Τίτλος Εργασίας
Εφαρμογή τεχνικών αναγνώρισης ρόλων μηχανικών λογισμικού με χρήση τεχνικών εξόρυξης γνώσης από αποθετήρια ανοιχτού λογισμικού

Μπεκιάρης Θεοφάνης ΑΕΜ 8200
ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ
ΤΜΗΜΑ ΗΛΕΚΤΡΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ & ΜΗΧΑΝΙΚΩΝ ΥΠΟΛΟΓΙΣΤΩΝ
ΤΟΜΕΑΣ ΗΛΕΚΤΡΟΝΙΚΗΣ & ΥΠΟΛΟΓΙΣΤΩΝ

Emai author: bekiaris95@gmail.com ή theompek@ee.auth.gr

Abstract

Κατά την ανάπτυξη λογισμικού υπάρχουν παραδοσιακές προσεγγίσεις οι οποίες εστιάζουν στην εκτέλεση κάποιας ακολουθίας βημάτων όπως το μοντέλο Agile, το οποίο τονίζει τη σημασία της ευελιξίας και της διαρκούς, άμεσης επικοινωνίας μεταξύ των μελών της ομάδας ανάπτυξης. Λόγω της σημαντικότητας και του τρόπου της επικοινωνίας που περιγράφεται από αυτά τα μοντέλα, είναι επιθυμητά άτομα που διαθέτουν τόσο τεχνικές, όσο και επικοινωνιακές δεξιότητες. Ωστόσο, το πρόβλημα που ανακύπτει αναζητώντας τέτοια άτομα είναι η δυσκολία της αξιολόγησης αυτών των δεξιοτήτων, καθώς δεν μπορούν να αποτελέσουν αντικείμενο κάποιας συνηθισμένης αξιολόγησης.

Keywords—mining software repositories; change metrics;

I. Introduction

Σκοπός της εργασίας είναι η ανάπτυξη μιας μεθοδολογίας η οποία εφαρμόζει τεχνικές εξόρυξης γνώσης για την αναγνώριση των διαφορετικών ρόλων που μπορεί να έχει ένας μηχανικός λογισμικού σε έργα ανάπτυξης λογισμικού. Πιο συγκεκριμένα, το βασικό ζητούμενο του προβλήματος προς επίλυση αποτελεί η κατασκευή ενός συστήματος αναγνώρισης των ρόλων που αναλαμβάνουν οι μηχανικοί λογισμικού βασισμένοι σε μια σειρά από χαρακτηριστικά τα οποία μετρήθηκαν από τη δραστηριότητά τους στο GitHub.

II. Research Overview

Το πρόβλημα που καλούμαστε να επιλύσουμε απαιτεί την κατασκευή ενός μοντέλου αναγνώρισης ρόλων με χρήση δεδομένων από προηγούμενα έργα λογισμικού που περιέχουν πληροφορία σχετικά με τις δραστηριότητες των μηχανικών κατά την ανάπτυξη του εν λόγω έργου.
Όσον αφορά την επιλογή των repositories, αυτή έγινε με βάση τη δημοφιλία τους, όπως αυτή αποτυπώνεται στον αριθμό των stars, καθώς και από το μέγεθός τους.
Με χρήση των παραπάνω δεδομένων θα πρέπει να κατασκευαστεί ένα μοντέλο για την αναγνώριση των εξής 3 ρόλων μηχανικού:

- dev αναφέρεται στους «καθαρούς» developers,
- ops στα operations
- devops σε μηχανικούς που επιτελούν και τους δύο ρόλους

Τα ερωτήματα που καλούμαστε να απαντήσουμε κατά την διαδικασία κατασκευής του μοντέλου είναι:

(1) Ποια από τα χαρακτηριστικά των δεδομένων περιέχουν πληροφορία για την εξαγωγή των ρόλων;

(2) Μπορεί οποιοσδήποτε συνδυασμός χαρακτηριστικών να περιέχει χρήσιμη πληροφορία;

(3) Περιέχουν όλα τα repositories αξιόπιστα δεδομένα;

