρόνο, οι εταιρείες εκδίδουν πάνω από 18 δισεκατομμύρια τιμολόγια στην Ευρώπη και την Αμερική. Η Ευρωπαϊκή Επιτροπή εκτιμά ότι το 80% έως 90% των τιμολογίων (2017) είναι χαρτί. Οι υπεύθυνοι για την επεξεργασία των τιμολογίων είναι τα λογιστικά τμήματα και οι πάροχοι λογιστικών υπηρεσιών και η επεξεργασία τιμολογίων γίνεται χειροκίνητα ή ημιαυτόματα.

Ο μηχανισμός που θα προταθεί, βασίζεται σε ένα υπάρχον σύνολο human – driven διαδικασιών κατά τις οποίες τα στοιχεία ενός παραστατικού αναγνωρίζονται με τη χρήση χώρου συντεταγμένων και αφορούν μεγάλο πλήθος προμηθευτών για διάφορους παραλήπτες. Η παρούσα εργασία προτείνει την υιοθέτηση τεχνικών computer vision και classification για τον εντοπισμό και εξαγωγή των πεδίων ενδιαφέροντος (πχ ΑΦΜ, Κωδικοί προϊόντων κλπ) από ένα παραστατικό.

1.2 Ορισμός προβλήματος

Για τη δημιουργία του μοντέλου που περιγράφεται, θα δημιουργηθεί αρχικά ένας μηχανισμός συλλογής των υπαρχόντων δεδομένων των human – driven διαδικασιών, τα οποία αναφέρονται αναλυτικότερα στο Κεφάλαιο 5.

Στη συνέχεια, με βάση τα δεδομένα που έχουν συλλεγεί, αλλά και της εμπειρίας των χρηστών, θα δημιουργηθούν βιβλιοθήκες λεκτικών και φράσεων που περιέχονται σε κάθε τμήμα των παραστατικών ανά γλώσσα. Για παράδειγμα, για την ελληνική γλώσσα, γνωρίζουμε πως στο τμήμα της κεφαλίδας του παραστατικού υπάρχει σχεδόν πάντα το

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

λεκτικό «Τύπος παραστατικού». Η γλωσσική προ-επεξεργασία αποτελεί ένα από τα πρώτα στάδια ενός συστήματος επεξεργασίας φυσική γλώσσας, και συνεπώς εξαγωγής πληροφορίας. Το παραπάνω οδηγεί στην ανάλυση των λέξεων ή φράσεων πάνω στο παραστατικό, με σκοπό το text classification που διαχωρίζει αυτόματα τις ενότητες που περιέχονται σε αυτό.

Το επόμενο στάδιο είναι η εύρεση των συσχετίσεων των λεκτικών μεταξύ τους, προκειμένου να ανακαλύψουμε αν μία λέξη βρίσκεται κοντά σε μία άλλη μέσα στο χώρο. Αυτή η συσχέτιση έχει ως αποτέλεσμα να γνωρίζουμε αν η πληροφορία που πρόκειται να εξαχθεί είναι χρήσιμη ή όχι για την εκπαίδευση του μοντέλου.

Τέλος, ύστερα από τις διαδικασίες που περιγράφονται, ο μηχανισμός που δημιουργείται ύστερα από την εκπαίδευση του μοντέλου, είναι ικανός να κατατάξει το νέο παραστατικό που θα εισαχθεί στη διαδικασία στην κατάλληλη κατηγορία και να αναγνωρίσει τα περιεχόμενα του.

Κεφάλαιο 2: Βασικές έννοιες και μέθοδος

2.1 Πλαίσιο - Θέματα

Στη διπλωματική εργασία, θα ακολουθηθεί η διαδικασία CRISP-DM που αποτελεί ένα μοντέλο – μεθοδολογία εξόρυξης δεδομένων, θα αναλυθούν τεχνικές επιλογής μεταβλητών με βάση τη σημαντικότητά τους, θα ερευνηθεί η δυνατότητα αξιοποίησης μεθόδων μηχανικής μάθησης για επεξεργασία φυσικής γλώσσας με σκοπό την δημιουργία αξιοποιήσιμων συστημάτων, άμεσα εφαρμόσιμων για την επεξεργασία παραστατικών αλλά και μεθόδων classification για αυτόματη κατηγοριοποίηση και αναγνώριση παραστατικών.

Το παραπάνω έχει εξαιρετικό ενδιαφέρον, καθώς βοηθά στην κατανόηση των διαδικασιών, αλλά και αναλύει την κατευθυντήρια γραμμή της διπλωματικής εγρασίας. Όπως αναφέρει ο Colin Shearer [2], η CRISP-DM είναι μια ολοκληρωμένη μεθοδολογία εξόρυξης δεδομένων και ένα μοντέλο διαδικασίας που παρέχει σε οποιονδήποτε (είτε αρχάριους είτε όχι) ένα πλήρες σχέδιο για τη διεξαγωγή ενός έργου.

Προκειμένου να ξεκινήσει το παραπάνω εγχείρημα και προτού αναλυθούν οι αλγόριθμοι και οι τεχνικές που ακολουθούνται, χρειάζεται η εξαγωγή, περιγραφή και έλεγχος της ποιότητας των δεδομένων που υπάρχουν διαθέσιμα. Το αντικείμενο που θα εξεταστεί, είναι ένας μεγάλος αριθμός παραστατικών (περίπου ένα εκ. δείγματα) και οι πληροφορίες που έχουν εξαχθεί από αυτά ύστερα από την επεξεργασία τους από το σύστημα. Η προαναφερθείσα εξαγωγή, έχει πραγματοποιηθεί με την χειροκίνητη δημιουργία μηχανισμών αναγνώρισης για την κάθε μορφή παραστατικού, μέσω της οποίας είναι γνωστές οι μορφές ημερομηνιών και ποσών για την κάθε περίπτωση, η στοίχιση της σελίδας και τα πεδία που περιέχονται (ΑΦΜ, μοναδικά στοιχεία παραστατικού, σύνολα κλπ). Επίσης, λόγω των πολλαπλών Layers σε κάθε pdf, αλλά και των διαφορετικών μεγεθών των γραμματοσειρών που ενδέχεται να περιέχονται σε αυτά, έχουν αναπτυχθεί τεχνικές computer vision για τη μετατροπή των παραπάνω σε απλό κείμενο.

Η διαδικασία που θα αναλυθεί στα επόμενα κεφάλαια έχει ως εξής:

- Περιγραφή μεθοδολογίας Knowledge Discovery in Databases (KDD) και CRISP-DM (CRoss-Industry Standard Process for Data Mining) για την εξαγωγή και επιλογή των μεταβλητών.
- Κριτήρια επιλογής των κατάλληλων μεταβλητών

- Περιγραφή και επιλογή των κατάλληλων machine learning αλγορίθμων για την κατηγοριοποίηση των δεδομένων.
- Περιγραφή λειτουργίας του αλγορίθμου στο σύνολό του.
- Συμπεράσματα και μελλοντικές δράσεις.

Το Project που θα περιγράφει στην διπλωματική εργασία ανήκει και λαμβάνει χώρα στην εταιρία Information Systems Impact. Ο προσωπικός μου ρόλος σε αυτό το εγχείρημα, είναι η έρευνα των βέλτιστων αλγορίθμων και τεχνικών που θα εξεταστούν και Exploratory Data Analysis πάνω στα διαθέσιμα δεδομένα.

Διάφοροι φορείς που αντιμετωπίζουν το ίδιο ή παρόμοιο ζήτημα είναι οι παρακάτω:

- datamolino (https://www.datamolino.com)
- cloudtrade (https://www.cloudtradenetwork.com)
- ABBYY (https://www.abbyy.com/en-gb)

Οι διαδικασίες που ακολουθούνται από τις παραπάνω εταιρίες δεν είναι γνωστές, πέραν της ABBYY στη σελίδα της οποίας αναφέρεται πως χρησιμοποιούν αλγορίθμους μηχανικής μάθησης.

2.2 Βασικές Τεχνικές

Παρακάτω, παρουσιάζονται οι βασικές έννοιες που απαιτούνται για την κατανόηση των μεθόδων και τεχνικών που αναπτύχθηκαν στα πλαίσια της εργασίας.

2.2.1 Μηχανική Μάθηση

Η μηχανική μάθηση (machine learning) αποτελεί έναν από τους σημαντικότερους τομείς έρευνας της τεχνητής νοημοσύνης. Στόχος της, είναι η δημιουργία συστημάτων που να μπορούν να εκπαιδεύονται από δεδομένα που έχουν παρατηρηθεί στο παρελθόν, ώστε να εκτελούν την εργασία για την οποία προορίζονται αποτελεσματικότερα. Η διαδικασία εκμάθησης μπορεί να αναλυθεί στα παρακάτω στάδια:

- Απόκτηση δεδομένων από την αλληλεπίδραση με το περιβάλλον.

- Επεξεργασία των δεδομένων, ούτως ώστε να βρεθούν πιθανές γενικεύσεις ή εξειδικεύσεις.
- Χρήση των αποτελεσμάτων της επεξεργασίας για την εκτέλεση της εργασίας στόχου.

Η μηχανική μάθηση μπορεί να διαχωριστεί στην επιβλεπόμενη μάθηση (supervised learning) και στη μη επιβλεπόμενη μάθηση (unsupervised learning). Ένα σύστημα επιβλεπόμενης μάθησης, εκπαιδεύεται αρχικά σε ένα σύνολο παραδειγμάτων, όπου κάθε παράδειγμα χαρακτηρίζεται από μια κατηγορία. Ένα παράδειγμα επιβλεπόμενης μάθησης αποτελούν τα προβλήματα ταξινόμησης (classification). Σε ένα πρόβλημα ταξινόμησης, κάθε παράδειγμα αντιστοιχεί σε ένα διάνυσμα που αποτελείται από ένα σύνολο τιμών χαρακτηριστικών, ή αλλιώς γνωρισμάτων, το οποίο περιέχει και μια τιμή κατηγορίας η οποία περιγράφει το επιθυμητό αποτέλεσμα, ή αλλιώς, την έννοια στόχο. Στη μη επιβλεπόμενη μάθηση δεν υπάρχει προκαθορισμένο σύνολο τιμών. Τα παραδείγματα εκπαίδευσης χωρίζονται σε ομάδες με βάση τα χαρακτηριστικά τους, μια διαδικασία που αναφέρεται σαν κατηγοριοποίηση (clustering).

Στην παρούσα διπλωματική εργασία, θα αναλυθεί μία επιβλεπόμενη προσέγγιση του εγχειρήματος, καθώς οι μορφές παραστατικών που χρησιμοποιούνται τείνουν να διαχωρίζονται σε συγκεκριμένες περιπτώσεις με ελαφρές διαφοροποιήσεις.

2.2.2 Feature selection

Στη μηχανική μάθηση, η επιλογή χαρακτηριστικών, γνωστή ως επιλογή μεταβλητών ή επιλογή μεταβλητού υποσυνόλου, είναι η διαδικασία επιλογής μεταβλητών για χρήση στην κατασκευή μοντέλων. Οι τεχνικές επιλογής χαρακτηριστικών χρησιμοποιούνται για τέσσερις λόγους:

- απλοποίηση των μοντέλων ώστε να διευκολυνθεί η ερμηνεία τους από τους ερευνητές / χρήστες.
- συντομότερους χρόνους εκπαίδευσης.
- Για τη μείωση των διαστάσεων – μετασχηματισμός των διαστάσεων.
- βελτιωμένη γενίκευση με τη μείωση της της διακύμανσης.

Η επιλογή των μεταβλητών, μπορεί να γίνει με τους παρακάτω τρόπους:

- Variable Ranking: είναι η διαδικασία διάταξης των χαρακτηριστικών από την αξία κάποιας λειτουργίας βαθμολόγησης, η οποία συνήθως μετρά τη συνάφεια των

χαρακτηριστικών. Η βαθμολογία S(fi) υπολογίζεται από τα δεδομένα του training, μετρώντας ορισμένα κριτήρια του χαρακτηριστικού fi. Κατά σύμβαση, ένα υψηλό σκορ είναι ενδεικτικό για ένα πολύτιμο χαρακτηριστικό.

Μια απλή μέθοδος επιλογής χαρακτηριστικών χρησιμοποιώντας το Variable Ranking, είναι η επιλογή των χαρακτηριστικών k με την υψηλότερη κατάταξη σύμφωνα με το S. Αυτό συνήθως δεν είναι βέλτιστο, αλλά συχνά προτιμάται από άλλες, πιο πολύπλοκες μεθόδους. Είναι υπολογιστικά αποδοτικό, καθώς χρειάζεται μόνο υπολογισμός και διάταξη των βαθμολογιών n.

