Performance measures in discrete supervised classification

Ana Sousa Ferreira and Anabela Marques

Abstract The evaluation of results in Cluster Analysis frequently appears in the literature, and a variety of evaluation measures have been proposed. On the contrary, in supervised classification, particularly in the discrete case, the subject of results evaluation is relatively rare in the literature of the area. This is the motto for the present study. The evaluation of the performance of any model of supervised classification is, generally, based in the number of cases correctly or incorrectly predicted by the model. However, these measures can lead to a misleading evaluation when data is not balanced. More recently, another type of measures had been studied as coefficients of association or agreement, the Huberty's index, Mutual information or even ROC curves. Exploratory studies have been made to understand the relationship between each measure and data characteristics, namely, samples size, balance and separability of classes. For this purpose, we resort to simulated data and use a Beta regression model on the performance of the models.

Key words: Balanced classes, Performance measures, Separability of classes, Supervised classification.

1 Introduction

In Statistics, a supervised classification problem exists when the aim is to identify to which, of a set of classes defined a priori, a new observation belongs, on the basis of a training set of data containing subjects whose class membership is known. For

Ana Sousa Ferreira
Faculdade de Psicologia, Universidade de Lisboa, Business Research Unit (BRU-IUL), Lisboa, Portugal, e-mail: asferreira@psicologia.ulisboa.pt

Anabela Marques
Escola Superior de Tecnologia do Barreiro, IPS, CIIAS, Barreiro, Portugal e-mail: anabela.marques@estbarreiro.ips.pt

example, consider a breast cancer dataset that contains nine variables describing 300 women that have suffered breast cancer and whether or not breast cancer recurrence within five years. So, we are facing a a binary classification problem: of the 300 observed women, how many will or will not suffer a recurrence of breast cancer within five years?

Performance evaluation allows either to evaluate the quality of a new classification model either to choose the most appropriate technique to solve a specific supervised classification problem. In fact, performance evaluation is fundamental in supervised classification: ”it is almost unthinkable to carry out any research work without an experimental section where the performance of the new proposed algorithm is tested and compared with other already proposed methods” ([9], p.1).

In the breast cancer example, are *False Negatives* (women wrongly diagnosed with no breast cancer recurrence) probably worse than *False Positives* (women wrongly diagnosed with recurrence) in this problem? In fact, more detailed screening will certainly clear the *Positives*, but the *False Negatives* women will be sent home and probably will lost follow-up evaluations.

The results of a supervised classification problem can be resumed in a contingency table named the confusion matrix. In Table 1 is presented the confusion matrix for the breast cancer data:

Table 1 Breast cancer confusion matrix

		Predicted classes		
		Recurrence	No Recurrence	
True	Recurrence	25 (TP)	75 (FN)	100
Classes	No Recurrence	18 (FP)	182 (TN)	200
		43	257	300

In the Medicine field, the 25, 75, 18 and 182 values are habitually referred to as *True Positives (TP)*, *False Negatives (FN)*, *False Positives (FP)* and *True Negatives (TN)*, respectively. This terminology, which has been extended to many other fields of application, stems, for example, from the fact that a diagnostic exam indicates that a given woman suffers a recurrence while in reality, the woman didn’t. Therefore, here we are dealing with a *False Positive* case. Some evaluation measures in classification are associated with this type of classification problem.

In supervised classification, global accuracy (or misclassification error) is widely used in classification problems since it is easy to compute and understand. Sometimes, accuracy is selected without considering in depth whether it is the most appropriate score to measure the quality of a classifier for the specific classification problem at hand. For Table 1, the global accuracy value is 0.69 (and the misclassification error 0.31).

In the discrete field there is often a problem of dimensionality, due to the number of parameters to be estimated in each model being too large, frequently samples being small and sparseness. So, most of the discrete models perform poorly, especially when classes are unbalanced and there is also a class separability problem. Thus, in discrete supervised classification the evaluation of results becomes even more relevant, when comparing new proposed models with other important models of the supervised classification literature. It is the aim of this paper to explore the evaluation of results in supervised classification, by comparing the correct classification rate with other types of measures ([3], [8]).