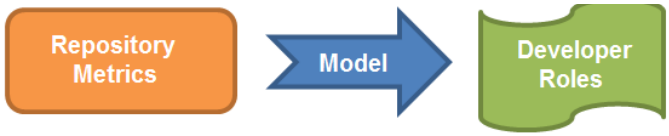
(4) Μπορούμε να χρησιμοποιήσουμε όλα τα repositories ταυτόχρονα στην κατασκευή του μοντέλου ή υπάρχει ανεξαρτησία μεταξύ των διαφορετικών έργων λογισμικού;

(5) Με ποιον τρόπο θα εξάγουμε από τα χαρακτηριστικά τις διαφορετικές ομάδες μηχανικών;

(6) Το μοντέλο που κατασκευάσαμε είναι αξιόπιστο και με ποιον τρόπο θα γίνει η αξιολόγηση του;

III. System Design

A. System Overview



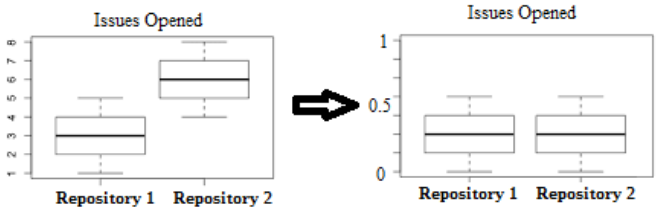
Η δομή του συστήματος ακολουθεί την μορφή του παρακάτω σχήματος. Από ένα σύνολο δεδομένων που περιέχουν πληροφορίες για την δραστηριότητά των μηχανικών κατά την ανάπτυξη ενός έργο λογισμικού θα πρέπει να κατασκευαστεί ένα μοντέλο το οποίο θα χρησιμοποιεί κάποια από αυτά τα χαρακτηριστικά-δεδομένα για να κάνει εκτίμηση των ρόλων που έχουν οι μηχανικοί πάνω στο έργο.

B. Data Preprocessing

Ο τρόπος με τον οποίο έγινε διαχείριση των δεδομένων των repositories προσπαθεί να επιλύσει τα προβλήματα των ερωτημάτων (3) και (4). Αρχικά πρέπει να αναφερθεί ότι τα δεδομένα από διαφορετικά repositories δεν μπορούν απευθείας να χρησιμοποιηθούν σαν ένα σύνολο καθώς κάθε repository αναφέρεται σε διαφορετικό έργο λογισμικού με αποτέλεσμα τα δεδομένα από διαφορετικά repositories να είναι ανεξάρτητα μεταξύ τους ως προς την απόλυτη τιμή τους. Αυτό που μας ενδιαφέρει δεν είναι η απόλυτη τιμή τους αλλά η εξάπλωση τους γύρο από το κέντρο τους. Για παράδειγμα το χαρακτηριστικό Issues Opened για κάποιο έργο λογισμικού μπορεί να εξαπλώνεται γύρο από μία μικρή μέση τιμή για ένα μικρό έργο ενώ για ένα μεγαλύτερο να έχουμε μεγαλύτερη μέση τιμή, παρόλα αυτά μέσα και στα δύο σύνολα δεδομένων ένας dev θα έχει μικρή

συνεισφορά(μικρότερη από την μέση τιμή του set δεδομένων) ενώ ένας ops θα έχει μεγαλύτερη συνεισφορά(μεγαλύτερη από την μέση τιμή του set δεδομένων).

Επιπλέον κάποιοι από τους μηχανικούς θα εμφανίζουν αποκλίνουσα ή ακραία συμπεριφορά ως προς κάποιο χαρακτηριστικό σε σχέση με το σύνολο μηχανικών που έχουν των ίδιο ρόλο στο έργο, αυτές οι τιμές ονομάζονται εξωκείμενες τιμές και εισάγουν "θόρυβο" στα δεδομένα και θα πρέπει να τις αφαιρέσουμε από το δεδομένα.



