Ονοματεπώνυμο Α.Μ. Ονοματεπώνυμο Α.Μ.	Ψαλλίδας Φώτης 1115200600170 Πολυχρονίου Ορέστης 1115200600176
Μάθημα	Data Mining
Αντικείμενο Εργασίας	**Classifying**
Καθηγητής	κ. Γουνόπουλος Δημήτρης
Ακαδημαϊκό Εξάμηνο	2008-2009

ΕΘΝΙΚΟΝ & ΚΑΠΟΔΙΣΤΡΙΑΚΟΝ ΠΑΝΕΠΙΣΤΗΜΙΟΝ ΑΘΗΝΩΝ
NATIONAL & KAPODISTRIAN UNIVERSITY OF ATHENS

- «Παράμετροι σωστής χρήσης των αλγορίθμων ***Naïve Bayes J48*** και ***k-nn***»

	Flags						
Naive Bayes	Debug		useKernellEstimator		useSupervisedEstimator		
J48	binarySplits	debug	reducedErrorPruning	saveInstanceData	subtreeRaising	unpruned	useLaplace
k-nn	None						

	Parameters			
Naïve Bayes	None			
J48	confidenceFactor	minNumObj	numFolds	seed
K-nn	k	DataSet	TestData	

Αναφορικά με τον knn αναφερόμαστε στο αλγόριθμο που υλοποιήσαμε εμείς και όχι αυτόν του weka.Παρακάτω αναφερόμαστε και στις τιμές που πάρθηκαν για να τρέξουμε τους αλγορίθμους.

- « **Compare time and accuracy for algorithms Naïve Bayes J48 Knn**»

  - ***Naive Bayes values token***

    Flags
    - Debug = false
    - useKernellEstimator=false
    - useSuperviseDiscretzation=false

    Parameters
    - None

  - ***J48 values token***

    Flags
    - binarySplits=false
    - debug=false
    - reduceErrorPruning=false
    - saveInstanceData=false
    - subtreeRaising=true
    - unpruned=false
    - useLaplace=false

    Parameters
    - confidenceFactor=0.25
    - minNumObj=2
    - numFolds=3
    - seed=1

  - ***Knn values token***

    Flags
    - None

    Parameters
    - k=1-2000
    - DataSet=Dataset.arff
    - TestSet=Testset,arff

Οι τιμές που χρησιμοποιήθηκαν για τους Naive Bayes και J48 είναι oi default που δίνει το weka Καθώς οι by default τιμές των αλγορίθμων naïve bayes και J48 κρίθηκαν ικανοποιήσιμες και δεν βρέθηκαν κάποιες καλύτερες χρησιμοποιήθηκαν αυτές ως παράμετροι.

		Time	accuracy
Naïve Bayes		0.58 seconds	50.15%
J48		0.63 seconds	50%
knn	k=1	14.78 seconds	48.95%
	k=3	14.76 seconds	49.95%
	k=5	14.83 seconds	50.05%
	k=7	14.89 seconds	50%
	k=9	14.89 seconds	50%
	k=11	15.19 seconds	50.05%
	k=13	14.87 seconds	50%
	k>=15	average 16	50%

Τα ολοκληρωμένα output μπορείτε να τα βρείτε στα αρχείο results.pdf

- “Επιρροή στην ακρίβεια εξαιτίας της ανομοιογένειας του ***DATASET.***”

Το Dataset που έχει δοθεί για το learning των αλγορίθμων περιέχει 25000 εγγραφές εκ των οποίων οι 22364 εγγραφές με τιμή -1 και 2636 εγγραφές με τιμή 1.Η συγκεκριμένη ανισοκατανομή στις τιμές προκαλεί πρόβλημα στους 3 αλγορίθμους που χρησιμοποιούμε ως προς την ορθότητα των αποφάσεων που παίρνουν για την τιμή της κάθε εγγραφή του Trainset.

Ο αλγόριθμος Naive Bayes επηρεάζεται άμεσα αφού η πιθανότητα μια τυχαία εγγραφή από το Trainset είναι εξορισμού μεγαλύτερη για την κλάση -1 λόγω των μεγαλύτερου πλήθους εγγραφών που υπάρχουν για αυτό.
Ο Naive Bayes αλγόριθμος προτιμάτε όταν υπάρχουν Bayesian κατανομές και όχι σε sets με ανισοκατανομές..

Ο αλγόριθμος J48 αποτελεί έναν αλγόριθμο ο οποίο δημιουργεί ένα pruned tree Βάση του οποίου γίνεται classify η κάθε εγγραφή του training set.Ωστόσο κατά τη δημιουργία του δέντρου αυτού ο αλγόριθμος μην μπορώντας να ορίσει μηδαμινές διαφορές δεν μπορεί να καταλάβει και άμεσα τη διαφορά που υπάρχει μεταξύ των -1 και των 1 οπότε τα βλέπει ίδια σχεδόν όλα και αποφασίζει να παράγει μόνο ένα κόμβο | -1.Έτσι όποια εγγραφή και αν έρθει από το trainingset θα γίνει classified στην κλάση -1 με αποτέλεσμα το 50% accuracy .Κοινώς δηλαδή θερούμε πως φταίνε η μηδαμινές διαφορές στα πολλά attributes.
Έγιναν κάποιες προσπάθειες μειώνοντας το confidence factor ώστε να υπάρξει ένα πιο δραστικό pruning αλλά το πρόβλημα παραμένει.

Ο αλγόριθμος Knn επηρεάζεται και αυτός όχι τόσο λόγω της μεγάλης ανισοκατανομής των δύο κλάσεων αλλά από την πυκνότητα των -1 και 1 στο χώρο όπου τα -1 και 1 είναι δίπλα το ένα στο άλλο. Οπότε όταν πάμε να ελέγξουμε τους γείτονες του στοιχείου που θέλουμε να γίνει classified συνήθως βρίσκουμε κοντά μας πολλά -1 και λιγότερα 1 λόγω της ανισοκατανομής σε συνεργασία με την πυκνότητα των -1 και 1 στον ίδιο χώρο.
Πιθανές λύσεις για τα παραπάνω προβλήματα θα μπορούσαν να είναι είτε το κόψιμο του dataset έτσι ώστε να υπάρχει μια ομοιόμορφη κατανομή είτε η εύρεση attributes ενδεικτικών των κλάσεων.