Εξόρυξη και Προετοιμασία Δεδομένων

Μεταπτυχιακή απαλλακτική εργασία ΦΕΒ. 2023-24

Καθηγητής:
Μ. Φιλιππάκης

Καμπάσης Παναγιώτης
Πανεπιστήμιο Πειραιώς
Τμήμα Ψηφιακών Συστημάτων
Μεγάλα Δεδομένα και Αναλυτική
17 Φεβρουαρίου 2024

Εξόρυξη και Προετοιμασία Δεδομένων

Πάνος Καμπάσης
27/01/2024

Περιεχόμενα

Περίληψη—
Στο παρόν θα δοκιμάσουμε τεχνικές συσταδοποίησης (*clustering*), και κατηγοριοποίησης (*classification*), σε βάση δεδομένων με πολλαπλά χαρακτηριστικά. Ως επί το πλείστον χρησιμοποιούμε ως εργαλείο προεπεξεργασίας, ανάλυσης και εφαρμογής μοντέλων το *Weka*. Η ανάλυση αφορά στους τρόπου που επιτυγχάνεται η καλύτερη δυνατή πρακτική εξόρυξης δεδομένων από βάση δεδομένων με πολλαπλές διαστάσεις για την αναλυτική και διαυγή κατηγοριοποίηση των δεδομένων. Στοχεύουμε στην καλύτερη αλλά συνάμα αρκετά γενικευμένη και πρακτική ομαδοποίηση των δεδομένων.
Λέξεις κλειδιά:
Ανάλυση δεδομένων, Εξόρυξη δεδομένων, $classification$, $clustering$, $Weka$, $Random\ Forest$

I. Εισαγωγή

Η βάση δεδομένων μας περιέχει πολλαπλές μετρήσεις για διάφορα χαρακτηριστικά σε μανιτάρια. Από 8124 μανιτάρια έχουμε δύο κατηγορίες, $'p'$ ($poisonous$) και $'e'$ ($edible$). Υπάρχουν ακόμα 22 άλλα χαρακτηριστικά που αφορούν το πλήθος το μανιταριών που πληρούν τις προϋποθέσεις που περιγράφει το εκάστοτε χαρακτηριστικό. Για παράδειγμα μερικά χαρακτηριστικά είναι το σχήμα της κουκούλας του μανιταριού ($cap-shape$), μυρωδιά ($odor$), χρώμα του εσωτερικού της κουκούλας ($gill-color$), κ.α. .
Αρχικά αναζητούμε στην βάση για ελλειπούσες τιμές και προεπεξεργαζόμαστε τα δεδομένα. Ως στόχος η μείωση της διάστασης των 23 χαρακτηριστικών, παραλείποντας εκείνα που συμμετέχουν λιγότερο στην συνολική μεταβλητότητα των δεδομένων. Έπειτα θα εκπαιδεύσουμε μοντέλα μηχανικής μάθησης για να κατηγοριοποιήσουμε τα δεδομένα σε 2 κατηγορίες $'p'$ ($poisonous$) και $'e'$ ($edible$) και θα εξάγουμε μετρικές χρησιμοποιώντας τα $ground\ data$. Επιλέγοντας το πιο αποδοτικό μοντέλο θα προχωρήσουμε σε μεθόδους συσταδοποίησης για την οπτικοποίηση και καλύτερο διαχωρισμό τον μανιταριών.

	$class$	$cap-shap$	$ecap-surface$	$cap-colorbruises$	$odor$	$gill-attachment$	$gill-spacing$
0	p	x	s	n	t	p	f
1	e	x	s	y	t	a	f
2	e	b	s	w	t	l	f
3	p	x	y	w	t	p	f
4	e	x	s	g	f	n	f

[1] : $The\ dataset$

II. Η διαδικασία σε βήματα

- Αναζητούμε την βάση για την καλύτερη κατανόηση της, την δομή, την μορφή των χαρακτηριστικών και τι ακριβώς περιγράφουν.
- Εφαρμόζουμε τεχνικές προ-επεξεργασίας.
- Δοκιμάζουμε μοντέλα μηχανικής μάθησης με σκοπό την κατηγοριοποίηση. Χαρακτηριστικό-στόχος: $class$. 2 κατηγορίες.
- Καταγράφουμε τα ποσοστά επιτυχίας των κατηγοριοποιήσεων και επιλέγουμε το μοντέλο που αποδίδει καλύτερα.
- Στα μοντέλα που μπορούνε να εμφανίσουν το δέντρο απόφασης μελετάμε το σχεδιάγραμμα του δέντρου και εξετάζουμε τεχνικές κλαδέματος ($Pruning$)
- Μέσω του $Weka$ επιλέγουμε $Attributeselection$ όπου παρέχεται πληθώρα μεθόδων μείωσης διάστασης για να βρούμε τα χαρακτηριστικά που έχουν καθοριστικότερη σημασία στην κατηγοριοποίηση. Με άλλα λόγια επιδιώκουμε να πετύχουμε ίδια ή παραπλήσια αποτελέσματα με πολύ μικρότερο πλήθος χαρακτηριστικών και πιο απλές μορφές σχεδιαγραμμάτων δένδρων απόφασης.

III. Προ-επεξεργασία

Α΄. Μετατροπή κατηγορικών χαρακτηριστικών σε αριθμητικά

Όπως φαίνεται στο πίνακα [1] οι τιμές που έχουμε είναι κατηγορικές. Το $Weka$ χρειάζεται τις αριθμητικές επομένως εφαρμόζουμε το φίλτρο $NominalToBinary$ σε όλα τα χαρακτηριστικά της βάσης μας. Αυτό θα μετατρέψει κάθε κατηγορία σε αληθή η ψευδή. Τότε ένα μανιτάρι ή θα ανήκει λόγου χάρη στην κατηγορία $'s'$ του χαρακτηριστικού $cap-surface$ (το οποίο πλέον υποδηλώνεται από το εάν το χαρακτηριστικό $cap-surface=s$ είναι 1) ή όχι (που υποδηλώνεται με 0). Τότε θα έχουμε 122 νέα χαρακτηριστικά και θα αναζητήσουμε για κατηγορίες που είναι μόνο 0 καθώς αυτές δεν θα προσφέρουν στην ανάλυση μας.