Εικόνα 1 : Παράδειγμα: Δεδομένα στα repositories πριν και μετά την προ επεξεργασία

Για την επίλυση των παραπάνω προβλημάτων ακολουθούμε την εξής προ επεξεργασία στα δεδομένα:

- Απομονώνουμε το κάθε repository αφού όπως είπαμε το καθένα αναφέρεται σε διαφορετικό έργο αλλά με παρόμοια χαρακτηριστικά(διασπορά τιμών).
- Αφαιρούμε τις εξωκείμενες-αρκαίες τιμές από κάθε χαρακτηριστικό για κάθε repository.
- Κανονικοποιούμε κάθε ένα από τα χαρακτηριστικά των repositories στο διάστημα [0,1]
- Στην συνέχεια ενώνουμε ξανά τα repositories σαν ένα σύνολο δεδομένων.
- Τελικά για την κατασκευή του μοντέλου μας κρατάμε μόνο τα 10 πρώτα σε αριθμό δεδομένων repositories καθώς περιέχουν αρκετό αριθμό δεδομένων για την εξαγωγή συμπερασμάτων και επίσης μειώνουμε την πολυπλοκότητα των αλγορίθμων.
- Τα υπόλοιπα repositories θα τα χρησιμοποιήσουμε για την επαλήθευση των μοντέλων, δηλαδή σαν δεδομένα ελέγχου του μοντέλου που θα κατασκευάσουμε.

C. Model Construction

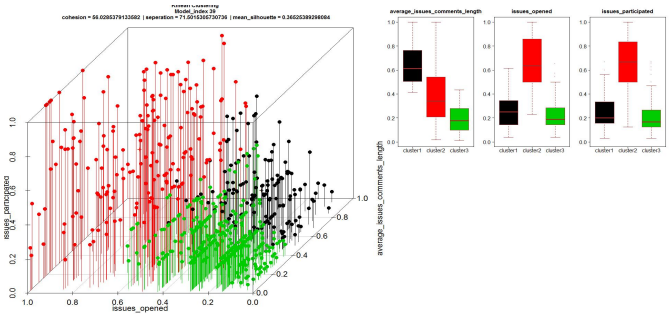
Στην συνέχεια θα προσπαθήσουμε να απαντήσουμε στα ερευνητικά ερωτήματα (1),(2),(5). Θα κατασκευάσουμε μοντέλα τα οποία με χρήση των δεδομένων θα μπορούν αν εκτιμήσουν τον ρόλο των μηχανικών. Όπως αναφέρθηκε έχουμε 3 βασικές ομάδες μηχανικών(dev,ops,devops). Για να εκτιμήσουμε λοιπόν αν τα χαρακτηριστικά των δεδομένων μπορούν να δώσουν πληροφορία για τον ρόλο του κάθε μηχανικού θα ακολουθήσουμε την εξής μεθοδολογία:

- Θα δημιουργήσουμε συνδυασμό χαρακτηριστικών ανά 3 για όλους τους δυνατούς συνδυασμούς. Η επιλογή των ανά 3 χαρακτηριστικών έγινε για μπορεί να γίνει δυνατή η οπτικοποίηση των αποτελεσμάτων σε 3D γράφημα.
- Στην συνέχεια χρησιμοποιήσουμε τεχνικές ομαδοποίησης για την αναγνώριση των 3 ρόλων μηχανικού. Συγκεκριμένα χρησιμοποιούμε τους αλγόριθμους Kmean και Hierarchical clustering complete linkage για την δημιουργία 3 ομάδων στο σύνολο των δεδομένων.
- Ο κάθε συνδυασμός χαρακτηριστικών αποτελεί και ένα διαφορετικό μοντέλο. Για να απαντήσουμε στο ερώτημα για το αν κάποιο χαρακτηριστικό περιέχει πληροφορία για κάποιον ρόλο, παρατηρώντας τα διαφορετικά μοντέλα και τις ομάδες που σχηματίζονται μέσα σε αυτά, προσπαθούμε να δούμε αν οι ομάδες αυτές έχουν κάποια λογική σύνδεση με τους ρόλους του μηχανικού.
- Αφού παρατηρήσουμε την συμπεριφορά των χαρακτηριστικών ανά ρόλο μηχανικού(clusters) τότε κατασκευάζουμε ένα τελικό μοντέλο χρησιμοποιώντας όλα τα χαρακτηριστικά.
- Η αξιολόγηση του μοντέλου θα γίνει στην επόμενη ενότητα.

