Αναγνώριση Προτύπων

Εργασία 1

Ζαμπόκας Γιώργος 7173 (geoz222222@gmail.com)
Μουρούζης Χρίστος 7571 (chrimour@auth.gr)
Μποσδελεκίδης Βασίλειος 7488 (vmposdel@auth.gr)

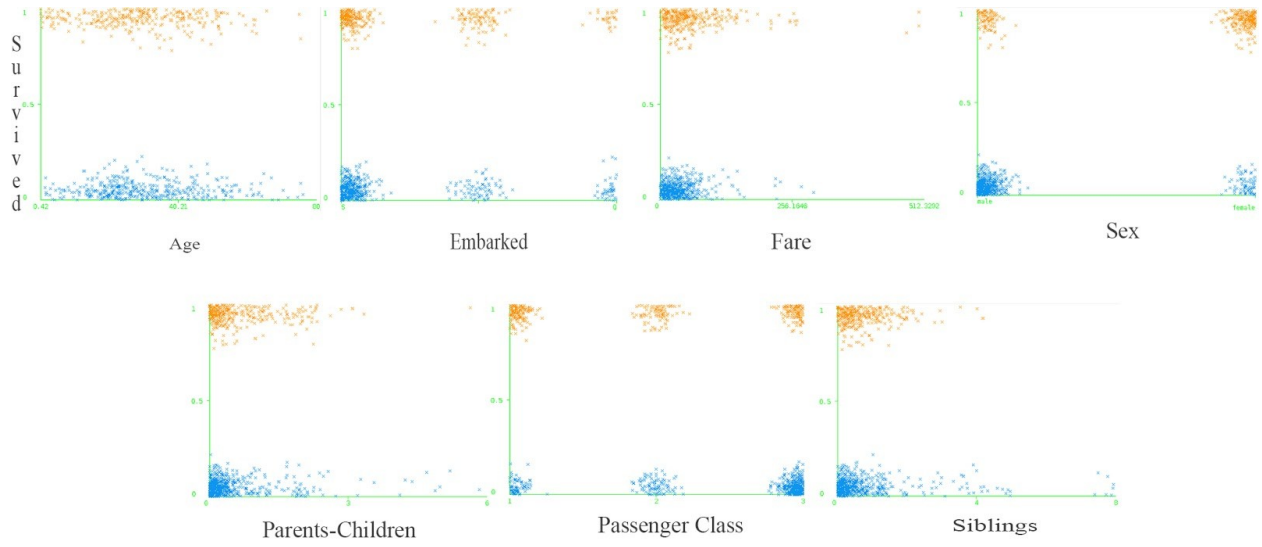
ΘΕΣΣΑΛΟΝΙΚΗ 2014

Εισαγωγή

Στην εργασία αυτή γίνεται η προσπάθεια να προβλεφθεί το αν κάποιος επιβάτης του τιτανικού επέζησε, γνωρίζοντας κάποια χαρακτηριστικά του. Τέτοια χαρακτηριστικά είναι η ηλικία, το φύλο, το κόστος του εισητηρίου του, η κλάση του επιβάτη, ο τόπος επιβίβασης του κ.α.

Πρώτη ματιά στα δεδομένα

Αναλύοντας το training dataset διαπιστώνονται κάποια χαρακτηριστικά που είναι περιττά για την ανάλυση (όνομα, κωδικός ειστηρίου, αριθμός καμπίνας) και κάποια άλλα ότι έχουν κενές εγγραφές (ηλικία, τόπος επιβίβασης).



