

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ

ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ



Μάθημα: Τεχνητή νοημοσύνη.

Όνομα: Μουστάκας Αριστείδης

AEM: 2380

email: moustakas@csd.auth.gr

1) Nominal or numeric attributes.

1. pclass : nominal
2. sex : nominal
3. age : numeric
4. sibsp : numeric
5. parch : numeric
6. ticket : nominal
7. fare : numeric
8. cabin : nominal
9. embarked : nominal
10. survived : nominal

2) Missing values for each attribute.

1. pclass : 0 (0%)
2. sex : 0 (0%)
3. age : 177 (20%)
4. sibsp : 0 (0%)
5. parch : 0 (0%)
6. ticket : 0 (0%)
7. fare : 0 (0%)
8. cabin : 687 (77%)
9. embarked : 2 (0%)
10. survived : 0 (0%)

3)

- ❖ Για τον αλγόριθμο ibk με 1 φίλο (με percentage split 66%) έχουμε την παρακάτω ακρίβεια:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.767	0.298	0.81	0.767	0.788	0.735	0
	0.702	0.233	0.645	0.702	0.672	0.735	1
Weighted Avg.	0.743	0.274	0.748	0.743	0.744	0.735	

όπου παρατηρούμε ότι το ποσοστό των παραδειγμάτων με τιμή 0 που ταξινομούνται σωστά είναι $TPR = 76.7\%$, αντίστοιχα ότι το ποσοστό των παραδειγμάτων με τιμή 1 που ταξινομούνται σωστά είναι $TPR = 70.2\%$. Γενικότερα το ποσοστό των παραδειγμάτων που ταξινομούνται σωστά είναι 74.3% . Επίσης παρατηρούμε ότι το ποσοστό των παραδειγμάτων με τιμή 0 που ταξινομούνται λάθος (δηλαδή που έπρεπε να ταξινομηθούν ως 1) είναι $FPR = 29.8\%$ και τέλος το ποσοστό των παραδειγμάτων με τιμή 1 που ταξινομούνται λάθος (δηλαδή που έπρεπε να ταξινομηθούν ως 0) είναι $FPR = 23.3\%$. Γενικότερα το ποσοστό των παραδειγμάτων που ταξινομούνται λάθος είναι 27.4% . Όπως βλέπουμε τα στοιχεία που έχουν τιμή 0 ταξινομούνται καλύτερα. Επίσης το precision (μέσος όρος των precision του 0 και του precision του 1) είναι ίσος με $Precision = 74.8\%$. Τέλος ο αρμονικός μέσος των precision και recall είναι το $F-Measure = 74.4\%$.

- ❖ Για τον αλγόριθμο j48 (με percentage split 66%) έχουμε την παρακάτω ακρίβεια:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.862	0.342	0.807	0.862	0.834	0.781	0
	0.658	0.138	0.743	0.658	0.698	0.781	1
Weighted Avg.	0.785	0.265	0.783	0.785	0.783	0.781	

όπου παρατηρούμε ότι το ποσοστό των παραδειγμάτων με τιμή 0 που ταξινομούνται σωστά είναι $TPR = 86.2\%$, αντίστοιχα ότι το ποσοστό των παραδειγμάτων με τιμή 1 που ταξινομούνται σωστά είναι $TPR = 65.8\%$. Γενικότερα το ποσοστό των παραδειγμάτων που ταξινομούνται σωστά είναι 78.5% . Επίσης παρατηρούμε ότι το ποσοστό των παραδειγμάτων με τιμή 0 που ταξινομούνται λάθος (δηλαδή που έπρεπε να ταξινομηθούν ως 1) είναι $FPR = 34.2\%$ και τέλος το ποσοστό των παραδειγμάτων με τιμή 1 που ταξινομούνται λάθος (δηλαδή που έπρεπε να ταξινομηθούν ως 0) είναι $FPR = 13.8\%$. Γενικότερα το ποσοστό των παραδειγμάτων που ταξινομούνται λάθος είναι 27.4% . Όπως βλέπουμε τα στοιχεία που έχουν

τιμή 0 ταξινομούνται καλύτερα. Επίσης το precision (μέσος όρος των precision του 0 και του precision του 1) είναι ίσος με Precision=78.3% . Τέλος το precision (μέσος όρος των precision του 0 και του precision του 1) είναι ίσος με Precision=78.3% .

Παρατηρούμε ότι από τους 2 παραπάνω αλγορίθμους το μεγαλύτερο ποσοστό πρόβλεψης σωστών αποτελεσμάτων το έχει ο j48.