Β′. Αφαιρούμε τα χαρακτηριστικά που έχουν μόνο μηδενικές τιμές

Με το φίλτρο $RemoveUnless$ θέτουμε το $MaximumVariancePercentageAllowed$ στο 0 και αυτό θα αφαιρέσει κάθε χαρακτηριστικό που όλες του οι τιμές έχουν συνολική διακύμανση 0, με άλλα λόγια εκείνο που δεν περιέχει κάμια άλλη τιμή εκτός από 0. Αυτό αφαιρεί 9 χαρακτηριστικά και φέρνει το σύνολο των χαρακτηριστικών μας στο 113.

IV. Μοντέλα μηχανικής μάθησης και απεικονίσεις

Α′. *Random Forest*

Εφαρμόζοντας $Random\ Forest$ και $10\text{-}fold\ cross-validation$ το μοντέλο κάνει μια τέλεια κατηγοριοποίηση. Πετυχαίνουμε τα παρακάτω ποσοστά:

$CorrectlyClassifiedInstances$	8124	100%
$IncorrectlyClassifiedInstances$	0	0%
$Kappastatistic$	1	
$Meanabsoluteerror$		0.0005
$Rootmeansquarederror$		0.0037
$Relativeabsoluteerror$		0.1059%
$Rootrelativesquarederror$		0.7322%
$TotalNumberofInstances$		8124

Prediction outcome

Edible	*Poisonous*	
4208	0	*Edible*
0	3916	*Poisonous*

$ConfusionMatrix$

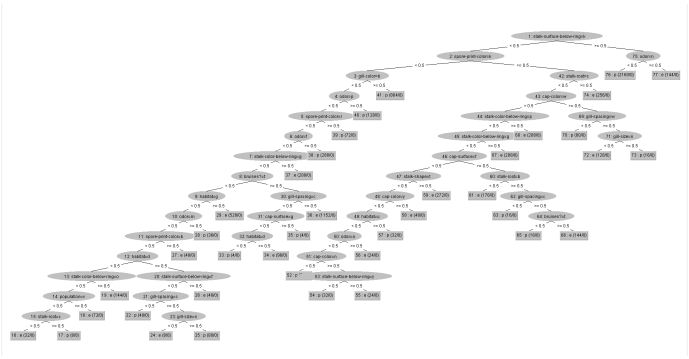
Detailed Accuracy By Class

	$TP\ Rate$	$FP\ Rate$	$Precision$	$Recall$	$F-Measure$	MCC	$ROC\ \ Area$	$PRC\ \ Area$	$Class$
	1.000	0.031	0.972	1.000	0.986	0.971	0.989	0.983	*Edible*
	0.969	0.000	1.000	0.969	0.984	0.971	0.989	0.992	*Poisonous*
W. Avg.	0.985	0.016	0.986	0.985	0.985	0.971	0.989	0.987	

Η πρώτη μας σκέψη παραπέμπει σε $overfit$ ωστόσο θα εξετάσουμε καλύτερα την βάση μας για την εξαγωγή αυτού του συμπεράσματος. Η φύση της βάσης δεδομένων είναι τέτοια ώστε κάθε μοντέλο μηχανικής μάθησης δύναται να δώσει υψηλά ποσοστά απόδοσης και αυτό επειδή αφενός έχουμε 2 κατηγορίες και αφετέρου κάθε δεδομένο μας είναι ένας συνδυασμός από 0 και 1. Ουσιαστικά η βάση μας είναι εντελώς ομογενής και κανονικοποιημένη.

Β′. *Random Tree*

Βλέπουμε το σχεδιάγραμμα ενός δέντρου απόφασης που σε συνδυασμό με άλλα πέτυχε τα παραπάνω ποσοστά:



Φυσικά η εικόνα δεν μας βοηθάει ιδιαίτερα στην κατανόηση της κατηγοριοποίησης. Ο αριθμός των χαρακτηριστικών είναι πολύ μεγάλος και για αυτόν τον λόγο καταφεύγουμε στην απλούστευση της βάσης δεδομένων μας.

V. Μείωση διαστάσεων

Με την τεχνική $Attributeselection$ της $Weka$ δοκιμάσουμε τις παρακάτω τεχνικές:

- $BestFirst + CfsSubtestEval$
- $GreedyStepwise + CfsSubtestEval$

Και κατά συμφωνία επιλέγουμε τις μεταβλητές :

- $odor = a$
- $odor = c$
- $odor = f$
- $odor = l$
- $odor = n$
- $odor = p$
- $gill - color = b$
- $stalk - surface - above - ring = k$

Το παραπάνω έγινε επιλέγοντας το $Filter\ Remove$ εφόσον δώσαμε στο $attributeIndices$: 1-21,26,29-35,37-53,55-112

Α′. *Random Forest* μετά την μείωση διαστάσεων

Εκπαιδεύοντας και πάλι το μοντέλο έχουμε:

Correctly Classified Instances	8004	98.5229%
Incorrectly Classified Instances	120	1.4771%
Kappa statistic		0.9704
Mean absolute error		0.0276
Root mean squared error		0.1176
Relative absolute error		5.5208%
Root relative squared error		23.5448%
Total Number of Instances		8124

DetailedAccuracyByClass

	TP Rate	*FP Rate*	*Precision*	*Recall*	*F − Measure*	*MCC*	*ROC Area*	*PRC Area*	*Class*
	1.000	0.031	0.972	1.000	0.986	0.971	0.989	0.983	*Edible*
	0.969	0.000	1.000	0.969	0.984	0.971	0.989	0.992	*Poisonous*
W. Avg.	0.985	0.016	0.986	0.985	0.985	0.971	0.989	0.987	

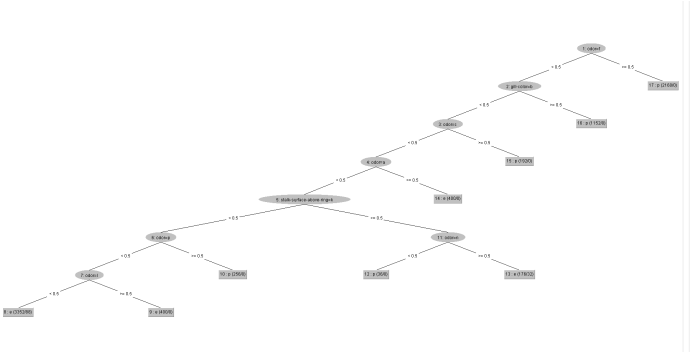
ConfusionMatrix

Prediction outcome

Edible	*Poisonous*	
4208	0	*Edible*
120	3796	*Poisonous*

Τα αποτελέσματα αυτά είναι πολύ καλά καθώς με μόλις 9 χαρακτηριστικά από τα αρχικά 113 επιτυγχάνεται ποσοστό επιτυχίας 98.52%.