Μοντέλα με χρήση του Kmean

Αφού προ επεξεργαστούμε τα δεδομένα δημιουργούμε όλους τους συνδυασμούς τριάδων δεδομένων και χρησιμοποιούμε τον Kmean για την δημιουργία 3 ομάδων μέσα στα δεδομένα και στην συνέχεια κατασκευάζουμε τα γραφήματα των συνδυασμών στα οποία παρατηρούμε τις ομάδες που κατασκευάστηκαν και το νόημα αυτών των ομάδων για τους ρόλους του μηχανικού. Από τα διαγράμματα αυτά παρουσιάζονται ενδεικτικά 3 μοντέλα, δηλαδή 3 συνδυασμοί των τριών χαρακτηριστικών που μπορούν να δώσουν κάποια λογική πρόβλεψη για τους ρόλους, έτσι έχουμε:

Μοντέλο 1



Εικόνα 2: Μοντέλο 3 χαρακτηριστικών που δημιουργεί 3 ομάδες και στα δεξιά boxplot με τις τιμές των χαρακτηριστικών ανά ομάδα.

Στο παραπάνω μοντέλο φαίνονται 3 διαφορετικές ομάδες που σχηματίζονται από το συνδυασμό των χαρακτηριστικών issues_opened, average_issues_comments_length και issues_participated. Στα δεξιά έχουμε διαγράμματα boxplot τα οποία παρουσιάσουν τις τιμές των χαρακτηριστικών ανά ομάδα. Για παράδειγμα στο πρώτο από τα αριστερά boxplot φαίνεται η διασπορά των τιμών του χαρακτηριστικού average_issues_comments_length για κάθε ομάδα σημείων, συγκεκριμένα η μαύρη ομάδα βλέπουμε ότι παίρνει τιμές περίπου στο διάστημα [0.5 , 1] ενώ η πράσινη στο διάστημα [0 , 0.4]. Με αυτό τον τρόπο μπορούμε να παρατηρήσουμε καλύτερα τις τιμές που παίρνει η κάθε ομάδα ανά χαρακτηριστικό.

Πριν συνεχίσουμε με την παρουσίαση των μοντέλων θα πρέπει σε αυτό το σημείο ορίσουμε μία κλίμακα σχετικά το μέγεθος των χαρακτηριστικών που θα χρησιμοποιήσουμε στην περιγραφή των μοντέλων. Όλες οι τιμές των χαρακτηριστικών είναι κανονικοποιημένες στο διάστημα [0,1], έτσι θα θεωρούμε ότι ένα χαρακτηριστικό παίρνει μικρή ή μεγάλη τιμή σύμφωνα με την παρακάτω κλίμακα.

[0 , 0.2) : Very Low ή VL

[0.2 , 0.4) : Low ή L

[0 .4, 0.6) : Medium ή M

[0.6 , 0.8) : High ή H

[0.8 , 0.1] : Very High ή VH

Έχοντα ορίσει επομένως την παραπάνω κλίμακα συνεχίζουμε παρουσιάζοντας τα συμπεράσματα για το μοντέλο ρόλων σύμφωνα με το παραπάνω σχήμα της εικόνας 2.

Μια εκτίμηση για τον ρόλο των μηχανικών είναι:

average_issues_ comments_length	issues_opened	Issues_ participated	Εκτίμηση ρόλου
M or H	VL or L	VL or L	DEVOPS
L or M or H	M or H	M or H	OPS
VL or L	VL or L	VL or L	DEV

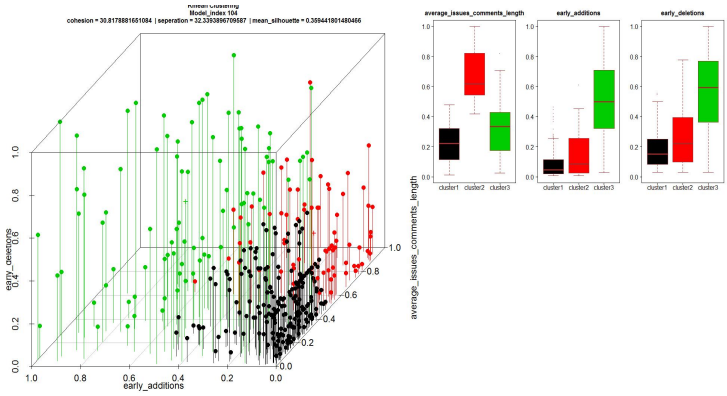
Πίνακας 1 : Εκτίμηση ρόλων, πρώτο μοντέλο

Τα χρώματα των κελιών αντιστοιχούν στα χρώματα των ομάδων της εικόνας 2 με αντικατάσταση του μαύρου με γκρι για να φαίνεται η γραμματοσειρά.

