Theory and Applications of Mathematics & Computer Science

(ISSN 2067-2764, EISSN 2247-6202)
http://www.uav.ro/applications/se/journal/index.php/tamcs

Theory and Applications of Mathematics & Computer Science 7 (1) (2017) 39 – 46

Evaluation of Classification Models in Machine Learning

Jasmina Dj. Novaković[a,*], Alempije Veljović[b], Siniša S. Ilić[c], Željko Papić[d], Milica Tomović[e]

[a]*Belgrade business school, Higher education institution for applied science, Belgrade, Serbia*
[b]*Faculty of technical sciences Čačak, University of Kragujevac, Čačak, Serbia*
[c]*Faculty of technical sciences K. Mitrovica, University of Priština, K. Mitrovica, Serbia*
[d]*Faculty of technical sciences Čačak, University of Kragujevac, Čačak, Serbia*
[e]*PE Post of Serbia, Belgrade, Serbia*

Abstract

We study the problem of evaluation of different classification models that are used in machine learning. The reason of the model evaluation is to find the optimal solution from various classification models generated in an iterated and complex model building process. Depending on the method of observing, there are different measures for evaluation the performance of the model. To evaluate classification models the most direct criterion that can be measured quantitatively is the classification accuracy. The main disadvantages of accuracy as a measure for evaluation are as follows: neglects the differences between the types of errors and it dependent on the distribution of class in the dataset. In this paper we discussed selection of the most appropriate measures depends on the characteristics of the problem and the various ways it can be implemented.

Keywords: accuracy, confusikon matrix, costs of misclassification, F-measure, ROC graph.
2010 MSC: 68T01, 68T05.

1. Introduction

Machine learning is a field of artificial intelligence that deals with the construction of adaptive computing systems that are able to improve their performance by using information from experience. Machine learning is the discipline that studies the generalization and construction and analysis of algorithms that can generalize. But as much as the applications of machine learning were diverse, there are tasks that are repetitive. Therefore, it is possible to talk about the types of learning tasks that often occur. One of the most common tasks of learning that occurs in practice is classification. Classification is an important recognition of object types, for example whether a particular tissue represents malignant tissue or not.

*Corresponding author
Email addresses: jnovakovic@sbb.rs (Jasmina Dj. Novaković), alempije@beotel.net (Alempije Veljović), sinisa.ilic@pr.ac.rs (Siniša S. Ilić)

Classification is one of the most common tasks of machine learning, and is a problem of classification unknown instance in one of the pre-offered categories - classes. The important observation in classification is that target functions are discrete. In general, the class label can't be meaningfully assigned numerical or some other values. This means that the class attribute, whose value should be determined, categorical attribute.

The classification of an object is based on finding similarities with predetermined objects that are members of different classes, with the similarity of the two objects is determined by analyzing their characteristics. In classifying every object is classified into one of the classes with certain accuracy. The task is that on the characteristics of objects whose classification is known in advance, make a model by which will be performed classification of new objects (Fawcett, 2003; Marzban, 2004; Vardhan *et al.*, 2012). In problem of classification, the number of classes is known in advance and limited.

A wide range of algorithms for classification is available, each with their own strengths and weaknesses. There is no such a learning algorithm which works best with all the problems of supervised learning. Machine learning involves a large number of algorithms such as: artificial neural networks, genetic algorithms, rule induction, decision trees, statistical and pattern recognition methods, k-nearest neighbors, Naïve Bayes classifiers and discriminatory analysis.

The main objective of this paper is to discuss the various classification models that can be used in the problem of classification. This paper presents the advantages and disadvantages of these models. For this purpose we have organized the paper in the following way. In the second part of this paper we present evaluation of classification models, in the third part of the paper we present measures for the evaluation of classification models. In the last part of the paper, we discuss the results and give directions for further research.