- Feature Subset Selection: Ο στόχος του παραπάνω, είναι να βρεθεί το υποσύνολο των βέλτιστων χαρακτηριστικών. Οι μέθοδοι επιλογής χαρακτηριστικών μπορούν να ταξινομηθούν σε τρεις ευρείες κατηγορίες:

    - Filter Methods: αξιολογεί τη συνάφεια των χαρακτηριστικών με την εξέταση των εσωτερικών ιδιοτήτων των δεδομένων. Στα κριτήρια φίλτρων, όλα τα χαρακτηριστικά βαθμολογούνται και ταξινομούνται βάσει ορισμένων στατιστικών κριτηρίων. Επιλέγονται τα χαρακτηριστικά με τις υψηλότερες τιμές κατάταξης και τα χαρακτηριστικά χαμηλής βαθμολογίας αφαιρούνται. Οι μέθοδοι φιλτραρίσματος είναι γρήγορες και ανεξάρτητες από τον ταξινομητή και μπορούν επίσης να κλιμακώνονται εύκολα σε datasets με πολύ υψηλές διαστάσεις. Ως αποτέλεσμα, η επιλογή χαρακτηριστικών πρέπει να γίνεται μόνο μία φορά και στη συνέχεια μπορούν να αξιολογηθούν διαφορετικοί ταξινομητές. Το μειονέκτημα των μεθόδων φιλτραρίσματος, είναι ότι αγνοούν την αλληλεπίδραση με τον ταξινομητή και κάθε χαρακτηριστικό θεωρείται ανεξάρτητο αγνοώντας έτσι τις εξαρτήσεις των χαρακτηριστικών.

    - Wrapper Methods: Αυτές οι μέθοδοι ενσωματώνουν την αναζήτηση υποδείγματος μοντέλου στην αναζήτηση υποσυνόλου χαρακτηριστικών. Σε αυτή τη ρύθμιση, ορίζεται μια διαδικασία αναζήτησης στο χώρο των πιθανών υποσυνόλων χαρακτηριστικών και παράγονται και αξιολογούνται διάφορα υποσύνολα χαρακτηριστικών. Η αξιολόγηση ενός συγκεκριμένου υποσυνόλου επιτυγχάνεται με την εκπαίδευση και τη δοκιμή ενός

συγκεκριμένου μοντέλου ταξινόμησης, κάνοντας αυτή την προσέγγιση προσαρμοσμένη σε έναν συγκεκριμένο αλγόριθμο ταξινόμησης. Για να αναζητηθεί ο χώρος όλων των υποσυνόλων χαρακτηριστικών, ένας αλγόριθμος αναζήτησης "τυλίγεται" γύρω από το μοντέλο ταξινόμησης. Ωστόσο, καθώς ο χώρος των υποσυνόλων χαρακτηριστικών αυξάνεται εκθετικά με τον αριθμό των χαρακτηριστικών, χρησιμοποιούνται μέθοδοι ευρετικής αναζήτησης (heuristics) για να καθοδηγήσουν την αναζήτηση ενός βέλτιστου υποσυνόλου.

- Embedded Methods: Οι ενσωματωμένες μέθοδοι πραγματοποιούν μεταβλητή επιλογή στη διαδικασία της εκπαίδευσης και συνήθως είναι συγκεκριμένες σε συγκεκριμένες μηχανές μάθησης (πχ classifications trees). Μοιάζουν με τις wrapper μεθόδους αλλά χρειάζονται λιγότερη υπολογιστική ισχύ. Κάνουν καλύτερη χρήση των διαθέσιμων δεδομένων, επειδή δεν χρειάζεται να χωρίζουν τα δεδομένα εκπαίδευσης σε σύνολα training και validation. Επιλύουν ταχύτερα, αποφεύγοντας την επανεκπαίδευση ενός προγνωστικού από το μηδέν για κάθε υποσύνολο μεταβλητών που ερευνήθηκε.

2.2.3 Mahalanobis Distance

Η απόσταση Mahalanobis (MD) είναι η απόσταση μεταξύ δύο σημείων στο χώρο πολλαπλών μεταβλητών. Για τις μη συσχετισμένες μεταβλητές, η ευκλείδεια απόσταση ισούται με την MD. Ωστόσο, εάν συσχετιστούν δύο ή περισσότερες μεταβλητές, οι άξονες δεν είναι εφικτές και οι μετρήσεις γίνονται αδύνατες. Το MD λύνει αυτό το πρόβλημα μέτρησης, καθώς μετρά αποστάσεις μεταξύ σημείων, ακόμη και συσχετισμένα σημεία για πολλαπλές μεταβλητές.

Η πιο συνηθισμένη χρήση για την απόσταση Mahalanobis είναι να βρεθούν outliers, τα οποία δείχνουν ασυνήθιστους συνδυασμούς δύο ή περισσοτέρων μεταβλητών.

Η απόσταση Mahalanobis μεταξύ δύο αντικειμένων ορίζεται ως:

$$d(\text{Mahalanobis}) = [(x_B - x_A)^T * \mathbf{C}^{-1} * (x_B - x_A)]^{0.5}$$

όπου x_A και x_B είναι ένα ζευγάρι αντικειμένων, και C είναι ο πίνακας των συνδιακύμανσης τους. Εναλλακτικά, χρησιμοποιείται και ως εξής, για την εύρεση της απόστασης από το μέσο όρο:

$$di = [(xi - \bar{x})^T C^{-1}(xi - \bar{x})]^{0.5}$$

2.2.4 Επίδοση αλγορίθμων Μηχανικής Μάθησης

Παρακάτω περιγράφονται μερικές μετρικές αξιολόγησης των αλγορίθμων μηχανικής μάθησης:

- Precision: Η μετρική της ακρίβειας (precision) εκτιμά την ορθότητα των αποτελεσμάτων. Αν σε ένα σύστημα, στο οποίο όλες οι ορθές απαντήσεις είναι Α και λαμβάνουμε απαντήσεις Β, από τις οποίες είναι σωστές οι Ν, τότε η ακρίβεια χαρακτηρίζεται ως ο λόγος Ν/Β.

- Recall: Αυτή η μετρική εκτιμά την πληρότητα των αποτελεσμάτων. Αν σε ένα σύστημα, στο οποίο όλες οι ορθές απαντήσεις είναι Α και λαμβάνουμε απαντήσεις Β, από τις οποίες είναι σωστές οι Ν, τότε η ακρίβεια χαρακτηρίζεται ως ο λόγος Ν/Α.

- F-measure: Η μετρική F-measure δίνει μια εκτίμηση της επίδοσης και συνδυάζει τις δύο παραπάνω μετρικές και αποτελεί τον αρμονικό μέσο όρο τους. Ορίζεται από τον παρακάτω τύπο:

$$F = \frac{2}{\left(\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}\right)}$$

- Confusion Matrix: δείχνει τον αριθμό σωστών και εσφαλμένων προβλέψεων που έγιναν από το μοντέλο ταξινόμησης σε σύγκριση με τα πραγματικά αποτελέσματα (τιμή στόχος) στα δεδομένα. Ο πίνακας είναι NxN, όπου N είναι ο αριθμός των τιμών στόχων (κλάσεις). Η απόδοση τέτοιων μοντέλων αξιολογείται συνήθως χρησιμοποιώντας τα δεδομένα στον πίνακα. Ο παρακάτω πίνακας εμφανίζει έναν πίνακα σύγχυσης 2x2 για δύο κατηγορίες (θετικό και αρνητικό).

Confusion Matrix		Target			
		Positive	Negative		
Model	Positive	a	b	*Positive Predictive Value*	a/(a+b)
	Negative	c	d	*Negative Predictive Value*	d/(c+d)
		Sensitivity	*Specificity*	**Accuracy** = (a+d)/(a+b+c+d)	
		a/(a+c)	d/(b+d)		

- **Accuracy** : the proportion of the total number of predictions that were correct.
- **Positive Predictive Value** or **Precision** : the proportion of positive cases that were correctly identified.
- **Negative Predictive Value** : the proportion of negative cases that were correctly identified.
- **Sensitivity** or **Recall** : the proportion of actual positive cases which are correctly identified.
- **Specificity** : the proportion of actual negative cases which are correctly identified.

2.2.5 Computer Vision

Το Computer Vision είναι ένας τομέας της επιστήμης των υπολογιστών που δουλεύει για να επιτρέπει στους υπολογιστές να βλέπουν, να αναγνωρίζουν και να επεξεργάζονται εικόνες με τον ίδιο τρόπο που κάνει η ανθρώπινη όραση και στη συνέχεια να παρέχουν την κατάλληλη απόδοση.

Ο στόχος είναι όχι μόνο η όραση, αλλά και η επεξεργασία και η παροχή χρήσιμων αποτελεσμάτων με βάση την παρατήρηση. Στο δυσδιάστατο χώρο που εξετάζεται σε αυτή την περίπτωση, ο στόχος είναι η αναγνώριση των στοιχείων του παραστατικού.

Κεφάλαιο 3: Δεδομένα

3.1 Τύποι, κατηγορίες και process επεξεργασίας

Οι τύποι των δεδομένων που θα χρησιμοποιηθούν, έχουν ως εξής:

Η αρχική εισαγωγή των δεδομένων θα γίνεται με τη χρήση ενός παραστατικού σε μορφή PDF, μέσα από το οποίο οι πληροφορίες «διαβάζονται» μέσω μηχανισμών αναγνώρισης οι οποίοι έχουν δημιουργηθεί από τους administrator χρήστες τις υπηρεσίας. Οι συγκεκριμένοι μηχανισμοί, παράγονται χειροκίνητα με τη χρήση συντεταγμένων μέσα στο εκάστοτε PDF και οι πληροφορίες που περιέχουν εξάγονται με τη χρήση regular expressions.

Ύστερα από αυτή την εξαγωγή, τα δεδομένα που έχουν εξαχθεί φορτώνονται στη βάση δεδομένων της υπηρεσίας. Τα πεδία που αποθηκεύονται, μπορούν να βρίσκονται στις εξής μορφές:

- Αριθμητικά πεδία (πχ συνολική αξία παραστατικού)
- Αλφαριθμητικά πεδία (πχ ΑΦΜ, Επωνυμία, Τύπος Παραστατικού)
- Πεδία ημερομηνιών (πχ Ημερομηνία παραστατικού, ημερομηνία παραγγελίας)
- Ακέραια πεδία (πχ συντεταγμένες πεδίων μέσα στο pdf)

Με την αποθήκευσή τους, τα παραπάνω δεδομένα είναι διαθέσιμα για την μεταφορά τους στον αλγόριθμο αυτόματης κατηγοριοποίησης και αναγνώρισης που αναλύεται στα επόμενα κεφάλαια.

Τέλος, τα δεδομένα που είναι διαθέσιμα στον χρήστη ύστερα από την επεξεργασία του παραστατικού, είναι όλες οι πληροφορίες που περιέχει, όπως ο αριθμός του, η ημερομηνία του, οι εμπλεκόμενοι φορείς, η πληρωτέα αξία του και άλλα στοιχεία που του χρησιμεύουν για τη χρήση και αρχειοθέτησή του.