Το παραπάνω μοντέλο έχει λογική για την εκτίμηση των ρόλων καθώς όπως φαίνεται ένας dev μηχανικός θα έχει πολύ χαμηλή συνεισφορά στα issues, ένας ops μηχανικός θα έχει μεγάλη συνεισφορά στα issues ενώ ο devops όπως φαίνεται έχει μικρή συνεισφορά στο να ανοίξει ένα issues ή στην συμμετοχή σε issues αλλά έχει σχετικά αυξημένο αριθμό στο μέγεθος των comments length.

Στην συνέχεια παρουσιάζονται άλλα δυο μοντέλα με την ίδια λογική όπως προηγουμένως.

Μοντέλο 2

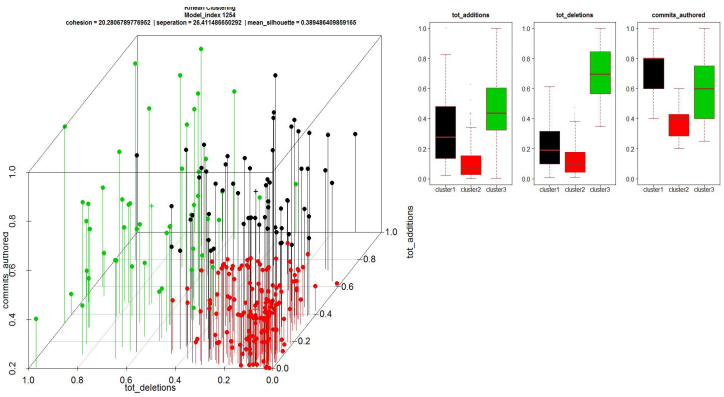


Εικόνα 3: Μοντέλο 3 χαρακτηριστικών που δημιουργεί 3 ομάδες και στα δεξιά boxplot με τις τιμές των χαρακτηριστικών ανά ομάδα.

average_issues_ comments_length	early_additions	early_deletions	Εκτίμηση ρόλου
VL or L	L	VL or L	DEV
M or H or VH	VL or L	VL or L or M	DEVOPS
L or M	M	L or M or H	OPS

Πίνακας 2 : Εκτίμηση ρόλων, δεύτερο μοντέλο

Μοντέλο 3



Εικόνα 4: Μοντέλο 3 χαρακτηριστικών που δημιουργεί 3 ομάδες και στα δεξιά boxplot με τις τιμές των χαρακτηριστικών ανά ομάδα.

tot_additions	tot_deletions	commits_authored	Εκτίμηση ρόλου
L or M	V or L	H	DEVOPS
VL	VL	L	DEV
M	M or H	M or H	OPS

Πίνακας 3 : Εκτίμηση ρόλων, τρίτο μοντέλο

Τα παραπάνω μοντέλα μπορούν να συνδυάσουν και να παράξουν ένα νέο μοντέλο πιο αυστηρό ως προς την εκτίμηση του για τον ρόλο του μηχανικού. Τα κοινά χαρακτηριστικά των παραπάνω μοντέλων θα πρέπει να συμφωνούν στο τελικό μοντέλο και επομένως στο τελικό μοντέλο θα συμπεριληφθούν μόνο τα πεδία(VL, L, M, H, VH) για το χαρακτηριστικό average_issues_ comments_length που συμφωνούν μεταξύ τους. Έτσι μπορούμε δημιουργήσουμε το παρακάτω μοντέλο:

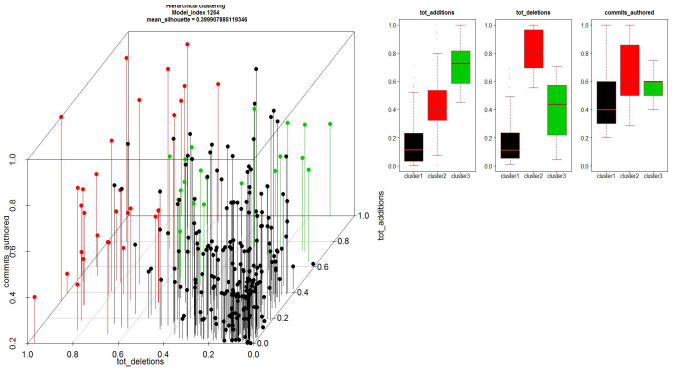
average_ issues_ commen ts_length	issues_ope ned	Issues_ particip ated	early_additio ns	early_deletio ns	tot_add itions	tot_dele tions	commit s_autho red	Εκτίμηση ρόλου
M or H	VL or L	VL or L	VL or L	VL or L or M	L or M	VL or L	H	DEVOPS
L or M	M or H	M or H	M	L or M or H	M	M or H	M or H	OPS
VL or L	VL or L	VL or L	L	VL or L	VL	VL	L	DEV

Πίνακας 4 : Αυστηρό μοντέλο εκτίμησης από τον συνδυασμό των προηγουμένων μοντέλων

Hierarchical clustering complete linkage

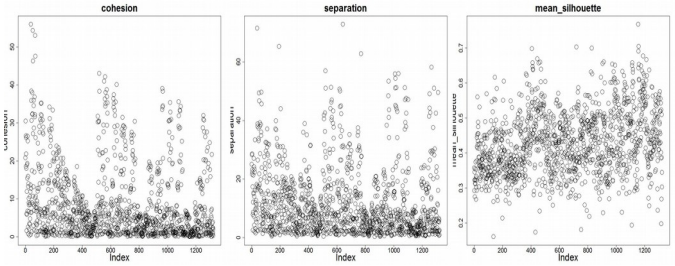
Τα αποτελέσματα από την ομαδοποίηση με ιεραρχική μέθοδο παράγουν παρόμοια αποτελέσματα με αυτά από τον Kmeam και αυτό επιβεβαιώνει την ορθότητα των παραπάνω μοντέλων. Παρακάτω παρατίθενται ενδεικτικά το διαγράμματα του τρίτου μοντέλου.

Προσοχή στα χρώμα για την αντιστοιχία των ομάδων, φαίνεται ότι οι ομάδες που σχηματίζονται με τον μοντέλο 3 είναι οι ίδιες απλά με διαφορετικό χρώμα.



Εικόνα 5: Hierarchical clustering, παρόμοιο με το μοντέλο 3 της προηγούμενης ενότητας του Kmean με διαφορετικά χρώματα.

Επιπλέον παρακάτω παρατίθεται τα διαγράμματα για τις μετρικές cohesion ,separation και silhouette για όλους τους ανά 3 συνδυασμούς χαρακτηριστικών που έγιναν. Οι μετρικές δεν είναι κατάλληλες για την σύγκριση των μοντέλων αλλά μπορούν να δώσουν μια εικόνα για την συνεκτικότητα των ομάδων μου δημιουργούνται.



Εικόνα 6: Μετρικές cohesion,separation, mean silhouette

IV. Evaluation

A. Evaluation Methodology

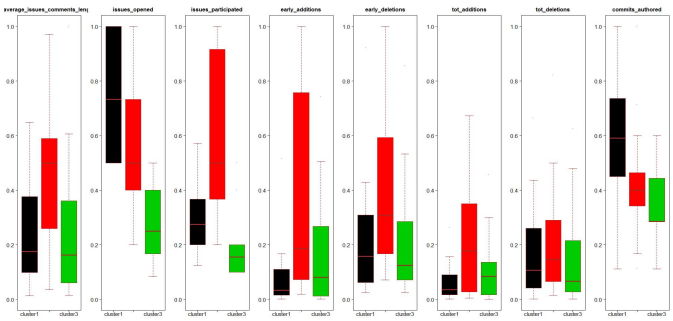
Σε αυτή την ενότητα θα προσπαθήσουμε να απαντήσουμε στο ερευνητικό ερώτημα (6). Για την αξιολόγηση του παραπάνω μοντέλου θα χρησιμοποιήσου τον συνδυασμό των παραπάνω χαρακτηριστικών σε ένα νέο σετ δεδομένων και θα παρατηρήσουμε τις τιμές των ομάδων που δημιουργούνται, αν σχηματιστούν ομάδες με παρόμοιες τιμές για τα χαρακτηριστικά που επιλέξαμε με αυτό το καινούργιο σετ δεδομένων τότε σημαίνει ότι το μοντέλο εκτίμησης των ρόλων είναι σωστό.