Β′. *Random Tree* των 8 χαρακτηριστικών



Το διάγραμμα από ένα *Random Tree* δίνει μια πολύ κατανοητή εικόνα για την κατηγοροιοποίηση μας. Μόνο με την απλή μελέτη αυτού του δυαδικού δέντρου μπορεί να συμπεράνει κανείς με ποσοστό επιτυχίας 98.52% εάν ένα μανιτάρι είναι δηλητηριώδες ή όχι.

VI. Ακρίβεια μοντέλου έναντι πρακτικής χρήσης του

Παρόλο που το παραπάνω αποτελεί ένα πολύ αποδοτικό μοντέλο μηχανικής μάθησης παρατηρούμε ότι το ποσοστό των *False Positive* είναι υψηλό. Στην πραγματικότητα εάν επρόκειτο για κάποιου άλλου είδους κατηγοριοποίηση το ποσοστό αυτό θα ήταν αμελητέο, ωστόσο παρατηρούμε ότι 120 από τα 3916 κατηγοριοποιήθηκαν ως βρώσιμα ενώ είναι δηλητηριώδη.

Θα προσπαθήσουμε να συνδυάσουμε άλλα μοντέλα ώστε να αποφύγουμε αυτό το συμβάν. Πρέπει πρακτικά να επικεντρωθούμε στα χαρακτηριστικά που μηδενίζουν τα *False Possitive* και όχι τα *True Negative*.

VII. Επιστροφή στη Προ-επεξεργασία

Δοκιμάζοντας μερικούς ακόμα κατηγοριοποιητές όπως *J*48, *RepTree*, *Naive Bayes*, *LogisticRegression* Καταλήγουμε στην ακριβώς ίδια κατηγοριοποίηση. Συμπέρασμα το οποίο δεν μας εκπλήσσει καθώς έχουμε 9 χαρακτηριστικά με 0,1 ως τιμές και κατηγοριοποίηση 2 κλάσεων. Μελετάμε τα αποτελέσματα μοντέλων επιλέγοντας άλλα χαρακτηριστικά από τα 113 αρχικά.

Α′. Επιλογή Χαρακτηριστικών με *Ranker Search Method*

Εφόσον δοκιμάσαμε *BestFirst* μεθόδους με πολλαπλά *AttributeEvaluators* καταλήγαμε πάντα στα ίδια χαρακτηριστικά. Έτσι καταφεύγουμε σε *Ranker Search Method* και συγκεκριμένα με *InfoGainAttributeEval* και έχουμε:

$Information\ Gain$	$Attribute\ Number$	$Attribute\ Name$
0.528778	27	$odor = n$
0.357168	24	$odor = f$
0.284429	54	$stalk - surface - above - ring = k$
0.27056	58	$stalk - surface - below - ring = k$
0.269398	36	$gill - color = b$
0.230154	35	$gill - size = n$
0.222702	90	$ring - type = p$
0.207825	92	$spore - print - color = h$

Β′. *Random Forest* με την μέθοδο *Ranker*

Filter Remove εφόσον δώσαμε στο *attributeIndices* : 1-23,25,26,28-34,37-53,55-57,59-89,91,93-112 .
Έτσι λοιπόν πετυχαίνουμε τον εξής πίνακα σύγχυσης:

Prediction outcome

Edible	*Poisonous*	
4112	96	*Edible*
88	3828	*Poisonous*

Τα ποσοστά αυτού του πίνακα μας ικανοποιούν περισσότερο καθώς παρόλο που θυσιάσαμε ακρίβεια *True Negative* από 0.0% σε 0.022% πετύχαμε να μειώσουμε το *False Possitive* από 0.03% σε 0.02%. Η διαφορά στα ποσοστά φαίνεται μικρή αλλά ισοδυναμεί με 32 μανιτάρια αληθώς κατηγοριοποιημένα σε δηλητηριώδη. Σημειώνουμε πώς η ακρίβεια έπεσε στο 97.7351 % από το προηγούμενο 98.5229%. Ποσοστό αμελητέο μπροστά στο προαναφερθέν γεγονός.
Με στόχο λοιπόν να πετύχουμε καλύτερα ποσοστά κατηγοριοποίησης αλλά και όσο το δυνατόν μικρότερο *False Positive* επιλέγουμε να συγκρατήσουμε επιπλέον 4 χαρακτηριστικά στην κατάταξη του *InfoGainAttributeEval* τα οποία είναι:

$Information\ Gain$	$Attribute\ Number$	$Attribute\ Name$
0.192379	21	$bruises? = t$
0.191691	88	$ring - type = l$
0.183563	55	$stalk - surface - above - ring = s$
0.147108	104	$population = v$

Filter Remove : 1-20,22,23,25,26,28-34,37-53,56-57,59-87,89,91,93-103,105-112
Και παίρνουμε:

Prediction outcome

Edible	*Poisonous*	
4112	96	*Edible*
72	3844	*Poisonous*

που βελτιώνει τη συνολική ακρίβεια κατά 0.2% αλλά αυτό που είναι πολύ πιο σημαντικό είναι ότι μειώνεται ο αριθμός των *False Possitive* από 88 σε 72.
προσθέτοντας τέλος άλλα 3 χαρακτηριστικά:

$InformationGain$	$AttributeNumber$	$AttributeName$
0.139342	94	$spore - print - color = n$
0.135077	59	$stalk - surface - below - ring = s$
0.126043	93	$spore - print - color = k$

Με αυτό το τρόπο επιτυγχάνουμε το βέλτιστο δυνατό αποτέλεσμα του να μειώσουμε τα *False Possitive* στο 0. Επιπρόσθετα πετύχαμε την καλύτερη δυνατή ακρίβεια για αυτό τον αριθμό χαρακτηριστικών 99.22%:

Πίνακας I: *Confusion Matrix and Performance Metrics*

Confusion Matrix

Actual Class	*Predicted as Edible*	*Predicted as Poisonous*	*Total*
Edible	4144	64	4208
Poisonous	0	3916	3916
Total	4144	3980	8124