2. Evaluation of classification models

For modeling regularity in the data there are a number of methods. Also, methods on the same set of examples for learning result in different models changing the parameters of the method. Due to the same problem and the same set of training data can produce a higher number of different models; it emphasizes the need for the evaluation of the quality model with respect to the given problem. That is why the evaluation of discovered knowledge is one of the essential components of the process of intelligent data analysis. Since this work deals with the classification problems, hereinafter will discuss the evaluation of classification models.

The task of evaluating classification models is to measure the degree to which the classification suggested using the model corresponding to the actual classification of the case. Depending on the method of observing, there are different measures for evaluation the performance of the model. Selection of the most appropriate measures shall be done depending on the char

acteristics of the problem and ways of its implementation.

3. Measures for the evaluation of classification models

In the evaluation of classification models basic concept is the notion of fault. If the application of the classification models in selected case leading to the prediction of a class that is different from

the actual class examples then there is an error in classification. If any mistake is equally important, then the total number of errors in the observed set can be an indicator of work a classifier.

This approach is based on accuracy as a measure for evaluating the quality of the classification model. This measure can be defined as the ratio of the number of correctly classified examples according to the total number of classified examples.

$$Accuracy = \frac{number\ of\ correctly\ classified\ examples}{total\ number\ of\ cases} \tag{3.1}$$

The main disadvantages of accuracy as a measure for evaluation are as follows: (1) neglects the differences between the types of errors; (2) dependent on the distribution of class in the dataset.

It is often important in practical problem solving distinguish certain types of errors. It is often the case in medicine, for example detecting the existence of disease in a patient. If system needs to classify breast tissue on malignant and benign based on mammography image, then if the system incorrectly marked diseased tissue as healthy tissue, the error is more important, because it will not notice the existence of the disease and will not apply the appropriate therapy. In case that the system recognizes healthy tissue as sick, error has less importance because it will further surgery and diagnosis to determine that the patient is not diseased.

In cases where it is necessary to distinguish more types of errors result of the classification is shown in the form of two-dimensional matrix, where each row of the matrix corresponds to one class and record number of examples where it is forecasted class, and each column of the matrix is also marked by a class and x h

+ c+aald number of examples where it is an actual class. lllvvvv mI f vf we look for example classification problem with five classes, where we need to classify the emotional state of the person appearing in the video in five different emotional categories: happy, sad, angry, gentle and frightened, then we confusion matrix display as in Figure 1.

		Actual class				
		happy	sad	angry	gentle	frightened
Predicted class	happy	51	2	1	1	1
	sad	3	23	1	1	0
	angry	2	2	17	0	0
	gentle	0	1	2	9	1
	frightened	1	0	1	1	18

Figure 1. Illustration of confusion matrix for the classification problem of recognizing emotional states.

On the diagonal of the matrix is the number of correct classified examples, while other elements of the matrix indicate the number of examples that were incorrectly classified as some of the other classes. Figure 1 shows that the six examples of class *happy* wrongly classified as follows: three are classified as class *sad*, two in class *angry*, zero in class *gentle*, and one in class *frightened*. It can be concluded that the use of a confusion matrix allows better analysis of different types of errors.

The largest number of measures for evaluation of classification models related to classification problems with two classes. This is not a particular limitation for the use of these measures, given

that problems with larger number of classes can be displayed as a series of problems with two classes. Each of these measures in particular stands out one of the class as a target class, with the data set is divided into positive and negative examples of the target class. The negative examples include examples of all other classes. That is why below we consider a classification problem with two classes.

Confusion matrix in classification problem with two classes is shown in Figure 2. It can be concluded from the figure that there are possible four different results forecasts. Really positive and really negative outcomes are correct classification, while the false positive and false negative outcomes are two possible types of errors.

False positive example is a negative example class that is wrongly classified as positive and false negative is a positive example of the class who is wrongly classified as negative. In the context of our research entrance to confusion matrix have the following meanings (Kohavi & Provost, 1998):

- a is the number of correct predictions that instances are negative,
- b is the number of incorrect predictions that instances are positive,
- c is the number of incorrect predictions that instances are negative,
- d is the number of correct predictions that instances are positive.