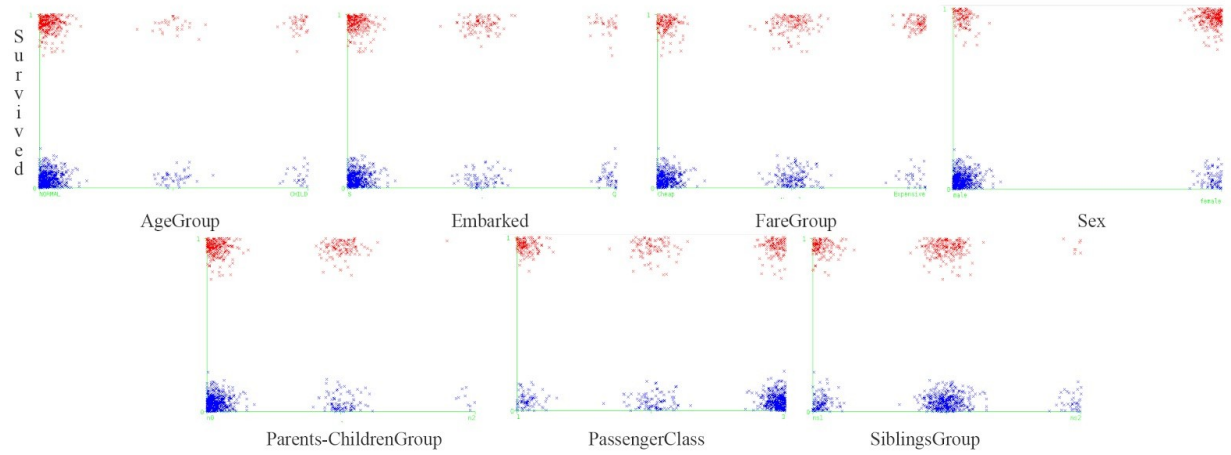
Σχήμα 1 - *Ανάλυση αρχικού dataset. Οι περιττές κλάσεις δεν απεικονίζονται.*

Προεπεξεργασία/Δημιουργία νέων κλάσεων

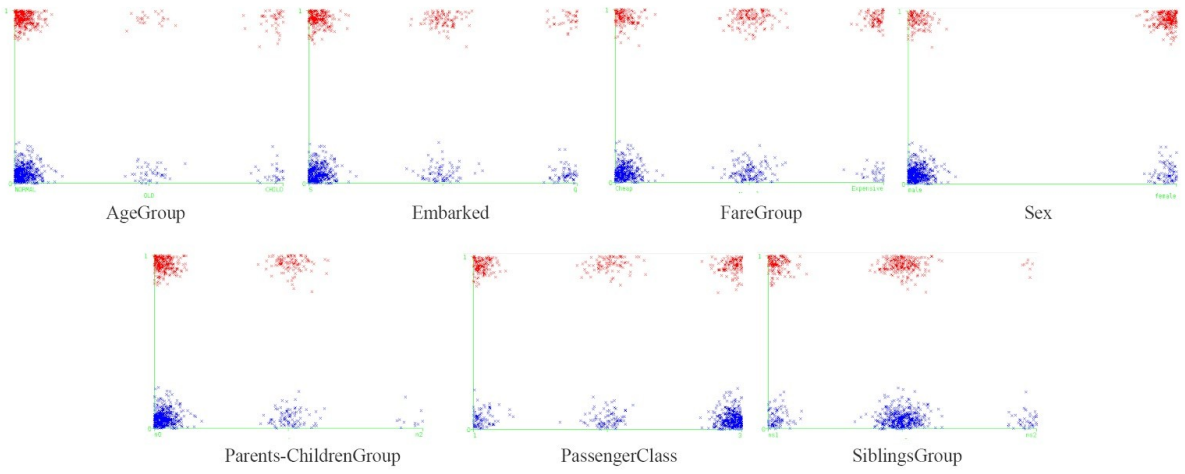
Αρχικά σβήσαμε τα περιττά για την ανάλυση δεδομένα και στη συνέχεια δημιουργήσαμε νέες κλάσεις για καλύτερη κατηγοριοποίηση των δεδομένων. Συγκεκριμένα η ηλικία χωρίστηκε σε 3 κατηγορίες (CHILD [0,10) , NORMAL [10,50) , OLD 50+), το κόστος εισιτηρίου σε επίσης 3 κατηγορίες (Cheap [0,19) , Normal [19,73) ,Expensive 73+), ο αριθμός των παιδιών/γονέων που ταξιδεύουν μαζί (n0 = μόνος , n1 = μέχρι και 3 παιδιά/γονείς , n2 = πάνω απο 3 παιδιά/γονείς) , ο αριθμός αδερφών που ταξιδεύουν μαζι (ns0 = κανένας αδερφός/ή , ns1 = μέχρι και 2 αδέρφια , ns2 = πάνω απο 2 αδέρφια).Στην παραπάνω κατηγοριοποίηση χρησιμοποιήθηκαν τα εξής φίλτρα στο WEKA έτσι ώστε να αποκτήσουν καλύτερη μορφή για την εκτέλεση των αλγορίθμων.

- replaceMissingValues: Αντικατάσταση κενών με το μέσο όρο των εγγραφών
- numericToNominal: Μετατροπή συνεχών attributes σε διακριτά
- resample: Αναδιανομή των δεδομένων για καλύτερη ισορροπία

Η διανομή των δεδομένων, πριν και μετά το resampling, φαίνεται στα παρακάτω σχήματα.