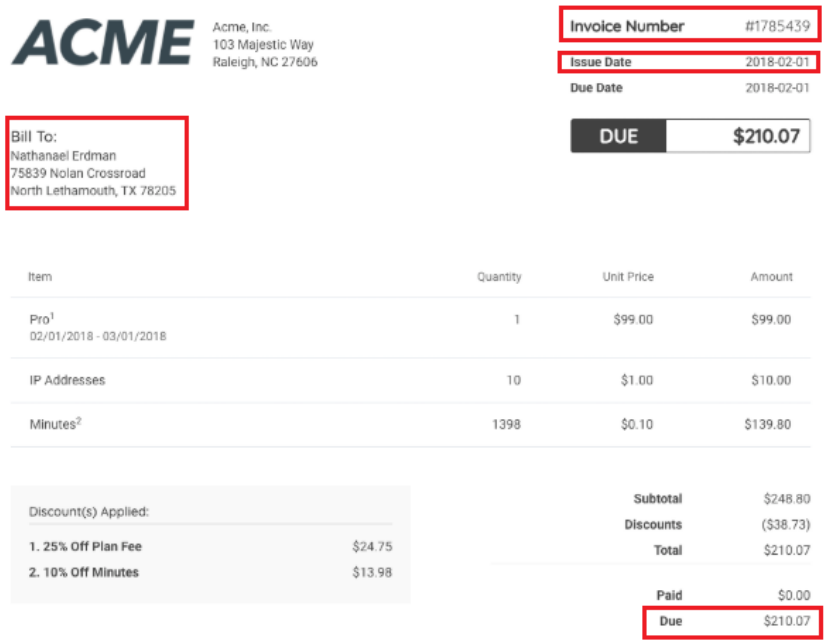
3.2 Σύνοψη και περιγραφή μηχανισμών αναγνώρισης

Προκειμένου να ξεκινήσει το παραπάνω εγχείρημα και προτού αναλυθούν οι αλγόριθμοι και οι τεχνικές που ακολουθούνται, χρειάζεται η εξαγωγή, περιγραφή και έλεγχος της ποιότητας των δεδομένων που υπάρχουν διαθέσιμα. Το αντικείμενο που θα εξεταστεί, είναι ένας

μεγάλος αριθμός παραστατικών (περίπου ένα εκ. δείγματα) και οι πληροφορίες που έχουν εξαχθεί από αυτά ύστερα από την επεξεργασία τους από το σύστημα.

Η προαναφερθείσα εξαγωγή, έχει πραγματοποιηθεί με την χειροκίνητη δημιουργία μηχανισμών αναγνώρισης για την κάθε μορφή παραστατικού, μέσω της οποίας είναι γνωστές οι μορφές ημερομηνιών και ποσών για την κάθε περίπτωση, η στοίχιση της σελίδας και τα πεδία που περιέχονται (ΑΦΜ, μοναδικά στοιχεία παραστατικού, σύνολα κλπ).

Για την εξαγωγή της πληροφορίας από τα παραστατικά, λόγω των πολλαπλών Layers σε κάθε pdf, αλλά και των διαφορετικών μεγεθών των γραμματοσειρών που ενδέχεται να περιέχονται σε αυτά, έχουν αναπτυχθεί τεχνικές computer vision για τη μετατροπή των παραπάνω σε απλό κείμενο.


ACME

Acme, Inc.
103 Majestic Way
Raleigh, NC 27606

Invoice Number #1785439
Issue Date 2018-02-01
Due Date 2018-02-01
DUE $210.07

Bill To:
Nathanael Erdman
75839 Nolan Crossroad
North Lethamouth, TX 78205

Item	Quantity	Unit Price	Amount
Pro[1] 02/01/2018 - 03/01/2018	1	$99.00	$99.00
IP Addresses	10	$1.00	$10.00
Minutes[2]	1398	$0.10	$139.80

Discount(s) Applied:
1. 25% Off Plan Fee $24.75
2. 10% Off Minutes $13.98

Subtotal $248.80
Discounts ($38.73)
Total $210.07
Paid $0.00
Due $210.07

Ως εκ των άνω, όταν ένα παραστατικό εισαχθεί στο μηχανισμό που έχει αναπτυχθεί για την αναγνώριση των παραστατικών με βάση τα παραπάνω, για τη δημιουργία της φόρμας αναγνώρισης, ο χρήστης ορίζει τα τμήματα στα οποία χωρίζεται το παραστατικό (Header, Details και Footer), δίνει ως input πληροφορίες όπως τη μορφή της ημερομηνίας για τη συγκεκριμένη περίπτωση και στη συνέχεια επιλέγει μέσα στο pdf για κάθε πεδίο του παραστατικού χωριστά, το χώρο μέσα στον οποίο βρίσκεται, προκειμένου στη συνέχεια να εξαχθεί η πληροφορία του.

Η κάθε μία από τις human – driven διαδικασίες που περιγράφονται σε προηγούμενο κεφάλαιο, περιέχει τα εξής στοιχεία:

- Αν το παραστατικό περιέχει τις γραμμές των προϊόντων σε σταθερή θέση σε όλες τις σελίδες ή αν η περιοχή μέσα στις οποίες υπάρχουν μετακινείται.
- Την γραμμή έναρξης των γραμμών και τη γραμμή λήξης τους.
- Τη μορφή των ημερομηνιών (πχ dd/MM/YYYY).
- Τη μορφή των αξιών (πχ X.XXX,XX).
- Τη διάταξη του παραστατικού (Portrait ή Landscape).
- Τη διάταξη των βασικών πεδίων (αν είναι διατεταγμένα από πάνω προς τα κάτω ή αριστερά προς τα δεξιά).
- Τη γλώσσα του παραστατικού για το κατάλληλο text classification, προκειμένου το παραστατικό να διαχωριστεί κατάλληλα στις περιοχές Κεφαλίδας, Γραμμών και Συνολικών Αξιών.
- Τον αριθμό των πεδίων σε κάθε περιοχή.

Όσον αφορά τα πεδία που περιέχονται σε κάθε περιοχή, υπάρχουν τα παρακάτω στοιχεία για το κάθε ένα:

Περιγραφή	Παράδειγμα
Όνομα πεδίου	ΑΦΜ Προμηθευτή
Συντεταγμένες στους άξονες X,Y	$X_1 = 5$, $X_2 = 25$, $Y_1 = 2$, $Y_2 = 20$
Περιοχή στην οποία ανήκει μέσα στο παραστατικό.	Κεφαλίδα, Γραμμές ή Σύνολα
Regular Expression εξαγωγής[11]	(?<=ΑΦΜ\s*)\d{9} (Παίρνει μετά το λεκτικό ΑΦΜ τα επόμενα 9 νούμερα – ελληνικά δεδομένα)
Regular Expression επικύρωσης	\d{9} (Επιβεβαιώνει την ορθότητα της πληροφορίας)
Την πληροφορία που υπάρχει γύρω από το πεδίο σε μορφή αντικειμένου με συγκεκριμένη κωδικοποίηση για την εισαγωγή στους αλγορίθμους.	Πιθανές λέξεις/φράσεις: Επωνυμία, Κωδικός, Δ.Ο.Υ. κτλ

Επιπροσθέτως, διαθέσιμος είναι και ένας πίνακας στην παρακάτω μορφή που αποτελείται από όλα τα διαθέσιμα πεδία και η ύπαρξή τους στον εκάστοτε μηχανισμό σημειώνεται με 0 (μη ύπαρξη) ή 1 (ύπαρξη).

[1] Το κάθε regular expression είναι μια ακολουθία συμβόλων και χαρακτήρων που εκφράζουν μια συμβολοσειρά ή μοτίβο που πρέπει να αναζητηθεί μέσα σε ένα μεγαλύτερο κομμάτι κειμένου.

Μηχανισμός	Καθαρή Αξία	Συνολική Αξία Φόρων	…	Πεδίο N
Μηχανισμός 1	1	0		1
Μηχανισμός 2	1	1		1
…				
Μηχανισμός N	0	1		0

Η ύπαρξη του παραπάνω πίνακα βοηθά στην κατανόηση των δεδομένων που υπάρχουν διαθέσιμα, καθώς και στους τρόπους classification που θα αναλυθούν παρακάτω και θα χρησιμοποιηθεί για το διαχωρισμό των περιπτώσεων. Ο έλεγχος της ποιότητας των δεδομένων που βρίσκονται διαθέσιμα δε θα γίνει εκ νέου για τη δημιουργία του μοντέλου, καθώς αυτή έχει ήδη επικυρωθεί από την απρόσκοπτη λειτουργία των μηχανισμών στην καθημερινή χρήση τους για την ανταλλαγή παραστατικών.

3.3 Επιλογή μεταβλητών

Στο αρχικό στάδιο ανάπτυξης για τα πεδία που βρίσκονται διαθέσιμα, είναι γνωστό πως τα κύρια από αυτά και που υπάρχουν πάντοτε σε ένα παραστατικό, είναι εκείνα που εξασφαλίζουν τη μοναδικότητά του, τα οποία είναι: ο αριθμός του παραστατικού, η ημερομηνία έκδοσης, τα ΑΦΜ των εταιριών που συναλλάσσονται, ο τύπος παραστατικού, η πληρωτέα αξία και στη συνέχεια η σειρά του παραστατικού και οι επωνυμίες των εταιριών.

Τα παραπάνω πεδία δεν έχουν επιλεγεί αυθαίρετα, αλλά βασίζονται στο UBL standard που είναι αποδεκτό και χρησιμοποιείται με βάση την ελληνική νομοθεσία για την ηλεκτρονική αποστολή εγγράφων, το οποίο τα θέτει απαραίτητα σε κάθε ηλεκτρονικό παραστατικό.

Ως εκ τούτου, οι μεταβλητές που θα χρησιμοποιηθούν για την εκπαίδευση του μοντέλου είναι οι παραπάνω, καθώς και οι συντεταγμένες της κάθε μίας στο χώρο, δοθέντος της συμμετοχής τους σε κάθε μηχανισμό αναγνώρισης, κάτι που οδηγεί στο συμπέρασμα πως αποτελούν τις πιο κρίσιμες μεταβλητές για την κατάταξη και δημιουργία των περιπτώσεων που έχουμε στη διάθεσή μας.

Στη συνέχεια, με βάση το Ranking που θα προκύψει από την ύπαρξη επιπλέον πεδίων στο participation matrix, κάτι που υπολογίζεται από τον αριθμό εμφανίσεων του εκάστοτε πεδίου

στο σύνολο των μηχανισμών, ενδέχεται να επιλεγούν επιπλέον πεδία ή σύνολο πεδίων για τη βελτιστοποίηση των τρόπων επιλογής του κριτηρίου διαχωρισμού των κατηγοριών.

3.4 Association Rules

Ως κριτήριο επιλογής μεταβλητών, ενδέχεται να χρησιμοποιηθούν και association rules, που θα βοηθήσουν στην κατανόηση της σημαντικότητας της κάθε μεταβλητής και θα εμφανίσουν πιθανές συσχετίσεις μεταξύ τους.

Το association rule learning, είναι μία machine learning μέθοδος βασισμένη σε κανόνες για την ανακάλυψη ενδιαφερόντων σχέσεων μεταξύ μεταβλητών σε μεγάλες βάσεις δεδομένων. Σκοπός του είναι να εντοπίσει τους ισχυρούς κανόνες που ανακαλύπτονται σε βάσεις δεδομένων χρησιμοποιώντας ορισμένα μέτρα ενδιαφέροντος. Αυτή η προσέγγιση βάσει κανόνων δημιουργεί επίσης νέους κανόνες καθώς αναλύει περισσότερα δεδομένα. Ο τελικός στόχος, είναι να βοηθήσουμε μια μηχανή να μιμείται την εξαγωγή χαρακτηριστικών του ανθρώπινου εγκεφάλου και τις αφηρημένες δυνατότητες σύνδεσης από νέα μη ταξινομημένα δεδομένα.