		Predicted class	
		Negatives	Positives
Actual class	Negatives	a	b
	Positives	c	d

Figure 2. Confusion matrix in classification problem with two classes.

A few standard terms are defined in a matrix with two classes: accuracy, true positive rate, false positive rate, true negative rate, false negative rate and precision. The accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined. Accuracy may be determined using the equation:

$$Accuracy = \frac{a+b}{a+b+c+d}. \tag{3.2}$$

True positive rate is the proportion of positive cases that are properly identified and can be calculated using equation:

$$True\ positive\ rate = \frac{d}{c+d}. \tag{3.3}$$

The false positive rate is the proportion of negative cases that were incorrectly classified as positive, and calculated with equation:

$$False\ positive\ rate = \frac{b}{a+b}. \tag{3.4}$$

The true negative rate was defined as the proportion of negatives cases which are classified correctly, and is calculated using the equation:

$$True\ negative\ rate = \frac{a}{a+b}. \tag{3.5}$$

The false negative rate is the proportion of positive cases that were incorrectly classified as negative, and are calculated using equation:

$$False\ negative\ rate = \frac{c}{c+d}. \tag{3.6}$$

Finally, precision or positive predictive value presents the fraction of predictive positive cases that are accurate, and is calculated using the equation:

$$Precision = \frac{d}{b+d}. \tag{3.7}$$

There are cases when accuracy is not adequate measures. The accuracy is determined by the equation (3.2) can't be an adequate measure of performance when the number of negative cases is much higher than the number of positive cases (Kubat *et al.*, 1998). If there are two classes and one is significantly smaller than the other, it is possible to obtain high accuracy if all instances are classified in larger class.

Suppose that there are 1000 cases, 995 negative cases and five cases which are positive. If the system classifies all of them negative, accuracy will be 99.5%, although classifier missed all positive cases. Or, for example, in tests which establish whether the patient is suffering from some disease, and the disease has only 1% of people in the population, a test should always reported that the patient has no disease would have an accuracy of 99%, but is unusable. In such cases, the accuracy as a measure of model quality is not adequate measure. In these cases the sensitivity of the classifier is an important measure and his ability to observe instances that are required, in this case ill patients.

In machine learning, most classifiers assumes equal importance of classes in terms of the number of instances and the level of importance, which means that all classes have the same significance. Standard techniques in machine learning are not successful when predicting a minority class in an unbalanced data set or when the false negatives are considered more important than false positives. In practical terms, unequal costs of inaccurate classifications are common, especially in medical diagnostics, so that the asymmetric misclassification costs must be taken into account as an important factor.

Cost-sensitive classifiers adapting models to costs of misclassification in the learning phase, with the objectives to reduce the costs of misclassification rather than to maximize the accuracy of classification. Because many practical problems of classifications have different costs associated with different types of errors, various algorithms for the evaluation of the sensitivity of classification is used.

Complementarity is one of the important characteristics of the evaluation of classification models. Using the pairs measures can be displayed specific accuracy of classification models with somewhat opposed positions. For example, by varying the parameter selected modeling techniques can be at the expense of one of the specific measures to increase the accuracy of the model

shown in another measure. This is an optimization problem in which the selection with the appropriate settings based on the one measure, maximize other measures. In some cases, the quality of the classifier needs expressed by a number, not a pair of dependent measures, which is achieved by using pairs measures. Using the pairs value of measures, one measure is fixed and is observed only second measure. Thus, for example, can be considered measures of accuracy with fixed value of the response to 20% and in this case the derived measure is called the precision of 20%.

