Validating Predictive Performance of Classifier Models for Multiclass Problem in Educational Data Mining

Ramaswami M

Department of Computer Applications
School of Information Technology
Madurai Kamaraj University
Madurai, Tamil Nadu
INDIA.

Abstract

Classification is one of the most frequently studied problems in data mining and machine learning research areas. It consists of predicting the value of a class attribute based on the values of other attributes. There are different classifications models were proposed in educational data mining (EDM) and it is used to evaluate student's academic performance in educational institutions and based on the results of the models, preventive measures to be taken in advance to enhance the students learning ability so that students' academic performance can be improved. The main objective of this study is to explore different predictive measures and assess the quality of predictive performance ability of the classifier models in educational data mining.

Keywords: *Overall Classification Rate, misclassification cost measure, ROC Measure, Volume Under ROC Surface, confusion matrix, Predictive Accuracy, classifier Performance.*

1. Introduction

Educational Data Mining (EDM) is a prominent interdisciplinary research domain that deals with the development of methods and models to explore the data originating in an educational context. EDM draws methods and theory from a number of disciplines, such as data mining, knowledge discovery, psychometrics, and statistical learning etc. It aims to contribute models and findings that can help design, develop and deployment of innovative learning applications and environments, as well as contributing to theory in educational psychology and other areas of education. EDM methods include classification, regression, factor analysis, clustering, relationship mining, knowledge prediction, correlation mining, association rule mining, visualization, domain structure discovery, discovery with models which leads to enhancement of students learning ability.

One of the potential areas of application of EDM is improvement of student models that would predict student's characteristics or academic performances in schools, colleges and other educational institutions. Prediction of student performance with high accuracy is useful in many contexts in all educational institutions for identifying slow learners and distinguishing students with low academic achievement or weak students who are likely to have low academic achievements. The end product of models would be beneficial to the teachers, parents and educational planners not only for informing the students during their study, whether their current behavior could be associated with positive and negative outcomes of the past, but also for providing advice to rectify problems. As the end products of the models would be presented regularly to students in a comprehensive form, these end products would facilitate reflection and self-regulation during their study.

2. Classifier Performance Measures

A classifier performance is a single index [1] that measures the goodness of the classifiers considered. Depending on the design / requirements, different problems may require different performance measures to ensure that the classifiers considered shall be compared properly and selected. To discover the subtle performance difference between one model and another, the performance measure used for classifier evaluation needs to better address the accuracy of the classifier performance. Student performance prediction models are used to predict the performance of the student based on some underlying factors that are given as input. In other words, the classifier model should classify a student into most appropriate class ***(pass, fail***) into which they actually belongs. But practically, most of the classifier model may predict incorrectly into another class, instead of actual class and it is referred as misclassification. Therefore, classifier evaluation should take account the different classifiers that have different misclassification cost for each fault prediction.

The most common measure used in classifier performance is the *overall classification rate*. The overall classification rate also called *predictive accuracy* is defined as the ratio of number of students that are correctly classified over the total number of students. Mathematically, let *CM* be an $M \times M$ confusion matrix, then the overall classification rate (*OCR*) is defined as

$$OCR = \frac{1}{N}\sum_{i=1}^{M} CM(i,i)$$

where *M* is the total number of classes and *N* is the total number of cases.

This type of performance measure can be calculated easily and is most ideal for all kinds of classifiers. The underlying assumption of the *OCR*, however, is that the classification errors for all classes have equal cost consequences. This assumption rarely meets the situation, as most of the real world problems are with unequal size of class distribution. Therefore the overall classification rate is often not an appropriate measure of the classifier performance [4]. The limitations of the overall classification rate as a performance measure include that it is sensitive to the unequal class size and then it does not reveal the performance of the classifier across the entire range of possible decision thresholds [6].

Breiman et al [7] have made OCR measure as useable by means of stratifying the classes based on the target cost and class distribution so that maximizing accuracy on the transformed data corresponds to minimizing costs on the target data. However, this strategy fits only to *two-class* problems and requires precise "true" class distribution, which is not ideal for most of the real-world problems. Alternatively, most of the researcher uses Receiver Operating Characteristics (*ROC measure)* for evaluating classifier performance. It is a well-established method for evaluating classifier performance in many fields. Originated from the field of signal detection to depict tradeoff between hot rate and false alarm rate [9], it prevail the most frequently used measure for evaluating classifier performance for *two-class* classifiers.