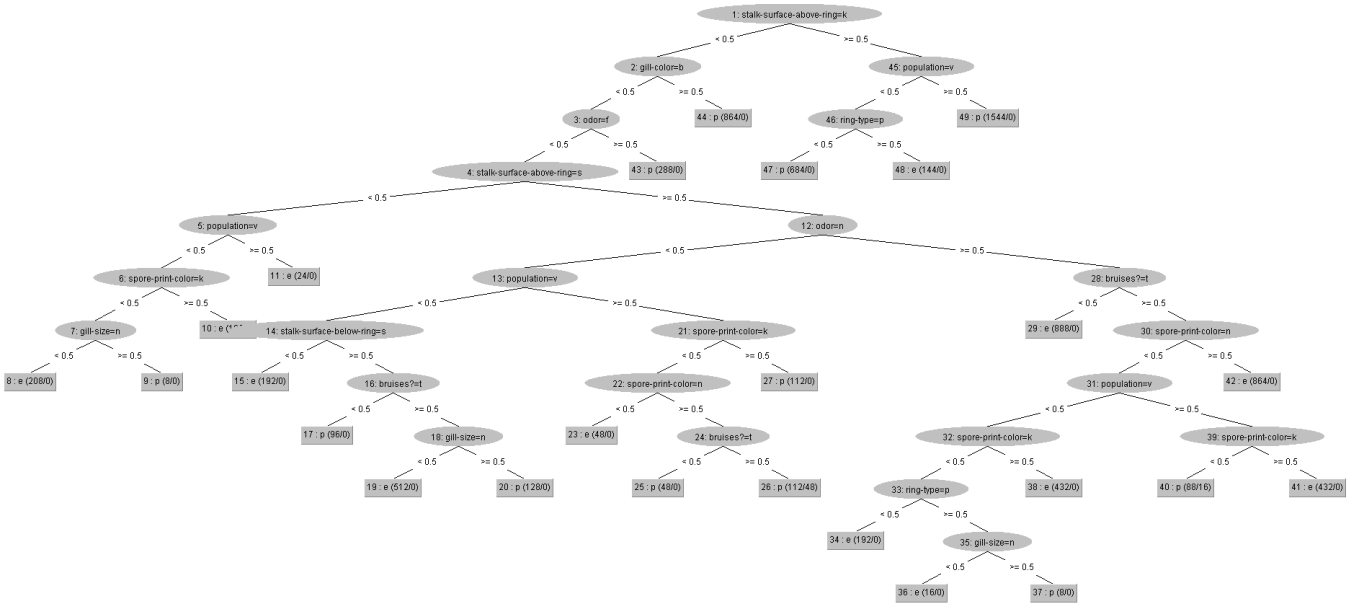
Overall Performance Metrics

Metric	*Value*	*Metric*	*Value*
Correctly Classified Instances	8060	*Relative Absolute Error*	2.0129%
Incorrectly Classified Instances	64	*Root Relative Squared Error*	14.2251%
Kappa statistic	0.9842	*Total Number of Instances*	8124
Mean Absolute Error	0.0101		
Root Mean Squared Error	0.0711		

Detailed Accuracy By Class

Class	*TP Rate*	*FP Rate*	*Precision*	*Recall*	*F − Measure*	*MCC*	*ROC Area*	*PRC Area*
Edible	0.985	0.000	1.000	0.985	0.992	0.984	1.000	1.000
Poisonous	1.000	0.015	0.984	1.000	0.992	0.984	1.000	1.000
WeightedAvg.	0.992	0.007	0.992	0.992	0.992	0.984	1.000	1.000

Γ′. Τελικό διάγραμμα *Random Tree* με τα βέλτιστα αποτελέσματα



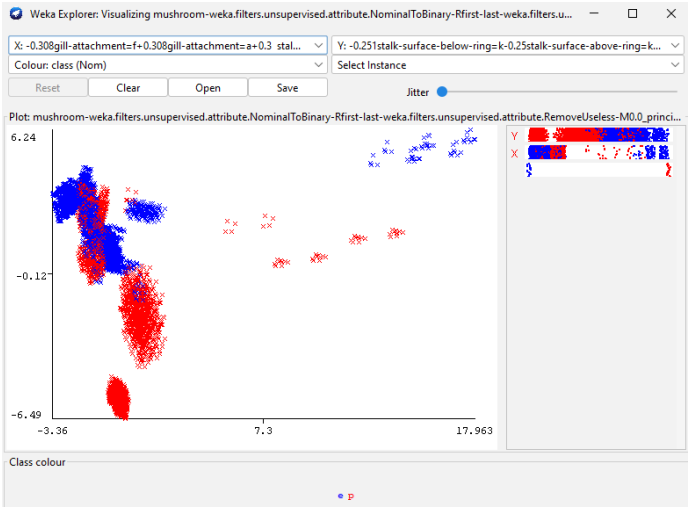
Το δέντρο δεν είναι τόσο απλό όσο το πρώτο αλλά καταφέρνομε τα βέλτιστα δυνατά αποτελέσματα με αυτό. Επιπρόσθετα παρόλο που οπτικά φαίνεται περίπλοκο είναι πολύ εύκολο στην χρήση τόσο υπολογιστικά όσο και στην κατανόηση διότι κάθε κόμβος είναι απλώς μια ερώτηση που απαντάται με ναι ή όχι αναλόγως εάν έχει ένα μανιτάρι ή όχι το εκάστοτε χαρακτηριστικό.

VIII. Συσταδοποίηση

Το αρχικό μας *dataset* των 113 χαρακτηριστικών παρουσιάζει μια ιδιαιτερότητα στην συσταδοποίηση του. Η βάση στην ολότητα της αποτελείται από 0 και 1 σε κάθε της χαρακτηριστικό, επομένως η οπτικοποίηση σε 2 ή 3 διαστάσεις δεν θα προσδώσει κάποια ιδιαίτερη αξία στην ανάλυση που επιχειρούμε. Σαν πρώτο βήμα θα επιλέξουμε μια μέθοδο μείωσης διάστασης για να οπτικοποιήσουμε το πως εκείνη διαχωρίζει τις 2 ομάδες μας *Edible* και *Poisonous*.