Το πρόβλημα μπορεί να γίνει ορατό ως εξής:

- *Έστω I = {i_1,i_2,….} ένα σύνολο δυαδικών χαρακτηριστικών που ονομάζονται στοιχεία.*
- *Έστω D = {t_1,t_2,….} ένα σύνολο συναλλαγών που ονομάζονται βάση δεδομένων.*

Κάθε μοναδική συναλλαγή D, περιέχει ένα υποσύνολο αντικειμένων I. Ένας κανόνας συσχέτισης, έχει τη μορφή:

$$X \rightarrow Y, \text{όταν } X, Y \subset I$$

$$Rule: X \Rightarrow Y \begin{cases} Support = \dfrac{frq(X,Y)}{N} \\ Confidence = \dfrac{frq(X,Y)}{frq(X)} \\ Lift = \dfrac{Support}{Supp(X) \times Supp(Y)} \end{cases}$$

Υπάρχουν διάφορα ποσοτικά μεγέθη που καθορίζουν πόσο ισχυρός είναι ο παραπάνω κανόνας. Αυτές οι μετρικές είναι οι **Support**, **Confidence** και **Lift.**

- Support: Αυτή η μετρική, εμφανίζει πόσο συχνά ένα σύνολο στοιχείων εμφανίζεται στη βάση δεδομένων. Ορίζεται ως το κλάσμα των εγγραφών που περιέχουν το X∪Y προς το συνολικό αριθμό εγγραφών στη βάση δεδομένων. Αν το support ενός στοιχείου είναι 0,1%, αυτό σημαίνει ότι μόνο το 0,1% των συναλλαγών περιέχει αυτό το στοιχείο.

$$Support\ (XY) = Count\ of\ (XY) / Total\ number\ of\ transactions\ in\ D$$

- Confidence: Είναι μία μετρική δύναμης των κανόνων σύνδεσης. Υπολογίζεται ως το κλάσμα των εγγραφών που περιέχουν το X∪Y, προς το συνολικό αριθμό εγγραφών που περιέχουν το X. Σε περίπτωση που το confidence ενός κανόνα είναι 80%, σημαίνει ότι το 80% των συναλλαγών που περιέχουν το X, τότε περιέχουν και το Y.

$$Confidence\ (X|Y) = Support\ (XY) / Support\ (X)$$

- Lift: Ορίζεται ως το κλάσμα του support(X∩Y) προς το γινόμενο των support των X και Y. Αν το Lift είναι 1, τότε θα σήμαινε πως οι πιθανότητες εμφάνισης του X και του Y είναι ανεξάρτητες η μία από την άλλη. Σε αυτή την περίπτωση κανένας κανόνας δεν μπορεί να εξαχθεί από αυτά τα 2 γεγονότα. Αν το Lift έχει τιμή μεγαλύτερη του 1, τότε εμφανίζεται το πόσο εξαρτώνται τα δύο ενδεχόμενα μεταξύ τους. Αν είναι μικρότερο του 1, τότε το ένα ενδεχόμενο είναι υποκατάστατο του άλλου. Πρακτικά, η παρουσία ενός στοιχείου έχει αρνητική επίδραση στην παρουσία του άλλου και αντίστροφα.

$$Lift\ (XY) = Support\ (X \cap Y) / (\ Support\ (X) * Support(Y))$$

Τα Association Rules βρίσκουν όλα τα σύνολα αντικειμένων που έχουν τιμή support μεγαλύτερη από την ελάχιστη και στη συνέχεια χρησιμοποιούν μεγάλα σύνολα δεδομένων για να δημιουργήσουν τα επιθυμητούς κανόνες που έχουν τιμή confidence μεγαλύτερη της

ATHENS UNIVERSITY OF ECONOMICS & BUSINESS
ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

ελάχιστης. Ένα τυπικό και ευρέως χρησιμοποιούμενο παράδειγμα εφαρμογής των κανόνων σύνδεσης είναι η Market Basket Analysis.

Κεφάλαιο 4: Αλγόριθμοι Classification

4.1 Τύποι αλγορίθμων και χρησιμοποίηση του καθενός

Το πρώτο πρόβλημα που προκύπτει ύστερα από τη συλλογή των δεδομένων, είναι ο τρόπος που θα διαχωριστούν. Οι αλγόριθμοι μηχανικής μάθησης διαχωρίζονται σε επιτηρούμενους και μη επιτηρούμενους που χρησιμοποιούνται για την ανάλυση μεγάλου όγκου δεδομένων. Οι επιτηρούμενοι αλγόριθμοι, χρησιμοποιούν τα δεδομένα που βρίσκονται διαθέσιμα και μία επεξήγηση δείγματος, με σκοπό να δημιουργήσουν ένα μοντέλο που είναι σε θέση να προβλέψει την τιμή της επεξήγησης του δείγματος σε ένα νέο δείγμα, έχοντας μόνο τα νέα δεδομένα – περίπτωση Classification. Όσον αφορά τους αλγορίθμους μη επιβλεπόμενης μάθησης, δεν χρησιμοποιούν καμία πληροφορία σχετικά με τις επεξηγήσεις του δείγματος και προσπαθούν να βρουν δομή και μοτίβα στα δεδομένα τα οποία στη συνέχεια μπορούν να ερμηνευθούν. Ένα παράδειγμα μη επιτηρούμενης μεθόδου είναι το Clustering, όπου ο στόχος είναι η εύρεση υποομάδων στα δεδομένα.

Σύμφωνα με τα υπάρχοντα δεδομένα και εφόσον οι χρήστες γνωρίζουν τη μορφή τους, η βέλτιστη τακτική φαίνεται πως είναι η επιβλεπόμενη μάθηση και ως εκ τούτου θα εξεταστούν αλγόριθμοι classification όπως οι:

- Naïve Bayes
- Decision Trees
- Support Vector Machines

4.2 Naïve Bayes

Πρόκειται για μια τεχνική classification βασισμένη στο Θεώρημα του Bayes με την παραδοχή της ανεξαρτησίας μεταξύ των predictors. Ο Naive Bayes υποθέτει ότι η παρουσία ενός συγκεκριμένου χαρακτηριστικού σε μια κλάση δεν σχετίζεται με την παρουσία οποιουδήποτε άλλου. Ακόμη και αν τα χαρακτηριστικά εξαρτώνται το ένα από το άλλο ή από την ύπαρξη των άλλων χαρακτηριστικών, όλες αυτές οι ιδιότητες συμβάλλουν ανεξάρτητα στην πιθανότητα να ισχύει μία υπόθεση και γι' αυτό είναι γνωστός ως «Naive».

Το μοντέλο Naive Bayes είναι εύκολο να κατασκευαστεί και ιδιαίτερα χρήσιμο για πολύ μεγάλα σύνολα δεδομένων όπως αυτό που αντιμετωπίζεται σε αυτή την περίπτωση. Μαζί με

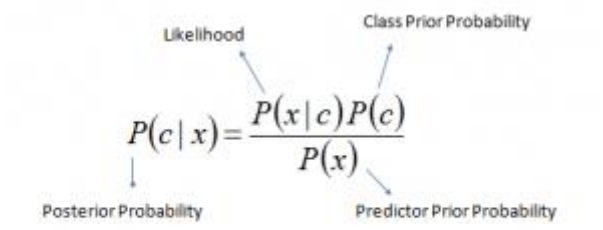
ATHENS UNIVERSITY OF ECONOMICS & BUSINESS ΟΙΚΟΝΟΜΙΚΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΑΘΗΝΩΝ

την απλότητα, το Naive Bayes είναι γνωστό ότι ξεπερνάει ακόμη και πολύ εξελιγμένες μεθόδους classification.

Τα πλεονεκτήματά του είναι πως είναι γρήγορος να εκπαιδευτεί, μπορεί να εφαρμοστεί ως χαρακτηριστικό πρόβλεψης πολλαπλών κατηγοριών και χρησιμοποιείται στην πλειοψηφία των περιπτώσεων σε text classification προβλήματα, κάτι που θα αντιμετωπιστεί σε μετέπειτα στάδια.

Τα μειονεκτήματα που τον διέπουν, είναι πως σε περίπτωση που στον έλεγχο του μοντέλου εμφανιστεί μία κατηγορική μεταβλητή που δεν είχε εμφανιστεί στα δεδομένα της εκπαίδευσης, τότε θα λάβει μηδενική πιθανότητα και η πρόβλεψή της είναι αδύνατη, εκτός αν εφαρμοστούν τεχνικές διόρθωσης. Ένας άλλος περιορισμός των Naive Bayes είναι η υπόθεση των ανεξάρτητων χαρακτηριστικών. Στην πραγματική ζωή, είναι σχεδόν αδύνατο να έχουμε ένα σύνολο χαρακτηριστικών που είναι εντελώς ανεξάρτητα.

Ο τρόπος λειτουργίας του βασίζεται στο θεώρημα του Bayes για την εύρεση της πιθανότητας εμφάνισης ενός ενδεχομένου.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Ο αλγόριθμός χρησιμοποιεί μια παρόμοια μέθοδο για να προβλέψει την πιθανότητα διαφορετικής κλάσης με βάση διάφορα χαρακτηριστικά.

4.3 Decision Trees

Τα δένδρα αποφάσεων είναι ένας ακόμα τύπος αλγόριθμου επιβλεπόμενης μάθησης ο οποίος έχει μια προκαθορισμένη μεταβλητή-στόχο. Λειτουργεί τόσο για κατηγορικές όσο και για συνεχείς μεταβλητές. Σε αυτό τον αλγόριθμο, χωρίζουμε τα δεδομένα σε δύο ή περισσότερα ομοιογενή σύνολα με βάση τον σημαντικότερο διαχωριστή στις μεταβλητές εισόδου, με βάση αυτές που έχουμε διαθέσιμες.

Τα Δένδρα Απόφασης λειτουργούν ως εξής:

- Βασική προϋπόθεση αποτελεί, το ότι κάθε δείγμα (objects ή cases) του data set πρέπει να μπορεί να εκφράζεται ως μία συλλογή από τα χαρακτηριστικά του data set.
- Ο αλγόριθμος λαμβάνει ως είσοδο κάθε case ως ένα input vector με τις τιμές των μεταβλητών και την αντίστοιχη κλάση που ανήκει. Αυτά τα cases – input vectors αποτελούν το training set του classifier.

Στη συνέχεια, παρατηρώντας και συγκρίνοντας αν τα vectors, των οποίων οι τιμές είναι κοντά η μία στην άλλη, ανήκουν ή όχι στην ίδια κλάση, κατασκευάζει ένα σετ από κανόνες αποφάσεων (decision rules) με σκοπό την ταξινόμηση μελλοντικών δειγμάτων (test set) στις κλάσεις που έχουν δημιουργηθεί.

Η απόφαση για τη δημιουργία στρατηγικών διαχωρισμού επηρεάζει σε μεγάλο βαθμό την ακρίβεια ενός δέντρου. Τα δέντρα αποφάσεων χρησιμοποιούν πολλαπλούς αλγόριθμους για να αποφασίσουν να χωρίσουν έναν κόμβο σε δύο ή περισσότερους υπό-κόμβους. Η δημιουργία υπό-κόμβων αυξάνει την ομοιογένεια των υπόλοιπων υπό-κόμβων που προκύπτουν. Με άλλα λόγια, μπορούμε να πούμε ότι η καθαρότητα (purity) του κόμβου αυξάνεται σε σχέση με τη μεταβλητή-στόχο. Το δέντρο απόφασης χωρίζει τους κόμβους σε όλες τις διαθέσιμες μεταβλητές και στη συνέχεια επιλέγει τη διάσπαση που έχει ως αποτέλεσμα τους πιο ομοιογενείς υπό-κόμβους. Οι τέσσερις πιο συχνά χρησιμοποιούμενοι αλγόριθμοι στο δέντρο αποφάσεων είναι οι:

- Gini Index: αν επιλέξουμε δύο στοιχεία από έναν πληθυσμό τυχαία τότε πρέπει να είναι της ίδιας τάξης και η πιθανότητα για αυτό είναι 1 εάν ο πληθυσμός είναι καθαρός. Όσο υψηλότερη είναι η τιμή του Gini, τόσο μεγαλύτερη είναι η ομοιογένεια.

- Chi-Square: Είναι ένας αλγόριθμος για την ανίχνευση της στατιστικής σημαντικότητας μεταξύ των διαφορών των υπό-κόμβων με τον κεντρικό κόμβο. Υπολογίζεται ως το άθροισμα των τετραγώνων των τυποποιημένων αποκλίσεων μεταξύ των παρατηρούμενων και των αναμενόμενων συχνοτήτων της μεταβλητής-στόχου. Όσο υψηλότερη είναι η τιμή του Chi-Square, τόσο μεγαλύτερη είναι η στατιστική σημασία των διαφορών μεταξύ του υπό-κόμβου και του κεντρικού κόμβου.

- Information Gain: Είναι ένα μέτρο για τον ορισμό του βαθμού αποδιοργάνωσης σε ένα σύστημα γνωστό ως Εντροπία. Εάν το δείγμα είναι εντελώς ομοιογενές, τότε η εντροπία είναι μηδέν και εάν το δείγμα είναι εξίσου διαιρεμένο (50% - 50%), έχει εντροπία ένα. Όσο μικρότερη είναι η εντροπία, τόσο το καλύτερο.

- Reduction in Variance: Ενώ οι τρεις παραπάνω αλγόριθμοι αφορούν κατηγορικές μεταβλητές, ο παρών χρησιμοποιείται για συνεχείς και προβλήματα παλινδρόμησης και χρησιμοποιεί την διακύμανση για να επιλέξει τον καλύτερο διαχωρισμό. Ο διαχωρισμός με χαμηλότερη διακύμανση επιλέγεται ως τα κριτήριο για τη διάσπαση του δείγματος.

Τα δένδρα απόφασης διαχωρίζονται σε δύο κατηγορίες. Αν η μεταβλητή-στόχος είναι κατηγορική, τότε το δένδρο λέγεται Categorical ή Binary Variable Decision Tree και καταλήγει σε είδος απόφασης ναι ή όχι και αν η μεταβλητή είναι συνεχής, ονομάζεται Continuous Variable Decision Tree.