Για την κατασκευή των προηγούμενων μοντέλων χρησιμοποιήσαμε τα πρώτα 10 σε αριθμό δεδομένων repositories. Για την επαλήθευση του μοντέλου μας θα χρησιμοποιήσουμε αυτή την φορά τα 10 επόμενα σε σειρά repositories σαν δεδομένα ελέγχου. Θα τρέξουμε των Kmean για αυτά τα δεδομένα και για τον συνδυασμό αυτών των 8 χαρακτηριστικών του παραπάνω μοντέλου.

B. Evaluation Results

Kmean για το σετ δεδομένων ελέγχου

Εκτελούμε τον αλγόριθμο Kmean για το νέο σετ δεδομένων για το μοντέλο των 8 χαρακτηριστικών και παίρνουμε το παρακάτω γράφημα στο οποίο όπως και πριν φαίνεται η διασπορά των τιμών κάθε χαρακτηριστικού για κάθε ομάδα που σχηματίζεται.



Εικόνα 7 : Διασπορά τιμών κάθε ομάδας για κάθε ένα από τα χαρακτηριστικά

Με βάση το παραπάνω διάγραμμα κατασκευάζου τον παρακάτω πίνακα.

average_issues_comments_length	issues_opened	Issues_participated	early_additions	early_deletions	tot_additions	tot_deletions	commits_authored	Εκτίμηση ρόλου
L or M	H or VH	L or M	VL	VL or L	VL	VL or L	H	DEVOPS
L or M	M or H	M or H	L or M or H	L or M	VL or L	L	M	OPS
VL or L	VL or L	VL	VL	VL or L	VL	VL	L or M	DEV

Πίνακας 5: Μοντέλο εκτίμησης με χρήση του νέου σετ δεδομένων ελέγχου

Συγκρίνοντας τον πίνακα 5 με τον μοντέλο του πίνακα 4, φαίνεται ξεκάθαρα ότι τα δεδομένα από δυο διαφορετικά dataset (διαφορετικά repositories) δίνουν τις ίδιες τιμές για τα χαρακτηριστικά των μοντέλων. Επομένως η υπόθεση για το μοντέλο μας πρέπει να είναι πραγματική.

V. Conclusions

Στην ανάλυση που προηγήθηκε παρουσιάστηκε η διαδικασία κατασκευής ενός μοντέλου εκτίμησης ρόλων μηχανικού κατά την ανάπτυξη ενός έργου λογισμικού με χρήση κάποιον χαρακτηριστικών σχετικά με την δραστηριότητά των εν λόγω μηχανικών κατά την ανάπτυξη του έργου. Αρχικά παρουσιάστηκε η διαδικασία προ επεξεργασίας των δεδομένων και ο τρόπος με τον οποίο θα χρησιμοποιηθούν ώστε να περιέχουν αξιόπιστη πληροφορία. Στην συνέχεια χρησιμοποιήσαμε ένα υποσύνολο των δεδομένων για την κατασκευή των αρχικών μοντέλων μας. Στην συνέχεια συνδυάσαμε τα μοντέλα για την εξαγωγή ενός τελικού πιο αυστηρού μοντέλου. Τέλος χρησιμοποιήσαμε ένα άλλο υποσύνολο δεδομένων για την επαλήθευση της αξιοπιστίας του μοντέλου μας. Τα τελικά μοντέλα από τα 2 διαφορετικά υπο-σετ δεδομένων τελικά συμφωνούν και επομένως φαίνεται ότι η υπόθεση για το μοντέλο που κατασκευάστηκε είναι αληθής.

References

[1] IEEE Manuscript Templates for Conference Proceedings, online: http://www.ieee.org/conferences_events/conferences/publishing/templates.

[2] I. Zafeiriou, "Software engineer profile recognition through application of data mining techniques on GitHub repository source code and comments," *Diploma thesis*, Aristotle University of Thessaloniki, Thessaloniki, Greece, 2017.