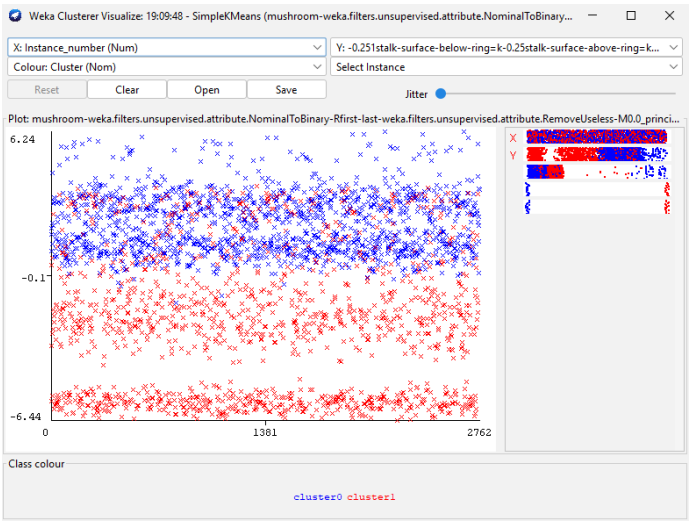
Α'. *PCA*

Με τη μέθοδος των κύριων συνιστωσών μέσω του *Weka*, αφού ρυθμίσουμε τις παραμέτρους της δοκιμάζουμε το πλήθος των συνιστωσών που επιθυμούμε να καταλήξουμε. Στην αρχή για μια πολύ απλή αλλά κατά μεγάλο βαθμό επεξηγηματική απεικόνιση δίνουμε στην *PCA* αριθμό κυρίων συνιστωσών 2 και πλήθος από χαρακτηριστικά 15 και από αυτά δημιουργείται ένα ζεύγος νέων χαρακτηριστικών που είναι γραμμικώς συνδυασμός των προηγούμενων και δίνει την παρακάτω εικόνα: Η



διαφοροποίηση είναι αρκετά καθαρή οπτικά, ωστόσο θυσιάζουμε πολύ πληροφορία για αρκετά ασαφή αποτελέσματα. Παρατηρούμε πως υπάρχουν αρκετές περιοχές συνκάλυψης κάτι που όπως είδαμε παραπάνω μας προβληματίζει ιδιαίτερα, ειδικά εφόσον το σφάλμα τύπου 1 έχει καταστροφικές συνέπειες. Επιπρόσθετα για κάθε νέα προσθήκη δεδομένων ο σωστός τρόπος συσταδοποίησης θα ήταν να εφαρμόσουμε και πάλι την μέθοδο κύριων συνιστωσών, που σημαίνει πως σπαταλούμε αρκετό χρόνο και πόρους. Ωστόσο η μέθοδο κύριων συνιστωσών βοηθά στο να έχουμε μια σαφή εικόνα του στόχου μας ως προς την συσταδοποίηση. Σε αυτές τις 2 συνιστώσες λοιπόν επιχειρούμε να εφαρμόσουμε την μέθοδο *SimpleKMeans* της *Weka* και καταγράφουμε τα αποτελέσματα της.

$$Within\ cluster\ sum of\ squared\ errors : 392.59$$



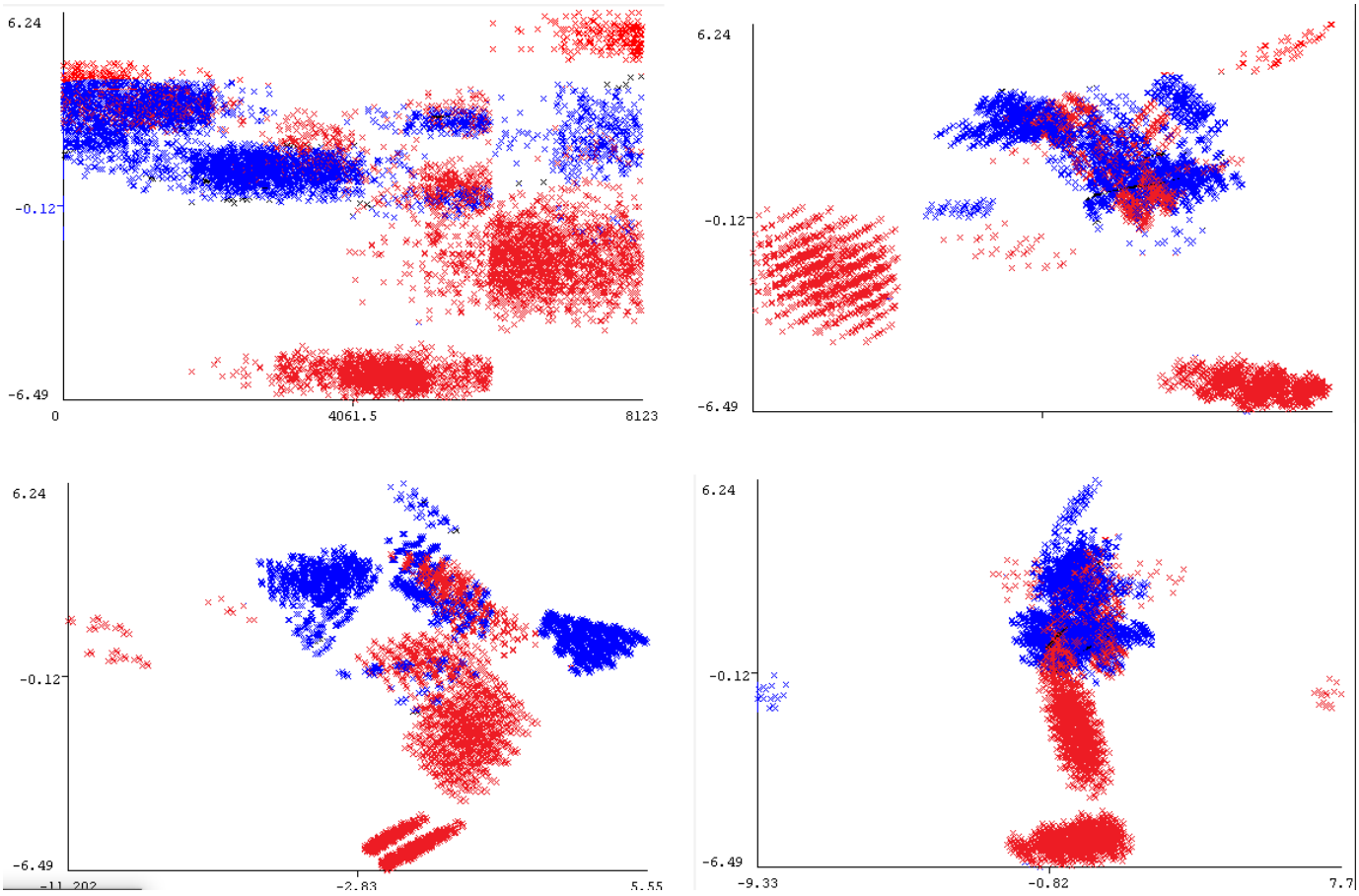
Η διαφοροποίηση εδώ είναι σαφώς καλύτερη και πιο ξεκάθαρη. Η μπλε ομάδα αποτελεί τα *Edible* ενώ η κόκκινη τα *Poisonous*. Και εδώ παρατηρούμε πως μια αρκετά σημαντική λωρίδα της κόκκινης ομάδας συμπίπτει με την μπλε. Σημειώνουμε πως η διαχωροποίηση είναι ικανοποιητική από την άποψη ότι ένα μικρό πλήθος της κόκκινης ομάδας (συγκριτικά με αυτό της μπλε) συμπίπτουν. Όμως και πάλι επειδή η συσταδοποίση δηλητηριωδών μανιταριών ως βρώσιμα είναι πολύ σημαντικό λάθος επιχειρούμε να εκπαιδεύσουμε καλύτερα το μοντέλο.