Ορισμένα από τα πλεονεκτήματα των δέντρων απόφασης, είναι πως αποτελούν ένα ευκολονόητο μοντέλο και είναι χρήσιμα για την κατανόηση των δεδομένων και των σημαντικών μεταβλητών του μοντέλου.

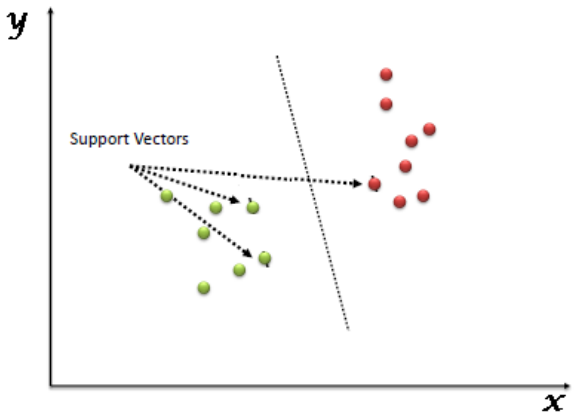
Το βασικό μειονέκτημά τους, είναι πως αν το μοντέλο λειτουργεί παραπάνω από καλά, υπάρχει ο κίνδυνος του overfitting και κατά συνέπεια της λανθασμένης κατηγοριοποίησης του αποτελέσματος. Το παραπάνω μπορεί να αντιμετωπιστεί θέτοντας ορισμένους κανόνες - περιορισμούς κατά τη δημιουργία του μοντέλου και με το «κλάδεμα» του δέντρου. Ένας άλλος τρόπος αντιμετώπισης, είναι η χρήση τυχαίων δασών (Random Forests).

Το Random Forest, δημιουργεί πολλαπλά δέντρα απόφασης που το καθένα εκπαιδεύεται με ένα τυχαίο υποσύνολο των χαρακτηριστικών εισόδου και τα συγχωνεύει για να πάρει μια πιο ακριβή και σταθερή πρόβλεψη. Προσθέτει επιπλέον τυχαιότητα στο μοντέλο και αντί να ψάχνει για το πιο σημαντικό χαρακτηριστικό για τη διαίρεση ενός κόμβου, αναζητά το καλύτερο δυνατό αποτέλεσμα μεταξύ ενός τυχαίου υποσυνόλου χαρακτηριστικών. Αυτό έχει ως απόρροια μια ευρεία ποικιλομορφία που γενικά οδηγεί σε ένα καλύτερο μοντέλο.

Ένα επιπλέον πλεονέκτημα που παρέχει το τυχαίο δάσος, είναι πως μπορεί να χρησιμοποιηθεί για να εντοπιστούν τα κυριότερα χαρακτηριστικά από το σύνολο των δεδομένων που εκπαίδευσαν το μοντέλο, για το βέλτιστο feature engineering.

4.4 Support Vector Machines

Σε αυτόν τον αλγόριθμο, κάθε στοιχείο δεδομένων σχεδιάζεται ως σημείο σε n-διάστατο χώρο (όπου n είναι ο αριθμός των χαρακτηριστικών που υπάρχουν διαθέσιμα) με την αξία κάθε χαρακτηριστικού να είναι η τιμή μιας συγκεκριμένης συντεταγμένης. Στη συνέχεια, πραγματοποιούμε την ταξινόμηση βρίσκοντας το υπερ-επίπεδο (hyper-plane) που διαφοροποιεί καλύτερα τις κατηγορίες.



Τα SVMs παίρνουν χώρο χαμηλών διαστάσεων εισόδου και μετατρέπουν το σε ένα υψηλότερο χώρο διαστάσεων και ως επί το πλείστον είναι χρήσιμο σε πρόβλημα μη γραμμικού διαχωρισμού, πραγματοποιώντας περίπλοκους μετασχηματισμούς δεδομένων και στη συνέχεια, ανακαλύπτει τη διαδικασία διαχωρισμού των δεδομένων με βάση τις ετικέτες ή τις εξόδους που έχουν οριστεί.

Τα πλεονεκτήματα των SVM είναι η αποτελεσματικότητά τους σε χώρους με μεγάλες διαστάσεις και χρησιμοποιούνται σε περιπτώσεις όπου ο αριθμός των διαστάσεων είναι μεγαλύτερος από τον αριθμό των δειγμάτων.

Το κύριο αρνητικό τους στοιχείο είναι πως δεν λειτουργούν καλά όταν το σύνολο δεδομένων έχει περισσότερο θόρυβο, δηλαδή οι κατηγορίες-στόχοι αλληλεπικαλύπτονται (όπως συμβαίνει στην παρούσα περίπτωση) και πως όταν έχουμε μεγάλο σύνολο δεδομένων, επειδή ο απαιτούμενος χρόνος εκπαίδευσης είναι υψηλότερος.

4.5 Εφαρμογή Mahalanobis Distance ως κριτήριο κατηγοριοποίησης

Σύμφωνα με όσα περιγράφονται στο υποκεφάλαιο 2.2.3, υπάρχει η δυνατότητα ένταξης της μετρικής Mahalanobis Distance προκειμένου να εισαχθεί ως κριτήριο στο μοντέλο που θα εκπαιδευτεί, για την καλύτερη κατηγοριοποίηση των περιπτώσεων.

Παρακάτω παρουσιάζονται τα γεωγραφικά δεδομένα για το πεδίο «ΑΦΜ Προμηθευτή» από 15 διαφορετικούς μηχανισμούς αναγνώρισης παραστατικών, πάνω στα οποία εφαρμόζεται MD.

	X1	X2	Y1	Y2
ΑΦΜ 1	8	19	8	15
ΑΦΜ 2	14	25	7	15
ΑΦΜ 3	8	19	8	14
ΑΦΜ 4	9	23	2	10
ΑΦΜ 5	14	29	8	17
ΑΦΜ 6	11	23	5	13
ΑΦΜ 7	7	19	9	16
ΑΦΜ 8	5	19	6	12
ΑΦΜ 9	6	18	6	13
ΑΦΜ 10	7	19	704	713
ΑΦΜ 11	6	18	6	11
ΑΦΜ 12	6	18	6	15
ΑΦΜ 13	14	28	10	16
ΑΦΜ 14	7	18	8	18
ΑΦΜ 15	85	105	4	11

Για τον υπολογισμό του Mahalanobis' Distance, δημιουργήθηκε πίνακας διακυμάνσεων, συνδιακυμάνσεων και ο αντίστροφος αυτού.

var(x1)	var(x2)	var(y1)	var(y2)		
397.6	483.3809524	32424.55238	32578.82857		
covarX1X2	covarX1Y1	covarX1Y2	covarX2Y1	covarX2Y2	covarY1Y2
437.7142857	-351.6142857	-353.5142857	-396.5952381	-398.8571429	32500.7
covar-mat	X1	X2	Y1	Y2	
X1	397.6	437.7142857	-351.6142857	-353.5142857	
X2	437.7142857	483.3809524	-396.5952381	-398.8571429	
Y1	-351.6142857	-396.5952381	32424.55238	32500.7	
Y2	-353.5142857	-398.8571429	32500.7	32578.82857	
inv-covar-mat	X1	X2	Y1	Y2	
X1	0.815116896	-0.738431392	0.059241902	-0.059295454	
X2	-0.738431392	0.671054426	-0.055192528	0.055263025	
Y1	0.059241902	-0.055192528	0.56295614	-0.561638969	
Y2	-0.059295454	0.055263025	-0.561638969	0.560355934	

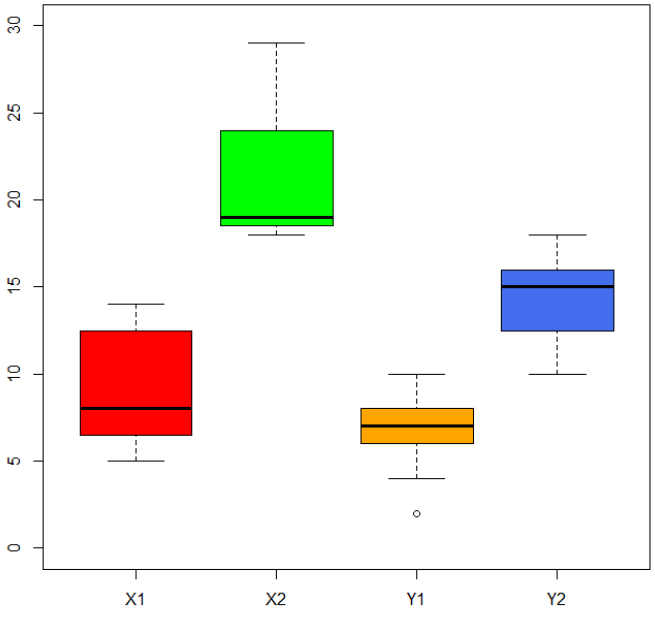
Επίσης, εφαρμόστηκε ο τύπος τόσο για τη διαφορά μεταξύ δύο διανυσμάτων των οποίων οι συντεταγμένες είναι κοντινές, όσο και για τη διαφορά μεταξύ δύο, τα οποία δεν είναι σε κοντινή απόσταση.

ΑΦΜ 1	8	19	8	15		ΑΦΜ 1 - ΑΦΜ 7	0.815170449	-0.738501888	0.057924731	-0.058012419
ΑΦΜ 7	7	19	9	16		1				
ΑΦΜ 13	14	28	10	16		0		0.815258137		
						-1	MD =	0.902916462		
						-1				
ΑΦΜ 1 - ΑΦΜ 7	1	0	-1	-1						
ΑΦΜ 1 - ΑΦΜ 13	-6	-9	-2	-1						
						ΑΦΜ 1 - ΑΦΜ 13	1.695992799	-1.55377945	-0.422991967	0.421327509
						-6				
						-9		4.232714679		
						-2	MD =	2.057356235		
						-1				

Από τα παραπάνω, φαίνεται πως το ΑΦΜ στον 1ο μηχανισμό κινείται σχεδόν στην ίδια περιοχή με του 7ου και αρκετά μακριά από του 13ου.

Τέλος, παρουσιάζονται ορισμένα στατιστικά (min, max, mean, median, 1ο και 3ο τεταρτημόριο), ύστερα από την εισαγωγή των δεδομένων σε γλώσσα R, καθώς και ένα boxplot για τη γραφική τους απεικόνιση.

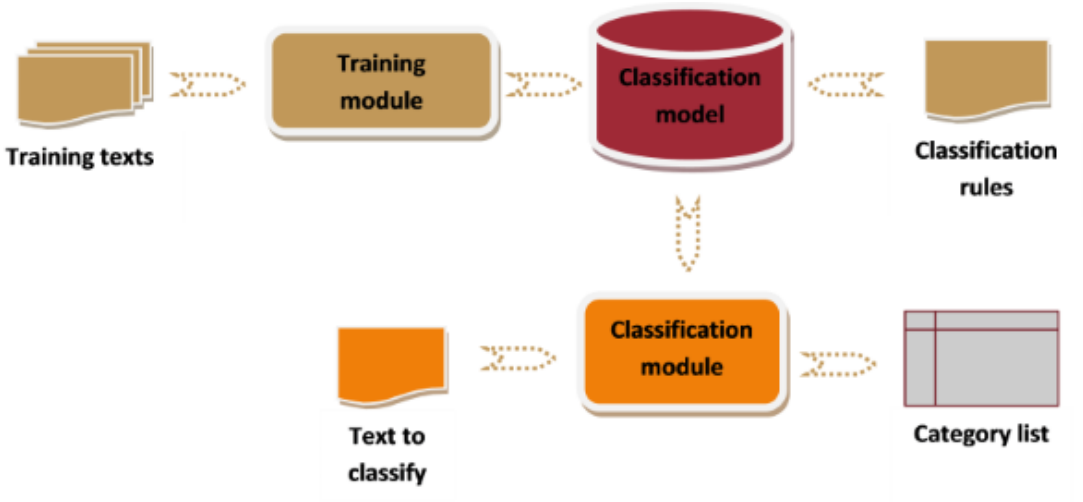
	x1	x2	y1	y2
Min.:	5.0	18.00	2.00	10.0
1st Qu.:	6.5	18.50	6.00	12.5
Median:	8.0	19.00	7.00	15.0
Mean:	13.8	26.67	53.13	60.6
3rd Qu.:	12.5	24.00	8.00	16.0
Max.:	85.0	105.00	704.00	713.0



4.6 Text Classification

Στο πρόβλημα του text classification, τα έγγραφα αντιπροσωπεύονται από ένα "bag-of-words", που είναι ένα διάνυσμα του μεγέθους του λεξιλογίου που περιέχει τη συχνότητα εμφάνισης των λέξεων. Τα λεξιλόγια εκατοντάδων χιλιάδων λέξεων είναι κοινά, αλλά το κλάδεμα των περισσότερων και λιγότερο συχνών λέξεων μπορεί να μειώσει τον αποτελεσματικό αριθμό λέξεων. Στην περίπτωση του παρόντος project, αυτές οι λέξεις ή φράσεις είναι συγκεκριμένες και οδηγούν στο διαχωρισμό των περιοχών του παραστατικού.