ROC curves are a valuable technique for visualizing classifier behavior over a range of decision rules. The ROC curve can be drawn by plotting true positive rate (*TPR*) on Y-axis and plotting false negative rate (*FPR*) on X-axis. Classifiers with high ROC value located in the upper-left corner of ROC curve are better. This is because of the fact that classifiers that have lower false positive rate and higher true positive rate than classifiers below them. The limitation of ROC analysis is that this measure will be confined to *two-class* problems only. This drawback limits the ROC analysis for much wider applications. The extended form of ROC curve is *Volume Under ROC Surface* (*VUS*), which is an alternative measure for evaluating multi-class classifiers. Only limited research articles are available on VUS. Due to elusiveness of its precise definition and complexity of calculation [5], it is not a widely acceptable method for evaluating performance of classifiers for multi-class problems.

To overcome these problems, an alternative measure called *misclassification cost measure* (*MCM*) suggested by Michie, et al [11] used as a general classifier performance measure for evaluating performance of *multi-class* classifier models. The misclassification cost is defined as the product of each element of the *normalized confusion matrix* (*NCM*) and the corresponding element of the cost matrix and summing the results, as follows

$$MCM = \sum_{i,j} \overline{cm}(i,j).C(i,j)$$

where $\overline{cm}(i,j) = CM(i,j)/\sum CM(i)$ is the normalized confusion matrix.

The misclassification cost (*MCM*) has been used by Yan et al.,[1][12] for designing cost-sensitive classifiers. Moreover, it is noted that, overall accuracy or OCR is a special case of the misclassification cost. When the cost matrix has a value of 1 on its diagonal elements and zeros on all off-diagonal elements, the misclassification cost becomes predictive accuracy of the classifier. Therefore, the misclassification cost measure is a general form of the accuracy measure. The most appealing merits of the misclassification cost measure are that it can be used for *multi-class* classifiers and take care of classifiers with different costs for different classes through proper definition of cost matrix.

The cost matrix is a matrix, where each element *C(i, j)* represents the cost incurred for misclassification of object in class ***i*** into class ***j***. Based on this information, it is noted that all diagonal elements of a cost matrix should have zero value. Moreover, different misclassification cost has different consequence on the problem domain.

For example, in student performance prediction model, misclassifying a student with "*excellent"* grade into "*fail"* is more critical than classifying "*excellent"* grade in to "*very good"* grade. Therefore, misclassifying high-achievers into low-achievers should have different cost consequence from misclassifying high-achievers into average-achievers. Capturing this difference into performance measure is the key for better evaluation of the classifier performance. Due to variation of the misclassification cost, the full cost matrix becomes a non-symmetric matrix.

The full cost matrix has to be constructed with the following two basic assumptions:

a). the cost of misclassifying i^{th} grade as j^{th} grade is different from that of misclassifying j^{th} grade as i^{th} grade if ***i*** and ***j*** are different.

b). the cost of misclassifying i^{th} grade as j^{th} grade is higher if ordered ranking of j^{th} grade is further away from that of i^{th} grade.

Based on this cost measure, the performance of the different classifiers has been evaluated by varying the number of cases of class variable *HScGrade*. For example, Table 1 shows the typical (fixed by user) ranking or penalty for ***n*** cases of grades of the class variable *HScGrade*.

Table 1: Grade list and Ranking

Grades	G_1	G_2	G_3	G_4	G_5	…	G_n
Ranking	R_1	R_2	R_3	R_4	R_5	…	R_n

Each R_i is the grade ranking for i^{th} grade and we define $d_{ij} = R_i - R_j$ as the distance measures, i.e., how far apart the two grades are in the ranking. The defined distance also represents the degree of misclassification when i^{th} grade is misclassified as j^{th} grade. Similar to confusion matrices, distance or degree of misclassification between each pair of grades can be represented as a matrix as shown in Table 2.

Based on the definition of d_{ij}, the value of d_{ij} can be either positive or negative. While a positive value of d_{ij} means that ranking for i^{th} grade is higher than that for j^{th} grade. Intuitively, the matrix representing the degree of misclassification should be directly related to the misclassification cost matrix.

Table 2: Matrix representing degree of misclassification Cost

		Predicted Grade			
		G_1	G_2	..	G_n
True Grade	G_1	*0*	d_{12}		d_{1n}
	G_2	d_{21}	*0*		d_{2n}
	..				
	G_n	d_{n1}	d_{n2}		*0*

Therefore, we compute the cost matrix C_{ij}, in terms of degree of misclassification D_{ij} as follows:

$$C_{ij} = \frac{d_{ij}}{\sum_i R} S \quad \text{for} \quad d_{ij} > 0 \text{ and}$$

$$C_{ij} = -m \frac{d_{ij}}{\sum_i R} S \quad \text{for} \quad d_{ij} \le 0$$

where $\sum R$ in the denominator is the sum of the values of the rankings and is used for normalization purpose. The factor m, $\{ m \le 1\}$ used for $d_{ij} < 0$ case in the equation captures the notion that misclassifying a higher grade as a lower grade is less costly than misclassifying higher grade as average grade. For classifier performance evaluation, only relative values of the cost matrix matter, i.e., scaling a cost matrix with a constant will not change the classifier evaluation results. Therefore, the relationship between C_{ij} and d_{ij} is unique but can be scaled. The particular scaling is performed with the domain-specific constant scaling parameter, *S*.