2 Performance measures

In the statistical literature, the most reported measure is accuracy that evaluates the overall efficiency of an algorithm. However, accuracy can be a misleading evaluation measure when data is not balanced ([9],[3], [5], [8]). Then, several measures have been defined in order to evaluate correctly the performance of each algorithm. The *Accuracy* (*Acc*) rate is the most commonly used measure, and quantifies the overall efficiency of the model. In fact, the *Accuracy* seeks to respond to the question: ”Overall, how frequently does the classification model decide correctly?”
The *Correctly classified rate* of cases in class 1 is also referred to as *Sensitivity*, and measures efficiency in class 1. The *Correctly classified rate* of cases in class 2 is also referred to as *Specificity*, and measures efficiency in class 2. In Table 2, some of the evaluation measures based on the confusion matrix are presented:

Table 2 Performance measures based on the confusion matrix

Measures	Definition
Correctly classified rate or *Accuracy (Acc)*	$\frac{TP+TN}{TP+TN+FP+FN}$
Accuracy of class 1 (Acc1) or *Sensitivity*	$\frac{TP}{TP+FN}$
Accuracy of class 2 (Acc2) or *Specificity*	$\frac{TN}{TN+FP}$
Precision (Pre)	$\frac{TP}{TP+FP}$

Clearly, a good classification model should be capable of identifying both *True Positive* and *True Negative* cases. In precise terms, *Sensitivity* is the rate of *True Positive* cases, while *Specificity* is the rate of *True Negative* cases. Finally, *Precision,* also referred to as the positive predictive value, measures the precision of the model, providing the answer to another question: ”Among the cases classified by the model as *Positive*, that is, belonging to Class 1, how many effectively are?” Thus, a high *Precision* value shows a model that it is a good predictor.

In general, the performance measures used do not provide a balance between the *False Positive* and *False Negative* cases. The combined performance measures, presented in Table 3, seek to obtain improved parity between them.

Table 3 Combined performance measures

Measures	Definition
Balanced accuracy (B_Acc)	$\frac{Sensitivity+Specificity}{2}$
Geometric mean (G_mean)	$\sqrt{Sensitivity \times Specificity}$
F measure (F)	$\frac{2\times Sensitivity\times Precision}{Sensitivity+Precision}$

Balanced accuracy is the arithmetic mean between *Sensitivity* and *Specificity* and, when compared with the *Overall accuracy*, will tend to be lower when the model is unable to classify both classes equally correctly. The *Geometric mean* measures the balance between the classification in the two classes. A low *Geometric mean* value indicates a weak performance in the class considered to be positive (usually deemed the class of most interest). Finally, the *F measure* combines the *Sensitivity* and *Precision* measures, even when the classes of data are really balanced. The afore-mentioned evaluation measures, which are generally simple or combined rates, naturally vary in the $[0,1]$ interval.
Evaluation measures of a different type ([9], [3], [5], [8]), which indicate association or agreement between real and predicted classes ([2]), have been referred to by several authors. On the other hand, an evaluation of the effective improvement the model brings to the majority rule appears to be of relevance ([6]). These less traditional measures in supervised classification are presented in Table 4.

The *Phi coefficient* (ϕ) is a known measure of association between two binary variables, and take values in the interval $[-1,1]$. The positive sign of this coefficient indicates a higher number of cases where the classification model has decided correctly. The negative sign, on the contrary, points to the existence of more incorrectly decided cases. *Cohen's Kappa statistic* ([2]) may be defined as the proportion of agreement between two classifications after removal of the agreement proportion owing to the random, and may also take values in the interval $[-1,1]$. Finally, *Huberty's index* ([6]) evaluates the performance of a model as the degree of classification correction achieved, in comparison with a percentage of correctly classified cases by the majority rule, defined as the ratio between effective improvement and possible improvement in the classification. This index is the only evaluation measure presented that take values outside the interval $[-1,1]$.

Table 4 Less traditional performance measures

Measures	Definition
Phi coefficient (ϕ)	$\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(FN+TN)(TP+FN)(FP+TN)}}$
Cohen's Kappa statistic (K)	$\frac{A_{cc}-P_{random}}{1-P_{random}}$, where $P_{random} = (\frac{TP+FN}{N} \times \frac{TP+FP}{N}) + (\frac{TN+FP}{N} \times \frac{TN+FN}{N})$ and $N = TP+TN+FP+FN$
Huberty's index (H)	$\frac{P_{cc}-P_m}{1-P_m}$, where P_{cc} - % correctly classified cases and P_m - % correctly classified cases in accordance with the majority rule

3 Numerical results

In order to understand the relationship between each performance measure and data characteristics, Beta regression ([7]) or Multiple linear regression models were used, according to the variation intervals of the performance measures. For this purpose, we resort to simulated data to predict performance measures based on data characteristics and understand the relative impact of each experimental complexity factor on performance. For the sake of simplicity , in our study, we will be focused in a two classes problem and four binary predictors.
Data have been simulated considering two levels of separability of the classes (Low and High) and according to the Multinomial distribution, with the occurrence probabilities of the four predicting binary variables, presented in Table 5:

Table 5 Parameters of the Multinomial distribution used in data simulation

Separability	C_1	C_2
Low	(0,5;0,5;0,5;0,5; 0,5;0,5;0,5;0,5)	(0,5;0,5;0,5;0,5; 0,5;0,5;0,5;0,5)
High	(0,1;0,9;0,7;0,3; 0,2;0,8;0,6;0,4)	(0,9;0,1;0,3;0,7; 0,8;0,2;0,1;0,9)

Two other characteristics were considered for simulated data: Sample size (small ($n = 60$), moderate ($n = 120$) and large ($n = 400$); Balance (classes with equal size), classes with moderate unbalanced and with severe unbalanced) and 30 classification runs in each scenario were considered. To implement the regression models, the three complexity factors were used in order to study the impact of each one in the performance measure:

- Separability of classes: The Affinity coefficient ([1]), defined in the interval [0,1], is used to measure separability of classes;
- Balance: The ratio between the minority and the majority class sizes is used to measure balance;
- Sample size: The ratio between the ”number of degrees of freedom” and sample size is used to measure sample size importance.

Based on the 270 sets of generated data, the performance measures referred previously were obtained using a reference model in discrete supervised classification, the First-Order Independence Model (FOIM) ([4]), and were estimated by twofold cross-validation. For the performance measures that assume values in the standard unit interval (0,1) Beta regression models were used and the estimated coefficients were obtained using the Betareg R package ([7]); For the performance measures that assume values elsewhere, Linear regression models were used. The estimated regression models are presented in Tables 6, 7 and 8.

Table 6 Estimated coefficients for performance measures based on the confusion matrix

	Accuracy − Pseudo $R^2 = 0.88$				*Sensitivity* − Pseudo $R^2 = 0.38$			
	Estimate	St. Error	z	Sig.	Estimate	St. Error	z	Sig.
Intercept	2.43	0.08	29.36	$***$	2.30	0.21	10.91	$***$
Separability	**-3.21**	0.08	**-41.63**	$***$	**-3.23**	0.18	**-17.93**	$***$
Balance	0.15	0.05	3.13	$**$	0.38	0.12	3.19	$**$
Sample size	**0.95**	0.11	**8.78**	$***$	0.80	0.27	3.01	$**$
	Specificity− Pseudo $R^2 = 0.68$				*Precision* − Pseudo $R^2 = 0.73$			
	Estimate	St. Error	z	Sig.	Estimate	St. Error	z	Sig.
Intercept	2.78	0.12	22.83	$***$	0.40	0.13	6.77	$***$
Separability	**-3.26**	0.11	**-30.50**	$***$	**-3.05**	0.11	**-27.55**	$***$
Balance	0.07	0.07	1.05	0.29	**2.35**	0.08	**22.31**	$***$
Sample size	**0.61**	0.15	**4.00**	$***$	0.70	0.16	4.38	$***$

$**p < 0.01$; $***p < 0.001$

The estimated regression models exhibited an adequate to good fit to data and the three complexity measures impacts significantly in all evaluation measures.

Separability, measured by the Affinity coefficient, emerges as the most important experimental factor with a negative impact on performance. Sample size is the second most important factor for *Accuracy*, *Sensitivity*, *Specificity*, *Balanced accuracy* and *Geometric mean* with a positive impact on performance. Balance is the second most important factor, for *Precision*, *F measure*, *Phi coefficient*, *Cohen’s Kappa statistics* and *Huberty’s index*, also with a positive impact on performance.

Table 7 Estimated regression coefficients for combined performance measures

	Balanced accuracy – Pseudo $R^2 = 0.81$				*Geometric mean* – Pseudo $R^2 = 0.74$			
	Estimate	St. Error	z	Sig.	Estimate	St. Error	z	Sig.
Intercept	2.17	0.10	20.96	$***$	1.79	0.13	13.61	$***$
Separability	**-2.95**	0.09	**-32.60**	$***$	**-2.87**	0.11	**-25.00**	$***$
Balance	0.27	0.05	4.52	$***$	0.38	0.08	4.97	$***$
Sample size	**0.88**	0.13	**6.69**	$***$	**1.11**	0.17	**6.54**	$***$
	F measure – Pseudo $R^2 = 0.84$							
	Estimate	St. Error	z	Sig.				
Intercept	0.87	0.13	6.77	$***$				
Separability	**-2.92**	0.11	**-27.55**	$***$				
Balance	**1.74**	0.08	**22.31**	$***$				
Sample size	0.71	0.16	4.38	$***$				