Β'. $K-means$

Κάνοντας πολλές δοκιμές με $K-means$ καταλήγουμε πως οι κλάση των *Poisonous* μανιταριών είναι πάντοτε πιο καθαρά διαχωρισμένη από εκείνη τον *Edible*, γεγονός που κάνει την ανάλυση μας δυσκολότερη. Σε όλες τις συσταδοποιήσεις χρησιμοποιούμε διαφορετικού πλήθους κύριες συνιστώσες και αυτό επειδή με αυτόν τον τρόπο δίνουμε πολύ καθαρότερη εικόνα στις ομάδες μας, εικόνα που δεν γίνεται να φανεί εάν παραμείνουν τα δεδομένα μας στις τιμές 0 και 1. Σε κάθε δοκιμή εφαρμόζουμε το φίλτρο *Standardize* της *Weka*. Αυτό βοηθά να είναι πιο συσπειρωμένες οι ομάδες μας και να γίνεται πιο διαυγής ο διαχωρισμός. Η βάση δεδομένων μας δεν χρειάζεται απαραίτητα αυτό το βήμα καθώς είναι αρκετά κανονικοποιημένη δεδομένου ότι η μέση τιμή και η διακύμανση κάθε χαρακτηριστικού είναι μεταξύ του 0 και του 1, ωστόσο δίνει μια πιο συμπαγή εικόνα και βοηθά στην καλύτερη απόδοση της *PCA*.
Παραδίδουμε μερικές απο τις δοκιμές της $K-means$ σε 60 κύριες συνιστώσες αφού πρώτα κάνουμε *Standardize*:

$$Within\ cluster\ sum of\ squared\ errors : 5059.15$$

$$Incorrectly\ clustered\ instances : 3118.0 \mid \ 38.38\%$$



Οι διαχωρισμοί δεν είναι πολύ ξεκάθαροι ωστόσο υπάρχουν ξεκάθαρα ευδιάκριτες ομάδες των *Poisonus* (κόκκινο) μανιταριών. Το κύριο ζητούμενο της ανάλυσης μας είναι να κάνουμε όσο πιο καθαρή γίνεται την ομάδα των *Edible* (μπλε). Δεν μπορούμε να συμπεράνουμε μόνο από το $Whithin\,class\,SSE$ εάν το μοντέλο είναι αποδοτικό αλλά μπορούμε να τα συγκρίνουμε το SSE μεταξύ των μοντέλων.
Οι πάνω περιοχές στα γραφήματα είναι αυτές που θέλουμε να ξεχωρίσουμε καθώς στην κάτω πλευρά υπάρχουν πολλά κόκκινα σημεία τα οποία πρέπει να αποφεύγονται.

Γ′. Ρυθμίσεις παραμέτρων και τελικό αποτέλεσμα

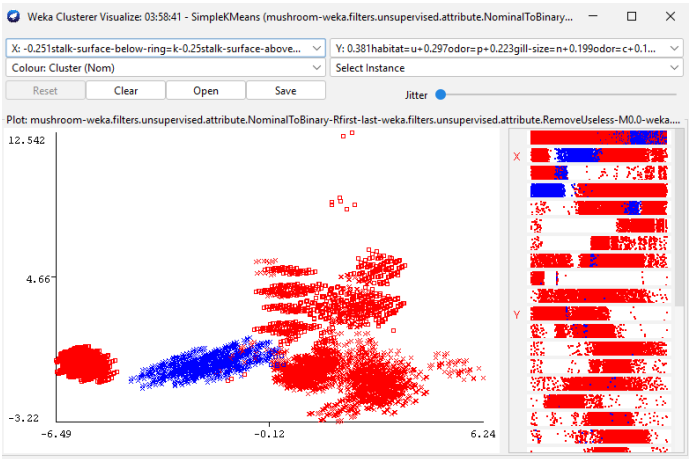
Μετά από δοκιμές με τα άλλα μοντέλα όπως $Density - based\, Clustering$ συμπεραίνουμε πως το $K - means$ αποτελεί την πιο αποδοτική μέθοδο για την βάση και έτσι ελέγχουμε εάν μπορούμε να βελτιώσουμε την απόδοση ρυθμίζοντας τις παραμέτρους.Πρώτα θα μειώσουμε τον αριθμό των κύριων συνιστωσών σε 30 καθώς επιτυγχάνεται μικρή μείωση στα $Incorrectly\, clustered\, instances = 3084.04$ (από το αρχικό 5059.15). Έπειτα αλλάζουμε την μέτρηση της απόστασης στην $K - means$ από ευκλείδεια στην απόσταση $Manhattan$ σε νόρμα 1 δηλαδή και περιμένουμε καλύτερα αποτελέσματα λόγω της μεγάλης διάστασης της βάσης δεδομένων. Πράγματι έχουμε σημαντική αύξηση στην ακρίβεια:

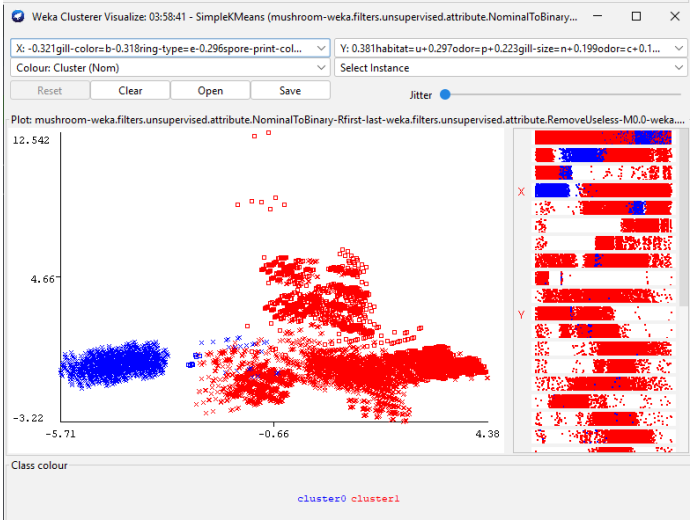
$$Incorrectly\, clustered\, instances : 2171.0 \mid \ 26.72\%$$

Θα παραθέσουμε και το ανάλογο του SSE το οποίο όμως δεν μας δίνει από μόνο του αποδοτική εικόνα για τη επίδοση του μοντέλου.