Ο αλγόριθμος συνδυάζει την ταξινόμηση εγγράφων με φιλτράρισμα βάσει κανόνων, το οποίο επιτρέπει την επίτευξη υψηλού βαθμού ακρίβειας σε ένα ευρύ φάσμα περιπτώσεων.



(https://www.meaningcloud.com/developer/resources/doc/models)

Το μοντέλο δημιουργείται από τα κείμενα που θα του δοθούν και του κανόνες κατηγοριοποίησης που θα ενταχθούν. Το σύστημα χρησιμοποιεί το παραπάνω μοντέλο προκειμένου να διαχωρίσει το κείμενο και να το κατατάξει σε μία ή παραπάνω κατηγορίες στις οποίες ενδέχεται να καταταγεί.

Υπάρχουν δύο είδη text classification και το μοντέλο μπορεί να χτιστεί είτε βάσει στατιστικών, είτε με τη χρήση κανόνων είτε με συνδυασμό των παραπάνω.

- Μοντέλο βάσει στατιστικών: Συγκρίνει το κείμενο εισαγωγής με τα ήδη υπάρχοντα κείμενα που συνετέλεσαν για το training του μοντέλου. Ως εκ τούτου, με βάση τα υπάρχοντα κείμενα, το σύστημα αποφασίζει ποιο από τα παραδείγματα είναι πιο κοντά στο κείμενο εισαγωγής.

Το πλεονέκτημα αυτής της μεθόδου είναι πως αν έχει γίνει σωστή κατηγοριοποίηση των λέξεων, η αρχική εκπαίδευση του μοντέλου είναι γρήγορη με καλά αποτελέσματα και προτείνεται για μεγαλύτερα κείμενα.

Το αρνητικό είναι πως αν η κάθε κατηγορία δεν έχει εκπαιδευτεί ομοιόμορφα, τότε το μοντέλο μπορεί να ευνοήσει τις κατηγορίες με περισσότερα δείγματα. Επίσης, το στατιστικό μοντέλο δεν προτείνεται για μικρά κείμενα, καθώς οι κατηγορίες ενδέχεται να αλληλοκαλύπτονται.

- Μοντέλο βάσει κανόνων: Το συγκεκριμένο μοντέλο χρησιμοποιεί αποκλειστικά κανόνες κατηγοριοποίησης. Αυτό σημαίνει πως ο διαχωρισμός γίνεται μόνο από τους κανόνες που έχουν οριστεί.

  Το πλεονέκτημα αυτής της μεθόδου είναι πως μειώνει σε μεγάλο βαθμό την πιθανότητα να κατηγοριοποιηθούν λανθασμένα τα δεδομένα ως false positive και την περιορίζει μόνο στις περιπτώσεις που έχουμε διφορούμενους όρους ή παρεμφερείς κανόνες σε περισσότερες από μία κατηγορίες.

  Το αρνητικό είναι πως αν υπάρχει περίπτωση που δεν έχει περιληφθεί σε κάποιο κανόνα, τότε δε θα ανιχνευθεί με αποτέλεσμα να αυξηθεί ο αριθμός των false negatives. Επίσης, οι κανόνες πρέπει να οριστούν υπερβολικά αναλυτικά, προκειμένου να επιτευχθεί υψηλή ακρίβεια.

- Υβριδικό μοντέλο: Σε περίπτωση που επιλεγεί αυτή η περίπτωση, τότε είναι ένας συνδυασμός των δύο παραπάνω μοντέλων, κατά τον οποίο μπορούν να χρησιμοποιηθούν οι στατιστικές για το ευρύτερο μέρος του classification και στη συνέχεια να εφαρμοστούν επιπλέον κανόνες προκειμένου να βελτιωθούν οι περιπτώσεις κατά τις οποίες το στατιστικό μοντέλο δεν είναι τόσο ακριβές.

  Ο ορισμός των κανόνων δεν χρειάζεται να είναι πολύ περίπλοκος και χρειάζεται να καλύψει μόνο τις περιπτώσεις στις οποίες το στατιστικό μοντέλο δίνει ανακριβή αποτελέσματα. Περιλαμβάνει τα πλεονεκτήματα και των δύο παραπάνω μοντέλων, αλλά είναι δυσκολότερο και πιο χρονοβόρο να στηθεί, καθώς πρέπει πρώτα να

δημιουργηθεί το στατιστικό μοντέλο και ύστερα να μπορέσει να βελτιωθεί με τους κανόνες που θα οριστούν.

Πριν από τη δημιουργία οποιουδήποτε μοντέλου text classification, πρέπει να οριστούν βιβλιοθήκες λέξεων από τις οποίες θα προκύπτει ο εκάστοτε διαχωρισμός, καθώς και να οριστούν τα multiwords (συνδυασμοί λέξεων που εμφανίζονται μαζί όπως πχ «Τύπος Παραστατικού») για πραγματοποιηθεί tokenization των λέξεων, με το οποίο θα βοηθηθεί η γλωσσική ανάλυση.

Κεφάλαιο 5: Project Methodology

5.1 Μεθοδολογία Knowledge Discovery in Databases

Η KDD μεθοδολογία, είναι η διαδικασία ανεύρεσης χρήσιμων γνώσεων από μια συλλογή δεδομένων. Αυτή η ευρέως χρησιμοποιούμενη τεχνική εξόρυξης δεδομένων είναι μια διαδικασία που περιλαμβάνει προετοιμασία και επιλογή δεδομένων, καθαρισμό δεδομένων, ενσωμάτωση προηγούμενων γνώσεων σε σύνολα δεδομένων και ερμηνεία ακριβών λύσεων από τα παρατηρούμενα αποτελέσματα.

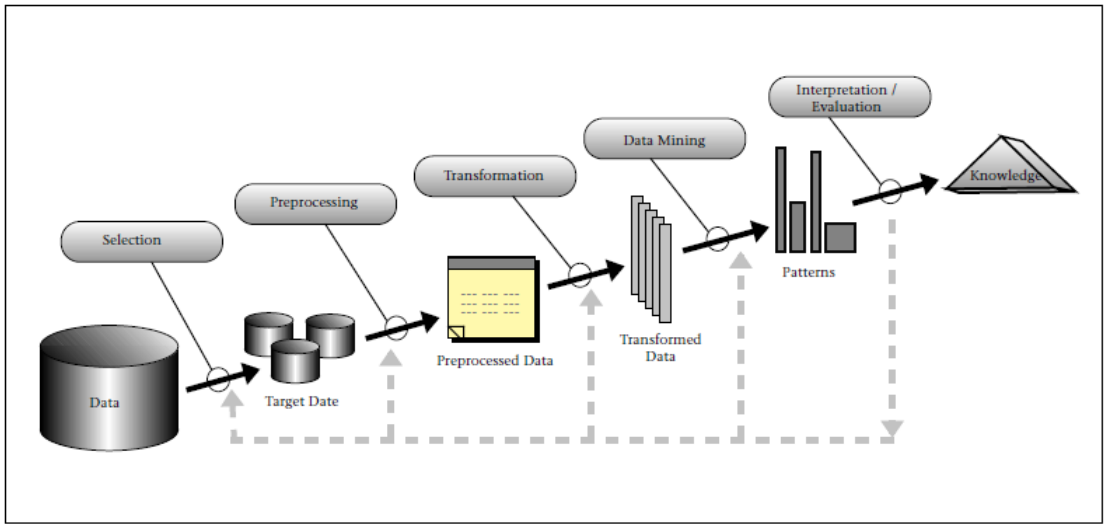
Το βασικό πρόβλημα που αντιμετωπίζει η διαδικασία KDD είναι η χαρτογράφηση δεδομένων χαμηλού επιπέδου (τα οποία είναι συνήθως πολύ ογκώδη για να ερμηνευτούν εύκολα) σε άλλες μορφές που θα μπορούσαν να είναι πιο συμπαγείς (για παράδειγμα, μια σύντομη αναφορά), πιο αφηρημένη όπως για παράδειγμα, μια περιγραφική προσέγγιση ή μοντέλο της διαδικασίας αυτής ή πιο χρήσιμη (για παράδειγμα ένα μοντέλο πρόβλεψης). Στον πυρήνα της διαδικασίας είναι η εφαρμογή συγκεκριμένων μεθόδων εξόρυξης δεδομένων για την ανακάλυψη και την εξαγωγή μοτίβων.

Η KDD έχει εξελιχθεί και συνεχίζει να εξελίσσεται, από τη διασταύρωση ερευνητικών πεδίων όπως η μηχανική μάθηση, η αναγνώριση προτύπων, οι βάσεις δεδομένων, στατιστικά στοιχεία, AI, οπτικοποίηση δεδομένων, και υψηλής απόδοσης υπολογιστές. Ο στόχος είναι η εξαγωγή γνώσεων υψηλού επιπέδου από χαμηλού επιπέδου δεδομένα στο πλαίσιο μεγάλων συνόλων δεδομένων.

Τα βασικά στάδια της διαδικασίας είναι τα εξής:

1. Κατανόηση του τομέα εφαρμογής και της γνώσης που πρέπει να προϋπάρχει σε συνδυασμό με τον ορισμό του στόχου.
2. Δημιουργία του data set πάνω στο οποίο θα εφαρμοστούν οι διαδικασίες.
3. Καθαρισμός δεδομένων και προεργασία πάνω σε αυτά (αφαίρεση ακραίων τιμών και θορύβου, τεχνικές χειρισμού ελλειπόντων πεδίων κλπ)/
4. Μείωση του όγκου των δεδομένων με την εύρεση των κυριότερων μεταβλητών που θα είναι χρήσιμες στη διαδικασία.
5. Επιλογή του τύπου data mining (classification ή clustering κλπ).
6. Επιλογή του βέλτιστου αλγορίθμου ανάλογα με τον τύπο του data mining και των μεταβλητών που έχουμε διαθέσιμες.

7. Εφαρμογή του data mining για την εύρεση μοτίβων και εξαγωγή αποτελεσμάτων.
8. Ερμηνεία των αποτελεσμάτων που θα εξαχθούν.
9. Χρήση της γνώσης που αποκτήθηκε από την εφαρμογή των παραπάνω.



(Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.)

5.2 Μοντέλο CRISP-DM (CRoss-Industry Standard Process for Data Mining)

Το CRISP-DM (CRoss-Industry Standard Process for Data Mining), είναι ένα μη-αποκλειστικό, τεκμηριωμένο και ελεύθερα διαθέσιμο μοντέλο εξόρυξης δεδομένων. Αναπτύχθηκε από τις βιομηχανίες με εισροές από περισσότερους από 200 χρήστες εξόρυξης δεδομένων και εργαλεία εξόρυξης δεδομένων και παρόχους υπηρεσιών. Αυτό το μοντέλο ενθαρρύνει τις βέλτιστες πρακτικές και προσφέρει στους οργανισμούς τη δομή για την επίτευξη καλύτερων, ταχύτερων αποτελεσμάτων από την εξόρυξη δεδομένων.

Πιο απλά, το CRISP-DM είναι μία πλήρης μεθοδολογία που παρέχει στον οποιονδήποτε – από αρχάριους στο χώρο μέχρι έμπειρους αναλυτές – ένα πλήρες σχέδιο για τη διεξαγωγή ενός data mining project. Το CRISP-DM αναλύει τον κύκλο ζωής ενός τέτοιου project σε έξι φάσεις, οι οποίες είναι οι: business understanding, data understanding, data preparation, modeling, evaluation και deployment

1) Business Understanding: Αποτελεί μία από τις κυριότερες φάσεις οποιουδήποτε data mining project. Βασίζεται στην κατανόηση των κύριων στόχων από την πλευρά του business, προκειμένου δημιουργηθεί ένα πλάνο για την επίτευξή τους και την καλύτερη κατανόηση των δεδομένων.