Σχήμα 2 - *Το dataset μετά την δημιουργία νέων κλάσεων και την κάλυψη των κενών attributes.*



Σχήμα 3 - *Το τελικό dataset, μετά και το resampling. Οι οριζόντιοι άξονες έχουν την εξής δομή: AgeGroup(NORMAL-OLD-CHILD), Embarked(S-C-Q), FareGroup(CHEAP-NORMAL-EXPENSIVE), Sex(male-female), Parch(n0-n1-n2), PClass(1-2-3) , Sibsp(ns1-ns0-ns2).*

Όπως γίνεται αντιληπτό ,υπάρχει μια σημαντικά καθαρότερη κατηγοριοποίηση από το αρχικό dataset. Για παράδειγμα οι ηλικιωμένοι επέζησαν σε μικρό ποσοστο, ενώ τα παιδιά σε μεγάλο, άνθρωποι με ακριβό εισητήριο είχαν μεγαλύτερο ποσοστό επιβίωσης, ένω με τα φθηνά ή τα κανονικά υπάρχει πολύ μεγάλη ή μεγάλη συγκέντρωση στην κλάση 0, οι άνδρες σε αντίθεση με τις γυναίκες έιχαν άσχημη κατάληξη σε συντριπτικά μεγαλύτερο ποσοστό(διαπίστωση που ισχύει ακόμα και από το αρχικό dataset, οι οικογένειες με πολλά μέλη, ή αν υπήρχαν πολλά αδέρφια πάνω στο πλοίο, κατηγοριοποιούνται κυρίως ως αποθανόντες(σε αντίθεση με άτομα που ταξίδευαν μόνα. Παρόμοιες παρατηρήσεις ισχύουν και για την κλάση του επιβάτη (όσο καλύτερη τόσο περισσότεροι επιζόντες) ,ενώ πολύ μικρότερη αλλά υπαρκτή καθαρότητα υπάρχει στην ιδιότητα Embarked.

Εκτέλεση Αλγορίθμων

Με τη βοήθεια του WEKA experimenter βγάλαμε τους καλύτερους αλγόριθμους ταξινόμησης με κριτήριο το prediction accuracy (Πίνακας 1) .

Ως κριτήριο επίδοσης χρησιμοποιήσαμε την πιο κάτω σχέση δίνοντας ίσα βάρη σε όλες τις μετρικές:

Score= 0.25*Accuracy + 0.25*Precision + 0.25*Recall + 0.25*F-Measure

Τελικά ξεχώρισαν τρεις κατηγορίες και από κάθε κατηγορία επιλέχθηκαν οι δύο καλύτεροι όπως φαίνεται πιο κάτω:

Trees
- LMT
- PART

SVM
- SPegasos
- SMO

NN
- KStar
- IBk

ΠΙΝΑΚΑΣ 1 - Σύγκριση datasets και αλγορίθμων

Οικογένεια Αλγορίθμου	Αλγόριθμος	Αρχικό Score	Score Before Resample	Final Accuracy	Final Precision	Final Recall	Final F-Measure	**Final Score**
Trees	PART	78.44%	80.35%	0.8515	0.85	0.851	0.849	**85.04%**
	LMT	----------	81.00%	0.8416	0.84	0.842	0.839	**84.07%**
	Random Tree	71.61%	81.95%	0.8373	0.842	0.837	0.833	83.73%
SVM	SPegasos	82.05%	81.50%	0.835	0.833	0.835	0.833	83.40%
	SMO	82.65%	81.50%	0.835	0.833	0.835	0.833	83.40%
	LIBSVM	52.90%	81.73%	0.8382	0.837	0.838	0.837	**83.76%**
NN	IB1	75.00%	76.25%	0.787	0.791	0.787	0.788	78.83%
	IBk	77.18%	81.75%	0.8373	0.842	0.837	0.833	83.73%
	KStar	76.80%	81.78%	0.8462	0.85	0.846	0.843	**84.63%**
Bayesian	AODE	79.25%	80.70%	0.8092	0.808	0.809	0.808	80.86%
	WAODE	----------	80.45%	0.8047	0.803	0.805	0.802	80.37%
N. Networks	M.Perceptron	----------	80.58%	0.8284	0.826	0.828	0.825	82.69%
	V.Perceptron	79.45%	79.80%	0.7935	0.792	0.793	0.79	79.21%

Παρατηρήσεις: Στο αρχικό dataset οι αλγόριθμοι LMT και Multilayer Perceptron δεν έτρεξαν λόγω μεγάλου μεγέθους του dataset (σφάλμα στη μνήμη (heap size της java)). Επίσης ο αλγόριθμος WAODE δεν ήταν διαθέσιμος.