$**p < 0.01$; $***p < 0.001$

Table 8 Estimated regression coefficients for less traditional performance measures

	Phi[1] – Pseudo $R^2 = 0.78$				*Kappa*[1] – Pseudo $R^2 = 0.80$			
	Estimate	St. Error	z	Sig.	Estimate	St. Error	z	Sig.
Intercept	0.84	0.02	39.15	$***$	0.59	0.04	16.45	$***$
Separability	**-0.89**	0.03	**-30.08**	$***$	**-0.91**	0.03	**-31.63**	$***$
Balance	**0.21**	0.02	**9.36**	$***$	**0.22**	0.02	**10.67**	$***$
Sample size	-0.01	0.00	-3.41	$**$	0.35	0.05	7.38	$***$
	Huberty's index[1] Pseudo $R^2 = 0.56$							
	Estimate	St. Error	z	Sig.				
Intercept	-0.33	0.20	-1.62	$***$				
Separability	**-1.90**	0.16	**-11.62**	$***$				
Balance	**1.80**	0.12	**14.99**	$***$				
Sample size	0.45	0.27	1.65	$***$				

[1] – Multiple linear regression
$**p < 0.01$; $***p < 0.001$

4 Conclusions

This study has revealed to be an interesting contribution to the goal of understanding how to choose an evaluation measure that really takes into account the classification problem at hand. Separability of classes emerges as the factor that really influences classifier performance: the weakly separated the classes are, the higher the affinity coefficient and the weaker the classification performance is. Note that the estimated weight of this factor in the regression models is always in the order of three points. The size of the samples and the balance between them also have an important impact on the quality of the classifier performance. Sample size is the second most important factor for all but two simple and combined measures (*Precision* and *F mea-*

sure): the larger samples size are the better classification performance is. Balance is the second most important factor important factor for the less traditional measures, *Precision* and the *F measure*: the more balanced classes are the stronger the classification performance is. Naturally, classification results improve as the classification problem becomes easier (better separability, bigger samples and classes more balanced).

Although not presented here due to being a short article, exploratory analysis with real data showed that with balanced classes, all performance measures show similar results; on the opposite, with low separability, was observed large differences between results of association or agreement measures and all the others. In the unbalanced case with high separability of classes, measures tend to be similar but with low separability, measures values are discrepant. Finally, let's note that the Huberty's index it's a very demanding but interesting measure, hardly reaching high values in real life problems.

The evaluation of results in Discrete Supervised Classification will continue to be further explored, using both simulated and real data, particularly in the case of unbalanced classes, with a view to better understanding the interest of other performance measures almost always absent in the literature of the area.

References

1. Bacelar-Nicolau, H.: The affinity coefficient in cluster analysis. Methods of Operations Research, **53**, 507–512 (1985)
2. Cohen, J.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement, **20**, 37–46 (1960)
3. Ferreira, A. S., Cardoso, M. G.: Evaluating Discriminant Analysis Results. In: Lita da Silva J., Caeiro F., Natário I., Braumann C. (eds.): Advances in Regression, Survival Analysis, Extreme Values, Markov Processes and and Other Statistical Applications. Studies in Theoretical and Applied Statistics, pp. 155-162. Springer, Heidelberg (2013)
4. Goldstein, M., Dillon, W.R.: Discrete Discriminant Analysis, Wiley and Sons, New York (1978)
5. Ho, T.K., Basu, M.: Complexity measures of supervised classification problems. IEEE Transactions on Pattern Analysis and Machine Intelligence, **24**, 289–300 (2002)
6. Huberty, C.J., Olejnik, S.: Applied MANOVA and Discriminant Analysis. Wiley-Interscience. John Wiley and Sons, New Jersey (2006)
7. Cribari-Neto, F., Zeileis, A.: Beta Regression in R. Journal of Statistical Software, **34(2)**, 1–24 (2010). http://www.jstatsoft.org/v34/i02/
8. Paik, H.: The effect of prior probability on skill in two-group discriminant analysis. Quality and Quantity, **32**,201–211 (1998)
9. Santafe, G., Inza, I., Lozano, J.A.: Dealing with the evaluation of supervised classification algorithms. Artificial Intelligence Review, **44(4)**, 467–508 (2015)