2) Data Understanding: Αυτή η φάση ξεκινάει με μία αρχική συλλογή δεδομένων. Στη συνέχεια, πρέπει να εξεταστεί η ποιότητά τους, να βρεθούν οι αρχικές πληροφορίες που προσφέρουν και να δημιουργηθούν αρχικές υποθέσεις.

3) Data Preparation: Σε αυτό το σημείο, τα δεδομένα που έχουν συλλεγεί, οργανώνονται προκειμένου να μπορέσουν να προετοιμαστούν για να τροφοδοτήσουν το μοντέλο που θα χρησιμοποιηθεί για την εξαγωγή αποτελεσμάτων. Αυτή η φάση περιλαμβάνει και την εύρεση των κατάλληλων μεταβλητών για χρήση.

4) Modeling: Ύστερα από την προετοιμασία των δεδομένων, εξετάζονται οι διάφορες τεχνικές μοντελοποίησης ανάλογα με το πρόβλημα που αντιμετωπίζεται. Συνήθως, για το ίδιο είδος προβλήματος υπάρχουν διαφορετικές τακτικές για την επίλυσή του, όπως για παράδειγμα το είδος του αλγορίθμου που μπορεί να χρησιμοποιηθεί σε ένα πρόβλημα κατηγοριοποίησης. Ως εκ τούτου, ενδέχεται να χρειαστεί επανεξέταση των δεδομένων, ανάλογα με το μοντέλο που θα επιλεγεί να χρησιμοποιηθεί ανά περίπτωση.

5) Evaluation: Σε αυτό το στάδιο, αξιολογείται το μοντέλο που έχει δημιουργηθεί από τον αναλυτή, προκειμένου να διερευνηθεί και να επιβεβαιωθεί η ορθή λειτουργία του. Σε αυτή την περίπτωση, το μοντέλο πρέπει να καλύπτει επαρκώς όλους τους business στόχους που έχουν οριστεί. Σε διαφορετική περίπτωση πρέπει να εξεταστούν περαιτέρω τα προηγούμενα βήματα και να δοκιμαστούν διαφορετικοί τρόποι αντιμετώπισης.

6) Deployment: Μόλις το μοντέλο είναι έτοιμο και καλύπτει επαρκώς όλους τους business στόχους, είναι έτοιμο για παραγωγική χρήση. Σε αυτή τη φάση, πρέπει να οργανωθεί και να παρουσιαστεί με τέτοιο τρόπο προκειμένου ο πελάτης να μπορεί να το χρησιμοποιήσει.

Ανάλογα με τις απαιτήσεις, η παρούσα φάση μπορεί να είναι εξαιρετικά απλή ή αντίστοιχα εξαιρετικά δύσκολη, ανάλογα με τους στόχους που έχουν επιλεγεί και τον τελικό χρήστη του μοντέλου. Σημαντικό βήμα στην παρούσα φάση, είναι και η πρόβλεψη του τρόπου παρακολούθησης και συντήρησης των αποτελεσμάτων,

ειδικά σε περίπτωση που το αποτέλεσμα που θα προκύπτει θα χρησιμοποιείται καθημερινά από τον τελικό αποδέκτη.

Σύμφωνα με τις μεθοδολογίες – μοντέλα που αναφέρονται παραπάνω, για την περίπτωσή του τρέχοντος project θα χρησιμοποιηθούν οι παρακάτω Data Mining τεχνικές:

- Περιγραφή Δεδομένων και Σύνοψή τους: Παρέχει μία συνοπτική περιγραφή των δεδομένων σε στοιχειώδη μορφή, προκειμένου δοθεί στους χρήστες μια γενική εικόνα των δεδομένων. Από την παραπάνω περιγραφή ενδέχεται να προκύψουν και οι στόχοι ή μέρος αυτών.

- Segmentation: Διαχωρίζει τα δεδομένα σε ενδιαφέρουσες και σημαντικές υποομάδες ή κατηγορίες που μοιράζονται κοινά χαρακτηριστικά. Ένας αναλυτής μπορεί να ταξινομήσει ορισμένες υποομάδες ως σχετικές για το επιχειρηματικό ερώτημα, με βάση τις προηγούμενες γνώσεις ή με βάση το αποτέλεσμα της περιγραφής και της σύνοψης των δεδομένων. Ωστόσο, υπάρχουν επίσης τεχνικές αυτόματης ομαδοποίησης που μπορούν να εντοπίσουν ανυποψίαστες και κρυφές δομές στα δεδομένα που επιτρέπουν το segmentation. Η τμηματοποίηση μπορεί να είναι ένας τύπος προβλήματος εξόρυξης δεδομένων όταν ο εντοπισμός των τομέων είναι ο κύριος σκοπός.

- Concept Descriptions: Αυτή η διαδικασία στοχεύει σε μία κατανοητή περιγραφή των κλάσεων. Ο σκοπός σε αυτή την περίπτωση δεν είναι να δημιουργήσουμε ένα μοντέλο με μεγάλη προβλεπτική ακρίβεια αλλά να λάβουμε γνώση. Το παραπάνω μπορεί να χρησιμοποιηθεί και για classification, προκειμένου να δημιουργηθούν ξεκάθαρες κλάσεις που θα ανταποκρίνονται στο σύνολο των περιπτώσεων που αντιμετωπίζονται, με τη δημιουργία κανόνων που θα προκύψουν από τη γνώση και εμπειρία πάνω στο αντικείμενο.

- Classification: Το classification προϋποθέτει ότι υπάρχει ένα σύνολο αντικειμένων που χαρακτηρίζονται από κάποια ιδιότητα ή χαρακτηριστικό που ανήκουν σε διαφορετικές κατηγορίες. Η τιμή της εκάστοτε κλάσης είναι μια διακριτή τιμή και είναι γνωστή για κάθε αντικείμενο. Ο στόχος είναι να δημιουργηθούν μοντέλα ταξινόμησης που διαχωρίζουν ορθά τα δεδομένα και κατηγοριοποιούν την κάθε νέα περίπτωση στην ορθή κλάση.

- Dependency Analysis: Αυτή η τεχνική, βρίσκει ένα μοντέλο που περιγράφει σημαντικές συσχετίσεις ή εξαρτήσεις μεταξύ των δεδομένων. Αυτές οι εξαρτήσεις, μπορούν να χρησιμοποιηθούν για να προβλεφθεί η τιμή ενός στοιχείου, δοθέντος της πληροφορίας άλλων στοιχείων.

  Μία ειδική περίπτωση εξαρτήσεων είναι τα Associations που περιγράφουν συγγένειες – σχέσεις μεταξύ των στοιχείων, όπως για παράδειγμα συμβάντα που συνδέονται και συμβαίνουν συχνά ταυτόχρονα.

Κεφάλαιο 6: Προτεινόμενη Λύση

Σύμφωνα με τα προηγούμενα κεφάλαια, θα δημιουργηθεί ένας αλγόριθμος ο οποίος θα είναι ικανός μόλις λάβει ως είσοδο ένα παραστατικό σε μορφή PDF, να προχωρήσει στην αυτόματη κατηγοριοποίηση και αναγνώρισή του, σύμφωνα με τα αποτελέσματα του μοντέλου που θα εκπαιδευτεί.

6.1 Προαπαιτούμενα - Παραδοχές

Τα κομβικά σημεία για το παραπάνω εγχείρημα είναι τα εξής:

1. Δημιουργία βιβλιοθηκών λέξεων και φράσεων για κάθε περιοχή του παραστατικού: Από τα δεδομένα που υπάρχουν διαθέσιμα και περιγράφονται στο κεφάλαιο 5, για κάθε πεδίο του παραστατικού υπάρχουν εκτός από το χώρο που βρίσκεται (Header – Details – Footer) και τα λεκτικά που το απαρτίζουν ή βρίσκονται γύρω από αυτό. Με το παραπάνω είναι εφικτό να δημιουργηθούν κλάσεις για την κατηγοριοποίηση των λέξεων ή φράσεων που θα εξαχθούν.

2. Δημιουργία βιβλιοθηκών με τα regular expressions που χρησιμοποιούνται για την εξαγωγή των δεδομένων: Αυτές οι βιβλιοθήκες χρειάζονται για την εξαγωγή των πληροφοριών για κάθε πεδίο του παραστατικού, μόλις αυτό εντοπιστεί μέσα στο χώρο.

3. Δημιουργία Concept Descriptions: Τα παραπάνω θα βοηθήσουν για να περιγραφεί καλύτερα το data set που έχει προκύψει και να δημιουργηθούν κανόνες που θα οδηγήσουν σε classification.

4. Υπολογισμός των μετρικών: Σε αυτό το βήμα θα υπολογιστούν οι μετρικές που ενδέχεται να βοηθήσουν στην κατανόηση και περιγραφή του dataset, αλλά και στους πιθανούς διαχωρισμούς που θα προκύψουν (Association Rules, Mahalanobis Distance κλπ).

5. Επιλογή μεταβλητών: Μόλις υπάρξουν διαθέσιμες οι παραπάνω πληροφορίες, πρέπει να γίνει επιλογή των μεταβλητών που θα χρησιμοποιηθούν. Αυτό θα γίνει είτε μετρώντας τον αριθμό εμφάνισης κάθε μεταβλητής από το participation matrix, είτε

χρησιμοποιώντας τις παραπάνω μετρικές, είτε με συνδυασμό των παραπάνω μεθόδων και της διαδικασίας που αναλύεται στο Κεφάλαιο 5.

6. Επιλογή του κατάλληλου αλγορίθμου classification: Ύστερα από τα παραπάνω βήματα και με βάση τις μεταβλητές που έχουν επιλεγεί για να χρησιμοποιηθούν, θα δημιουργηθεί το κατάλληλο classification μοντέλο για το διαχωρισμό των κλάσεων των μηχανισμών αναγνώρισης. Οι δύο αλγόριθμοι που εμφανίζονται ως οι πλέον κατάλληλοι για την πραγματοποίηση του classification στην παρούσα περίπτωση είναι ο Naïve Bayes και το Decision Tree, με το 2^{ο} να υπερισχύει, καθώς λόγω της πολυπλοκότητας των δεδομένων και τη χρήση του τυχαίου δάσους αυξάνεται η ακρίβεια πρόβλεψης.

   Το παραπάνω μοντέλο δημιουργήθηκε για την αυτόματη κατηγοριοποίηση ενός παραστατικού όταν το λαμβάνουμε ως είσοδο, προκειμένου να μπορέσει να εντοπιστεί το κατάλληλο μοτίβο απεικόνισης πεδίων για την εξαγωγή των δεδομένων του μόλις αναγνωστεί και μετατραπεί στο σύνολό του σε κείμενο.

6.2 Overview of the technical solution - Tech Stack

Παρακάτω αναλύεται η δομή – τεχνικό υπόβαθρο με την οποία δημιουργήθηκε ο μηχανισμός.

Το Tech Stack, είναι ένας συνδυασμός γλωσσών προγραμματισμού, framework και εργαλείων που χρησιμοποιούν οι προγραμματιστές για να δημιουργήσουν μια διαδικτυακή ή κινητή εφαρμογή. Τα δύο βασικά στοιχεία κάθε εφαρμογής είναι από την πλευρά του πελάτη (front-end) και από την πλευρά του διακομιστή (back-end). Κάθε στρώμα της εφαρμογής είναι χτισμένο πάνω από το ένα κάτω, δημιουργώντας έτσι μια στοίβα.

Αναλύεται στο τμήμα του πελάτη (client-side) και στου διακομιστή (server-side). Η πλευρά του πελάτη είναι ό, τι βλέπουν οι χρήστες στις οθόνες τους και όπου συμβαίνει η αλληλεπίδραση. Η πλευρά του διακομιστή δεν είναι ορατή σε ένα χρήστη. Οι γλώσσες προγραμματισμού της, τροφοδοτούν την πλευρά του χρήστη και δημιουργούν τη λογική των ιστοσελίδων και εφαρμογών. Στην περίπτωση του συγκεκριμένου project, το tech stack έχει ως εξής:

- Client-side: HTML, CSS και JavaScript. Ο χρήστης θα έχει πρόσβαση στο δίκτυο WeDoConnect που παρέχεται από την εταιρία για να εντοπίζει και να λαμβάνει τα παραστατικά που τον αφορούν.

- Server-side: Microsoft SQL Server για την αποθήκευση και μεταφορά των δεδομένων, C# για τη δημιουργία των εργαλείων εξαγωγής πληροφορίας μέσα από τα PDF και του αλγορίθμου που θα διαχειρίζεται τα δεδομένα, Python για τη χρήση των Machine Learning αλγορίθμων και η βάση δεδομένων του δικτύου WeDoConnect για την περαιτέρω επεξεργασία των στοιχείων και την παροχή τους στους χρήστες της υπηρεσίας.