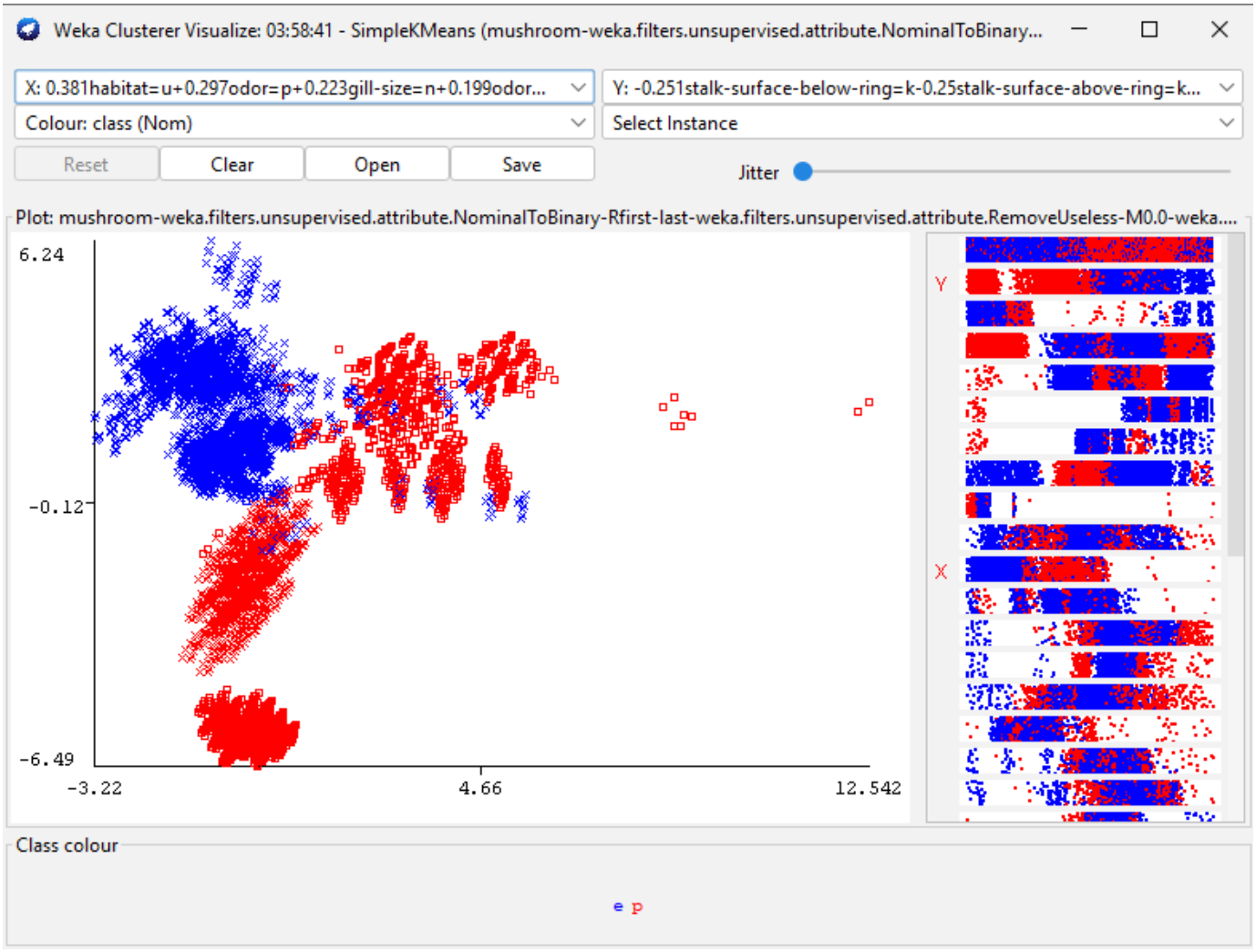
$$Sum\, of\, within\, cluster\, distances : 15600.41$$

Οπτικά τα αποτελέσματα είναι σαφώς καλύτερα. Είναι άξιο να σημειώσουμε πως τα χρώματα και οι ομαδοποιήσεις στις παρακάτω εικόνες δεν αποτυπώνουν απολύτως την πραγματική συσταδοποίηση, αλλά μόνο την απόδοση του μοντέλου.





Ωστόσο ολοκληρώνουμε την έρευνα μας επιλέγοντας τις 2 κύριες συνιστώσες που διαχωρίζουν καλύτερα τα δεδομένα μας και καταλήγουμε στην παρακάτω εικόνα η οποία αποτυπώνει τα δεδομένα σύμφωνα με το χαρακτηριστικό *class*. Αυτή είναι η καθαρότερη αποτύπωση που καταφέραμε να πάρουμε εάν θέλαμε να χαρτογραφήσουμε την κατηγορία ενός μανιταριού. Η παρακάτω είναι η καλύτερη συσταδοποίηση που μπορούσαμε να παράξουμε και αυτό επειδή η κλάση των *Edible* (*e*) είναι εντελώς καθαρή από δηλητηριώδη μανιτάρια (σημειώνουμε πως δεν είναι αυτή η καλύτερη δυνατή απόδοση αλλά επιτυγχάνει το καλύτερο αποτέλεσμα όσον αφορά το σφάλμα τύπου 1).