3. Penalty method

Percentage of accuracy is generally not preferred for classification, as values of accuracy are highly dependent on the base rates of different classes. For assessing the goodness of a predictor, an extensive study on the student data set was conducted by applying five individual classifiers J48 (J48), Bayesian Net (BN), Neural Net (NN), Decision Tree (DT), and Naïve Bayes (NB), are used in this study. These classifiers were chosen based on their reasonable performance in our preliminary study under student performance classification [3]. The performance of these classifiers can be compared in terms of their predictive accuracy against with misclassification cost measure (MCM). The outcome of this study leads to recommendation of ideal classifier for student performance prediction model in EDM. These five classifiers were used to design the student prediction models under *multi-class* class variable – *HScGrade*. *HScGrade* is declared as response variable indicates Marks/Grade obtained at higher secondary level in Tamil Nadu, India and outcome of the class variable is defined as *five-case* class variable with values *excellent, very-good, good, fair,* and *poor*. Group them into five classes, "*excellent*" for students who secured 90% marks and above, "*very-good*" for students who got marks between 75% - 90%, "*good*" for marks between 60% - 75%, "*fair*" for marks between 40% - 60% and "*fail*" for other cases.

All experiments reported in this study were conducted by using the *WEKA* [2][10] that facilitates all data mining

techniques. To access the predictive performances of five classifiers, a 10-fold cross-validation [8] was applied to each configuration. The performance evaluation of these five classifiers was carried out for five-class student data with the following possible outcome of the classifier are ("*excellent*", **"***very-good*", "*good*", "*fair*" and "*fail*").

Alternatively, the performance of these five classifiers was assessed through misclassification cost measure. The relative ranking for five-class problem was fixed as shown in Table 3 and its associated cost matrix for three-class has been given in Table 4. Heavy penalty was fixed for misclassification of "*excellent* "class into "*fail*" class.

Table 3: Relative Result Ranking for Five-Class

Results	excellent (90% and above)	very-good (75% and above)	good (60% and above)	fair (40% and above)	fail (less than 40% of mark)
Ranking	0.0	0.1	0.2	0.3	0.9

Table 4: Matrix representing Degree of Misclassification for Five-Class

			Predicted Results				
			excellent	very-good	good	fair	fail
True Results	excellent	0.0	0.0	-0.1	-0.2	-0.3	-0.9
	very-good	0.1	0.1	0.0	-0.1	-0.2	-0.8
	good	0.2	0.2	0.1	0	-0.1	0
	fair	0.3	0.3	0.2	0.1	0.0	-0.6
	fail	0.9	0.9	0.8	0.7	0.6	0.0

The final cost matrix for five-class problem was obtained from the degree of misclassification with m = 0.9 and S = 100 and it has been shown in Table 5.

Table 5: Cost Matrix for Five-Class

		Predicted Results				
		excellent	very-good	good	fair	fail
True Results	excellent	0	2	4	6	18
	very-good	3.33	0	2	4	16
	good	6.67	3.33	0	2	0
	fair	10	6.67	3.33	0	12
	fail	30	26.67	23.33	20	0

Table 6 shows the performance results of five classifiers against Full Subset (FSS), Correlation based (CFS), Consistency-Subset (CSS), CHI-Square (CHI), Gain Ratio (GAR) and Information Gain (ING) feature evaluation methods. The performance results of these classifiers showed that the rank value of both cost measure and predictive measures in filter-based approach were quit similar for MLP and J48 classifiers.