6.3 Συνολική λειτουργία αλγορίθμου

Το παραπάνω μοντέλο δημιουργήθηκε για την αυτόματη κατηγοριοποίηση ενός παραστατικού όταν το λαμβάνουμε ως είσοδο, προκειμένου να μπορέσει να εντοπιστεί το κατάλληλο μοτίβο απεικόνισης πεδίων για την εξαγωγή των δεδομένων του μόλις αναγνωστεί και μετατραπεί στο σύνολό του σε κείμενο.

Ως εκ των άνω, ο αλγόριθμος που δημιουργήθηκε λειτουργεί με βάση τα παρακάτω βήματα:

1. Το νέο παραστατικό εισέρχεται στο μηχανισμό. Με τεχνικές computer vision εξάγονται όλες οι πληροφορίες που περιέχει το παραστατικό, ανεξάρτητα από τη γραμματοσειρά ή τα layers που περιέχει το pdf και διαχωρίζεται σε γραμμές σειριακά, όπου στην κάθε μία αντιστοιχούν οι λέξεις/φράσεις που υπάρχουν σε αυτή.

2. Το μοντέλο που έχει εκπαιδευτεί για το text classification, κατατάσσει τα λεκτικά και κατ' επέκταση τις γραμμές που δημιουργήθηκαν παραπάνω στις περιοχές Κεφαλίδα – Περιγραφές – Σύνολα, προκειμένου να λάβει γνώση για τη χωροταξία των πεδίων και ως εκ τούτου να κατηγοριοποιήσει το παραστατικό σε μία ή περισσότερες από τις πιθανές κατηγορίες μορφής παραστατικών και τη μετέπειτα δοκιμή αναγνώρισής του, σύμφωνα με τους υπάρχοντες μηχανισμούς αναγνώρισης.

3. Σύμφωνα με την κατηγορία (ή κατηγορίες) που ενδέχεται να ανήκει το παραστατικό με βάση τον διαχωρισμό που έγινε στο βήμα 2, ο μηχανισμός αναζητεί τη γραμμή

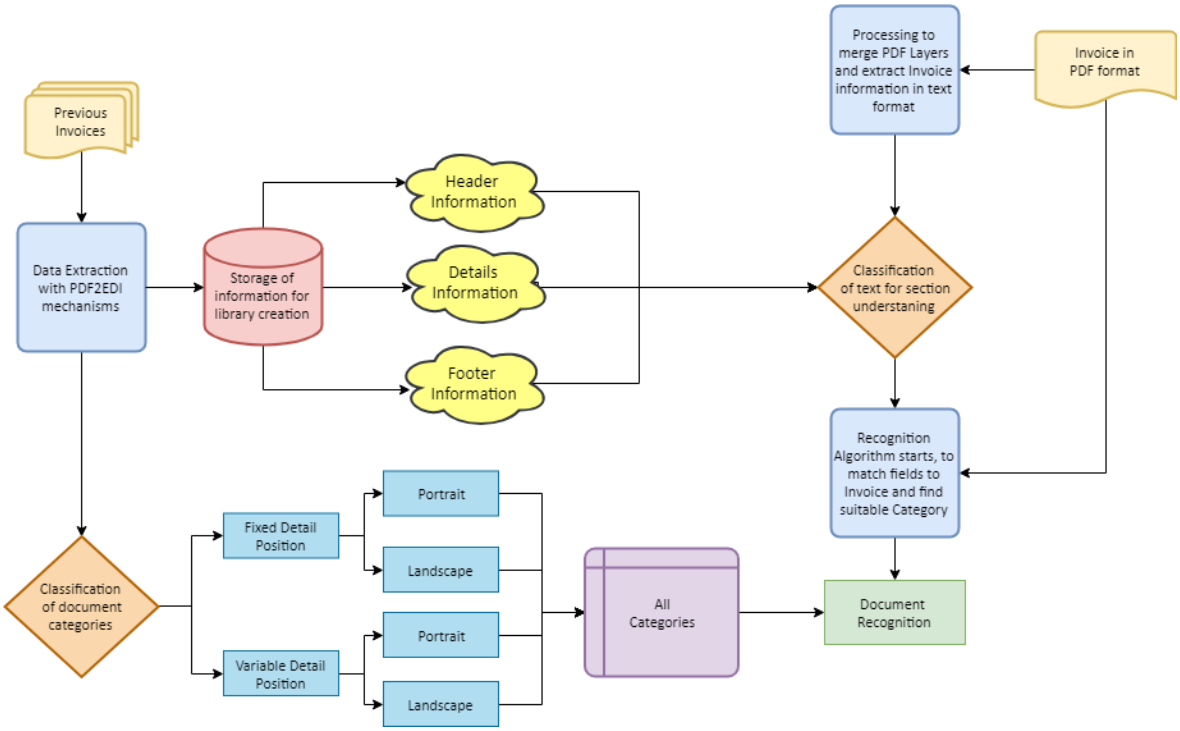
που εμφανίστηκε το λεκτικό που παραπέμπει στο πεδίο με τη μεγαλύτερη σημαντικότητα ή με το μεγαλύτερο ποσοστό εμφάνισης στους μηχανισμούς αναγνώρισης (πχ «Τύπος Παραστατικού») και ανάλογα με τη θέση του και την περιοχή που ανήκει, ξεκινά η διαδικασία αναγνώρισής του με τη χρήση των πιθανών regular expressions ανά περίπτωση. Αν το πεδίο αναγνωριστεί επιτυχώς, τότε με βάση το μοντέλο που έχει εκπαιδευτεί, ο μηχανισμός αναζητά το επόμενο πεδίο σύμφωνα με τις πιθανές κατηγορίες που ενδέχεται να ανήκει, αλλά και με τις μετρικές απόστασης που έχουν υπολογιστεί για τα πεδία μεταξύ τους. Αν δεν τα καταφέρει, τότε επεκτείνονται οι συντεταγμένες ελέγχου κατά ένα ποσοστό προκειμένου να καλυφθεί μεγαλύτερος χώρος για τον επανέλεγχο. Αν και αυτό αποτύχει, τότε δοκιμάζεται η επόμενη πιθανή περίπτωση κλάσης.

Τα παραπάνω, υπόκεινται στην περίπτωση που το πεδίο ανήκει είτε στην περιοχή της Κεφαλίδας του παραστατικού ή των Συνόλων. Όπως είναι φυσιολογικό, η περιοχή των Περιγραφών που περιέχει τις αναλύσεις προϊόντων αρχίζει στο πέρας της Κεφαλίδας και τελειώνει στην έναρξη της περιοχής των Συνόλων του παραστατικού. Για την περίπτωση των Περιγραφών, ο μηχανισμός λειτουργεί αντίστοιχα, με τη διαφορά πως το ύψος των πεδίων είναι σταθερό σε όλες τις περιπτώσεις.

4. Η παραπάνω διαδικασία επαναλαμβάνεται μέχρι να αναγνωριστεί το σύνολο των πληροφοριών του παραστατικού και γίνει ο αντίστοιχος συστημικός έλεγχος για την ορθότητα των αποτελεσμάτων. Σε αντίθετη περίπτωση, η μορφή του παραστατικού ή θεωρείται ως νέα περίπτωση με αποτέλεσμα την επανεκπαίδευση του μοντέλου ή τη βελτίωσή του.

Τα αποτελέσματα του αλγορίθμου θα εξεταστούν αρχικά με την εμπειρία χρήσης αλλά και με μετρικές αξιολόγησης classification αλγορίθμων όπως το Confusion Matrix και οι υπόλοιπες μετρικές που αναφέρονται σε προηγούμενα κεφάλαια.

Μία γραφική απεικόνιση του αλγορίθμου είναι η παρακάτω:


Previous Invoices
Data Extraction with PDF2EDI mechanisms
Storage of information for library creation
Header Information
Details Information
Footer Information
Processing to merge PDF Layers and extract Invoice information in text format
Invoice in PDF format
Classification of text for section understaning
Recognition Algorithm starts, to match fields to Invoice and find suitable Category
Document Recognition
Classification of document categories
Fixed Detail Position
Variable Detail Position
Portrait
Landscape
Portrait
Landscape
All Categories

Κεφάλαιο 7: Συμπεράσματα και Μελλοντική Δράση

Με την ορθή λειτουργία του αλγορίθμου που περιγράφεται στο Κεφάλαιο 6, επιτυγχάνεται η αυτόματη κατηγοριοποίηση και αναγνώριση των παραστατικών που αποσκοπεί στην μείωση του χρόνου επεξεργασίας αλλά και του κόστους διαχείρισης των εγγράφων.

Η προσέγγιση που παρουσιάστηκε σε αυτή την εργασία προετοιμάζει το έδαφος για περαιτέρω βελτιώσεις και επεκτάσεις στην επίλυση του προβλήματος που εξετάζεται. Μια επέκταση που θα μπορούσε να εξεταστεί, είναι ο συνδυασμός των Τυχαίων Δασών με μια πιο εξελιγμένη μέθοδο επιλογής χαρακτηριστικών, όπως ο αλγόριθμος Correlation Feature Selection που λαμβάνει υπόψη τις εξαρτήσεις των χαρακτηριστικών για να επιλέξει τα καταλληλότερα για χρήση από έναν ταξινομητή.

Επίσης, θα μπορούσε να εξεταστεί η δυνατότητα κατασκευής αλγορίθμου ομαδοποίησης (clustering) ενός συνόλου δεδομένων σε προκαθορισμένο αριθμό συστάδων ώστε να ομαδοποιεί τα δείγματα σε k συστάδες (clusters) βάσει των χαρακτηριστικών του κάθε δείγματος, έτσι ώστε τα clusters να είναι όμοια εσωτερικά και ανόμοια εξωτερικά.

Βιβλιογραφία

[1] Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI magazine, 17(3), 37.

[2] Colin Shearer (2000). The CRISP-DM Model: The New Blueprint for Data Mining. JOURNAL OF DATA WAREHOUSING, 5(3), 13.

[3] Isabelle Guyon, Andre Elisseeff (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research 3,* 1157-1182.

[4] Jennifer G. Dy, Carla E. Brodley (2004). Feature Selection for Unsupervised Learning. Journal of Machine Learning Research 5, 845–889.

[5] Yiming Yang, Jan O. Pedersen. A Comparative Study on Feature Selection in Text Categorization, USA.

[6] Sotiris Kotsiantis, Dimitris Kanellopoulos. Association Rules Mining: A Recent Overview (2006). GESTS International Transactions on Computer Science and Engineering. Vol.32 (1), 2006, pp. 71-82

[7] Quinlan, J. R. 1986. Induction of Decision Trees. Mach. Learn. 1, 1 (Mar. 1986), 81-106.

[8] Andrew McCallum, Kamal Nigam. A Comparison of Event Models for Naive Bayes Text Classification, Pittsburgh.

[9] Mehul Ved. (2018, July 20). Feature Selection in Machine Learning: Variable Ranking and Feature Subset Selection Methods. Retrieved from https://medium.com.

[10] Mahika Bansal. (2017, Mar 29). Association Rule Mining. Deciphered. Retrieved from https://medium.com.

[11] Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16–28.

[12] Battiti, R. (1994). Using mutual information for selecting features in supervised neural net learning. IEEE Transactions on Neural Networks, 5(4), 537–550

[13] Sunil R. (2015, Sept 11). 6 Easy Steps to Learn Naive Bayes Algorithm. Retrieved from https://www.analyticsvidhya.com.

[13] Sunil R. (2015, Jan 15). Decision Tree – Simplified. Retrieved from https://www.analyticsvidhya.com.

[14] Niklas Donges (2018, Feb 22). The Random Forest Algorithm. Retrieved from https://towardsdatascience.com.

[15] Synced (AI Technology & Industry Review). (2017, Oct 24). How Random Forest Algorithm Works in Machine Learning. Retrieved from https://medium.com.

[16] Sunil R. (2017, Sept 13). Understanding Support Vector Machine algorithm. Retrieved from https://www.analyticsvidhya.com.

[17] Ahmed Kachkach (2018, Jan 10). Problem-solving with ML: automatic document classification. Retrieved from https://cloud.google.com.

[18] Stephanie (2017, Nov 21). Mahalanobis Distance: Simple Definition, Examples. Retrieved from https://www.statisticshowto.datasciencecentral.com.