4)

- Για τον αλγόριθμο ibk με 1 φίλο (με percentage split 66%) το ποσοστό των παραδειγμάτων που ταξινομείται σωστά είναι:

```
Correctly Classified Instances      225          74.2574 %
```

- Για το αλγόριθμο j48 (με percentage split 66%) το ποσοστό των παραδειγμάτων που ταξινομείται σωστά είναι:

```
Correctly Classified Instances      238          78.5479 %
```

5)

- Για τον αλγόριθμο ibk με 1 φίλο (με cross validation 5) έχουμε ακρίβεια:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.825	0.354	0.789	0.825	0.807	0.747	0
	0.646	0.175	0.697	0.646	0.671	0.747	1
Weighted Avg.	0.756	0.285	0.754	0.756	0.755	0.747	

- Για τον αλγόριθμο j48 (με cross validation 5) έχουμε ακρίβεια:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.891	0.415	0.775	0.891	0.829	0.78	0
	0.585	0.109	0.769	0.585	0.664	0.78	1
Weighted Avg.	0.773	0.298	0.773	0.773	0.766	0.78	

Παρατηρούμε ότι ο j48 έχει και καλύτερο ποσοστό παραδειγμάτων που ταξινομούνται 0 ή 1 και είναι όντως 0 ή 1 (precision) αλλά και καλύτερο F-Measure που είναι ο αρμονικός μέσος των recall και precision.

Επίσης, για τον αλγόριθμο ibk με 1 φίλο (με cross validation 5) το ποσοστό των παραδειγμάτων που ταξινομείται σωστά είναι:

Correctly Classified Instances 674 75.6453 %

και αντίστοιχα για το αλγόριθμο j48 (με cross validation 5) το ποσοστό των παραδειγμάτων που ταξινομείται σωστά είναι:

Correctly Classified Instances 689 77.3288 %

Παρατηρούμε ότι ο αλγόριθμος j48 (με cross validation 5) έχει ταξινομή σωστά περισσότερα στοιχεία.

6)

Κάνοντας απαλοιφή των ελλιπών έχουμε τα παρακάτω αποτελέσματα:

- Για τον αλγόριθμο ibk με 1 φίλο (με cross validation 5) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.816	0.304	0.812	0.816	0.814	0.768	0
	0.696	0.184	0.702	0.696	0.699	0.768	1
Weighted Avg.	0.77	0.258	0.77	0.77	0.77	0.768	

Correctly Classified Instances 686 76.9921 %

- Για τον αλγόριθμο j48 (με cross validation 5) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.891	0.404	0.78	0.891	0.832	0.788	0
	0.596	0.109	0.773	0.596	0.673	0.788	1
Weighted Avg.	0.778	0.291	0.777	0.778	0.771	0.788	

Correctly Classified Instances 693 77.7778 %

Παρατηρούμε ότι και για τους 2 αλγόριθμους τα TPR,Precision,Recall και F-Measure αυξάνονται όπως και τα παραδείγματα που ταξινομούνται σωστά. Άρα συμπεραίνουμε ότι μετά την εφαρμογή του φίλτρου για την απαλοιφή των ελλιπών τιμών η ακρίβεια αυξάνεται.

7)

- ✚ Για τον αλγόριθμο SMO (με cross validation 10) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.885	0.254	0.848	0.885	0.866	0.815	0
	0.746	0.115	0.802	0.746	0.773	0.815	1
Weighted Avg.	0.832	0.201	0.83	0.832	0.83	0.815	

όπου βλέπουμε πως οι μετρικές TPR, Precision Recall και F-Measure είναι πάνω από 80%. Άρα ο SMO είναι ένας πολύ "δυνατός" αλγόριθμος.

- ✚ Για τον αλγόριθμο BFTtree (με cross validation 10) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.914	0.623	0.702	0.914	0.794	0.64	0
	0.377	0.086	0.733	0.377	0.498	0.64	1
Weighted Avg.	0.708	0.417	0.714	0.708	0.681	0.64	

όπου παρατηρούμε ότι οι μετρικές αυτού του αλγορίθμου για το συγκεκριμένο dataset είναι αρκετά χαμηλές.

- ✚ Για τον αλγόριθμο RBFNetwork (με cross validation 10) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.847	0.31	0.814	0.847	0.83	0.821	0
	0.69	0.153	0.738	0.69	0.713	0.821	1
Weighted Avg.	0.787	0.25	0.785	0.787	0.785	0.821	

όπου βλέπουμε πως οι μετρικές TPR, Precision Recall και F-Measure είναι λίγο κάτω από 80%. Άρα ο RBFNetwork είναι ένας καλός αλγόριθμος.

- ✚ Για τον αλγόριθμο Zero-R (με cross validation 10) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	1	1	0.616	1	0.763	0.497	0
	0	0	0	0	0	0.497	1
Weighted Avg.	0.616	0.616	0.38	0.616	0.47	0.497	

όπου βλέπουμε πως οι μετρικές αυτού του αλγορίθμου μας δίνουν πολύ χαμηλά ποσοστά ακρίβειας.

🗨 Για τον αλγόριθμο BayesNet (με cross validation 10) έχουμε τα παρακάτω αποτελέσματα:

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	ROC Area	Class
	0.849	0.327	0.806	0.849	0.827	0.838	0
	0.673	0.151	0.735	0.673	0.702	0.838	1
Weighted Avg.	0.781	0.26	0.779	0.781	0.779	0.838	

όπου βλέπουμε πως οι μετρικές TPR, Precision Recall και F-Measure είναι λίγο κάτω από 80%. Άρα ο BayesNet είναι ένας αρκετά καλός αλγόριθμος.