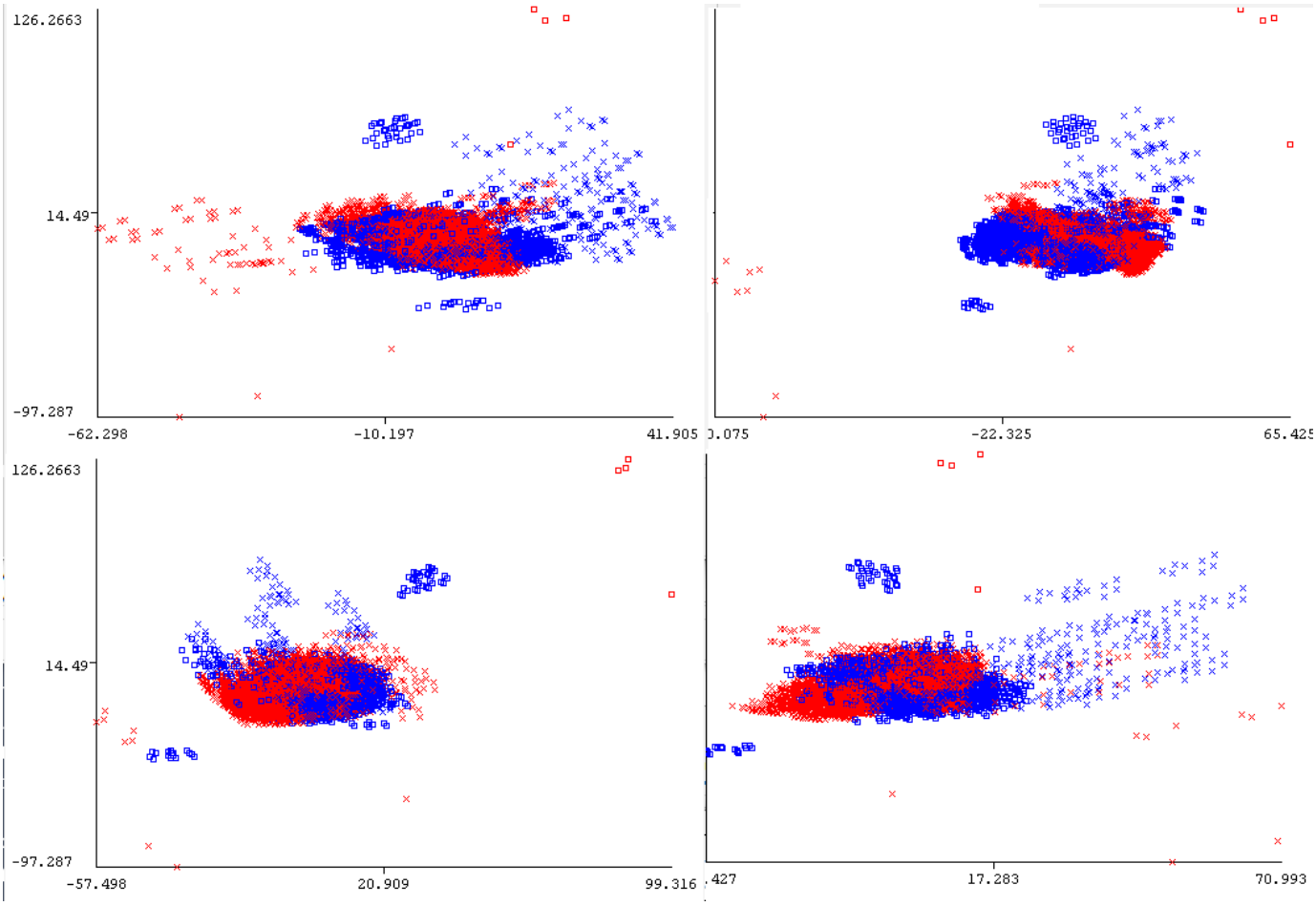
IX. Μετα-ανάλυση

Σημειώνουμε κάποια αξιόλογα αποτελέσματα που βρήκαμε με πειραματισμό με το *Weka*. Σε αναζήτηση μεθόδου περαιτέρω μείωσης διάστασης παράλληλα με την *PCA* δοκιμάσαμε την *RandomProjection*. Οι δύο μέθοδοι δεν βοηθούν η μία την άλλη ωστόσο καταγράφουμε τον πίνακα σύγχυσης που βρήκαμε καθώς και το χαμηλότερο *SSE*:

Poisonous	*Edible*	
4016	192	*Edible*
3912	4	*Poisonous*

Ο προβληματισμός μας είναι πως οπτικά οι συσταδοποιήσεις είναι εντελώς μη διαχωρισμένες

Within cluster sum of squared errors : 782.1



Αναφορές

[1] Γ. Εασον, Β. Νοβλε, ανδ Ι. Ν. Σνεδδον, 'Ον ζερταιν ιντεγραλς οφ Λιπσζηιτζ-Ηανκελ τψπε ινολινγ προδυζτς οφ Βεσσελ φυνζτιονς,' Πηιλ. Τρανς. Ροψ. Σος. Λονδον, ολ. Α247, ππ. 529–551, Απριλ 1955.

[2] Θ. Ψλερκ Μαξωελλ, Α Τρεατισε ον Ελεζτριζιτψ ανδ Μαγνετισμ, 3ρδ εδ., ολ. 2. Οξφορδ: Ψλαρενδον, 1892, ππ.68–73.

[3] Ι. Σ. Θαζοβς ανδ Ψ. Π. Βεαν, 'Φινε παρτιζλες, τηιν φιλμς ανδ εξζηανγε ανισοτροπψ,' ιν Μαγνετισμ, ολ. ΙΙΙ, Γ. Τ. Ραδο ανδ Η. Συηλ, Εδς. Νεω Ψορκ: Αζαδεμις, 1963, ππ. 271–350.

[4] Κ. Ελισσα, 'Τιτλε οφ παπερ ιφ κνοων,' υνπυβλισηεδ.

[5] Ρ. Νιζολε, 'Τιτλε οφ παπερ ωιτη ονλψ φιρστ ωορδ ζαπιταλιζεδ,' Θ. Ναμε Στανδ. Αββρε., ιν πρεσς.

[6] Ψ. Ψοροζυ, Μ. Ηιρανο, Κ. Οκα, ανδ Ψ. Ταγαωα, 'Ελεζτρον σπεζτροσζοπψ στυδιες ον μαγνετο-οπτιζαλ μεδια ανδ πλαστις συβστρατε ιντερφαζε,' ΙΕΕΕ Τρανσλ. Θ. Μαγν. Θαπαν, ολ. 2, ππ. 740–741, Αυγυστ 1987 [Διγεστς 9τη Αννυαλ ΔνφΜαγνετιζς Θαπαν, π. 301, 1982].

[7] Μ. Ψουνγ, Τηε Τεζηνιζαλ Ωριτερ΄ς Ηανδβοοκ. Μιλλ Ψάλλεψ, Ψα: Υνιερσιτψ Σζιενζε, 1989.