Προκύπτουν οι παρακάτω πίνακες σύγχυσης:

LMT

classified as -->	a	b
a=0	179	18
b=1	30	76

PART

classified as -->	a	b
a=0	181	16
b=1	29	77

SPegasos

classified as -->	a	b
a=0	176	21
b=1	29	77

SMO

classified as -->	a	b
a=0	176	21
b=1	29	77

KStar

classified as -->	a	b
a=0	507	35
b=1	102	247

IBK

classified as -->	a	b
a=0	507	35
b=1	110	239

Δίνοντας μεγαλύτερο βάρος σε αυτούς που επέζησαν

Παρατηρώντας τους πίνακες σύγχυσης, διαπιστώνουμε πως το ποσοστό αυτών που επέζησαν και κατατάχθηκαν λάθος είναι πολύ μεγαλύτερο σε σύγκριση με αυτών που δεν επέζησαν (αλλά και απόλυτα υπερβολικά μεγάλο). Αυτό επιδιώξαμε να το περιορίσουμε εκτελώντας κατηγοριοποίηση με τον “Cost Sensitive Classifier” της κατηγορίας των meta classifiers. Χρησιμοποιήσαμε τον παρακάτω πίνακα κόστους:

0	1
5	0

Συνεπώς, αυτοί που κατηγοριοποιούνται ως αποθανόντες ενώ επέζησαν έχουν μεγαλύτερη ποινή.

Στη συνέχεια συγκρίνοντας, παραθέτουμε τους δύο αλγορίθμους που πετυχαίνουν καλύτερη ικανοποίηση αυτής της συνθήκης, αλλά, επίσης ,δεν επιδεινώνουν ιδιαίτερα και το accuracy.

Αλγόριθμος:

- IBK
  Accuracy: 76.54%
  Πίνακας σύγχυσης:

classified as -->	a	b
a=0	366	176
b=1	33	316

- LMT
  Accuracy: 73.51%
  Πίνακας σύγχυσης:

classified as -->	a	b
a=0	343	199
b=1	37	312

Βλέπουμε ότι ο IBK προβλέπει σε πολύ μεγάλο βαθμό αυτούς που επέζησαν, διατηρώντας σχετικά ικανοποιητικό accuracy. Παρ' όλα αυτά, λόγω της σημαντικά μειωμένης ακρίβειας δεν μπορούμε να επιλέξουμε αυτό το "cost sensitive" μοντέλο σαν το καλύτερο.

Κατηγοριοποίηση των επιβατών του test set

Κάνοντας χρήση των καλύτερων μοντέλων μας μπορούμε να προβλέψουμε την απόληξη των επιβατών, όταν γνωρίζουμε μόνο τα γνωρίσματα τους. Αρχικά τα δεδομένα του test dataset μετατρέπονται στην ίδια μορφή με αυτά του training set (ίδιες στήλες και ίδιο σύνολο διακριτών τιμών). Τοποθετώντας ερωτηματικά στη θέση της κλάσης, μπορούμε να εφαρμόσουμε τα μοντέλα της διαδικασίας εκμάθησης πάνω σε αυτό το test set. Αυτό μας δίνει τις τελικές προβλέψεις για κάθε επιβάτη. Αυτό το αρχείο, μετά και τις απαραίτητες τροποποιήσεις στον τρόπο παρουσίασης των δεδομένων, αξιολογείται από την εφαρμογή Kaggle. Το μοντέλο με το οποίο προέκυψαν οι καλύτερες προβλέψεις ήταν το PART με training set 66% (δίνοντας σκορ 79.42%) και

το δεύτερο καλύτερο ήταν το LMT με cross validation (δινοντας σκορ περίπου κατά 0.4% μικρότερο) . Η ομάδα μας ονομάζεται “predictedOK” και κατατασσόμαστε μέχρι στιγμής στην θέση #557.

Συμπεράσματα

Ο καλύτερος αλγόριθμος στη διαδικασία εκμάθησης ήταν ο PART με δεύτερο τον KStar, όμως με πολλά False Negatives (negative -> class=1). Ωστόσω, παρόλο που ο PART δίνει τα καλύτερα αποτελέσματα και στη διαδικασία της πρόβλεψης (για το συγκεκριμένο test set στο οποίο δοκιμάστηκε), δεύτερος καλύτερος είναι ο LMT.