Table 6: Performance Evaluation Results of Filter-Based Five-Class Classifiers

Classifiers	Based on Misclassification Cost Measure		Based on Accuracy Measure	
	Cost	Ranking	Accuracy	Ranking
Bayes-CFS	25.54592	18	49.1025	17
Bayes-CHI	27.06650	21	47.4629	19
Bayes-CSS	27.59583	22	49.0162	18
Bayes-FSS	24.51467	15	42.7511	21
Bayes-GAR	29.30358	24	47.4629	19
Bayes-ING	29.30358	24	47.4629	19
DT-CFS	27.87417	23	49.4477	16
DT-CHI	24.51467	15	51.6741	14
DT-CSS	25.60515	19	49.7929	15
DT-FSS	24.43254	13	52.8133	12
DT-GAR	24.05142	11	51.9676	13
DT-ING	24.51467	15	51.6741	14
J48-CFS	24.06144	12	54.591	11
J48-CHI	15.66173	9	68.4674	9
J48-CSS	15.43349	7	70.8146	6
J48-FSS	15.13625	5	71.2806	5
J48-GAR	15.33592	6	68.5537	7
J48-ING	15.65809	8	68.4846	8
Naive-CFS	26.83961	20	44.6151	20
Naive-CHI	24.69793	17	40.3003	24
Naive-CSS	25.23449	18	41.8882	22
Naive-FSS	24.55009	16	39.5927	25
Naive-GAR	24.49796	14	41.0079	23
Naive-ING	24.69793	17	40.3003	24
MLP-CFS	21.82812	10	59.7169	10
MLP-CHI	11.84857	4	81.6362	4
MLP-CSS	9.863847	2	85.951	2
MLP-FSS	4.338674	1	92.7166	1
MLP-GAR	10.03112	3	82.6717	3
MLP-ING	10.03112	3	82.6717	3

The predictive performance of the five machine learning algorithms against diverse filter-based feature subsets with different cardinalities derived from five feature selection methods were evaluated. Filter based subset selection method have high impact on the predictive accuracy of the five machine learning algorithms, in particular, Neural Net and Decision-Tree (C4.5) algorithms could yield high predictive accuracy. Also the feature evaluation methods CHI and ING were significantly dominating other feature evaluation methods. The results of the predictive accuracy of the machine leaning algorithms further justifies using misclassification cost measure, which confirmed that, both Neural Net and Decision-Tree algorithms were best suited for student performance prediction model for the higher secondary students.

4. Conclusion

An extensive evaluation of five classifiers with different configurations settings was carried out and it was observed that the predictive accuracy of the classifiers ranged from

40% to 92% for five-class class variable. In addition, it was also observed that the Decision Tree and Neural network models showed better performance based on predictive accuracy as well as misclassification cost measure. In examining the problem of prediction of performance with this penalty method, it is possible to automatically select best classifier models to predict students' performance. The outcome of this study leads to recommendation of ideal classifier for student performance prediction model in EDM.

Acknowledgments

Author take this opportunity to express a deep sense of gratitude to University Grants Commission(UGC), New Delhi, India for their financial support through UGC Minor Project F.No.41-1353/2012(SR).

References

[1] Yan, W., Goebel, K. and Li, J. C.(2000), "*Classifier performance measures in multi-Fault Diagnostics for Aircraft Engines*," Proceeding of SPIE component and systems Diagnostics Prognostics and Health Management II," V4733, 88-97.

[2] Weka 3.5.6.(2009), "*An open source data mining software tool developed at university of Waikato, New Zealand*," downloaded from ***http://www.cs.waikato.ac.nz/ml/weka/*.**

[3]. Ramaswami, M. and Bhaskaran, R.(2010), " *A Effect of Feature Selection Techniques in Educational Data Mining* " Journal of Computing 1(1), 7- 11.

[4] Provost, F. and Fawcett, T. (1997), "*Analysis and visualization of classifier performance: Comparison under imprecise class and cost distributions*," Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97), pp. 43-48.

[5] Ferri, C., Hernández-orallo, J. and Salido, M. A. (2003), "*Volume Under the ROC Surface for Multi-class Problems-Exact Computation and Evaluation of Approximations*," Proc. of 14th European Conference on Machine Learning, pp. 108-120.

[6] Downey, T. J., Meyer, D.J., Price, R.K. and Spitznagel, E. L. (1999),"*Using the receiver operating characteristic to asses the performance of neural classifiers*," IJCNN'99-International Joint Conference on Neural Networks 5, 3642-3646.

[7] Breiman, L., Friedman, J. H., Olshen, R.A. and Stone, C.J. (1984), "*Classification and regression trees*," Chapman and Hall/CRC, Florida.

[8] Hastie, T., Tibshirani, R. and Friedman, J. (2001), "*The Elements of Statistical Learning: Data Mining, Inference and Prediction*," Springer-Verlag, New York, USA.

[9] Bradley, A. P. (1997), "*The use of the area under the ROC curve in the evaluation of machine learning algorithms*," Pattern Recognition, 30(7), 1145-1159.

[10] Witten, I. and Frank, E.(2005), "*Data Mining – Practical Machine Learning Tools and Techniques*," Morgan Kaufmann.

[11] Michie, D., Spiegelhalter, D.J.& Taylor, C.C (Eds.)(1994), "*Machine Learning, Neural and Statistical Classification*," Ellis Horwood, New York, NY.

[12] Margineantu, D.D. and Dietterich, T.G.(2000), "*Bootstrap methods for the cost-sensitive evaluation of classifiers*," Proceedings of International Conference on Machine Learning (ICML-2000), pp. 583-590.