Besides derived measure, there are measures that are not based on fixing one component of a pair of measure, for example *F-measure*, which is defined as follows:

$$F\text{-}measure = \frac{2 \times response \times accuracy}{response + accuracy}. \tag{3.8}$$

Another way to test the performance of the classifier is the ROC graph (Swets, 1988). ROC graph is the two-dimensional representation which on the X axis represents the false positive rate and the Y axis represents true positive rate. Item (0,1) is the perfect classifier: classifies all positive and all negative cases correctly. This is (0, 1), because the false positive rate is 0 (zero), a positive real rate is 1 (all). Point (0, 0) is a classifier that predicts all cases to be negative, while point (1, 1) corresponds to the classifier which provides that every case is positive. Point (1, 0) is a classifier that is incorrect for all classifications. In many cases, the classifier has a parameter which can be adjusted increasing the real positive rates at the cost of increasing false positive rates or reducing the false positive rate based on the dropping value of real positive rates.

Each setting parameters gives par value for a false positive rate and positive real rates and the number of such pairs can be used to represent the ROC curves. Nonparametric classifier is presented ROC to one point, which corresponds to the par value of the false positive rate and positive real rate.

Figure 3 shows an example of a ROC graph with two ROC curves and two ROC points marked P1 and P2. Nonparametric algorithms produce a single ROC point for a particular data set. Characteristics of ROC graph are:

- ROC curve or point is independent of the distribution of the class or the cost of errors (Kohavi & Provost, 1998).
- ROC graph contains all the information contained in the matrix of errors (Swets, 1988).
- ROC curve provides a visual tool for testing the ability of the classifier to correctly identify positive cases and negative cases that were incorrectly classified.

The area under of the one ROC curve can be used as a measure of accuracy in many applications, and it is called the measurement accuracy based on the surface (Swets, 1988).

Prevost and Fawcett in 1997 (Provost & Fawcett, 1997) argued that the use of the classification accuracy of the classifier comparison is not adequate measure unless the cost classification and distribution of class unknown, but one classifier must be chosen for each situation. They propose a method of assessing the classifier using the ROC graph, imprecise costs and distribution of class.

Another way of comparing ROC points is the equation that balances accuracy with Euclidean distances from perfect classifier, i.e. from the point (0, 1) on the graph. In this way we include

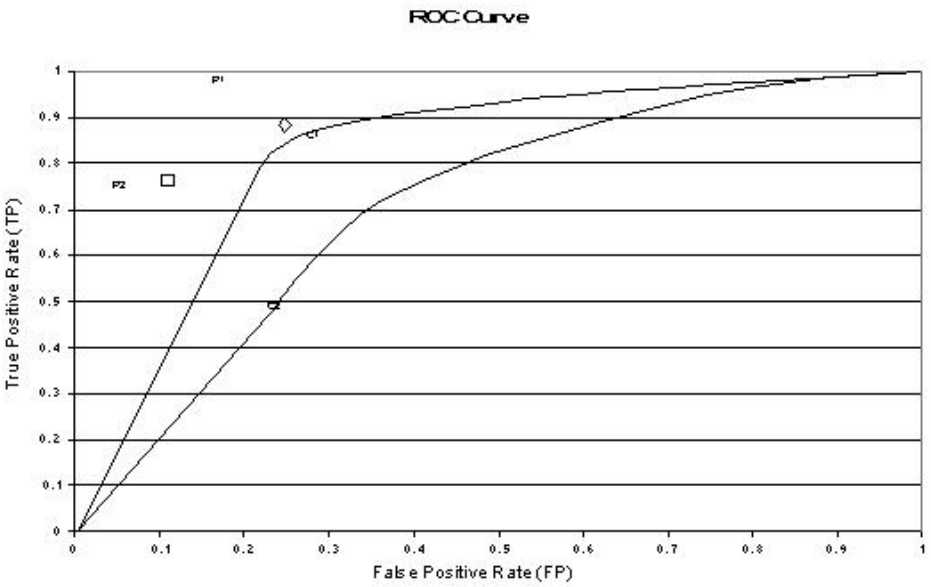

Figure 3. ROC graph
http://www2.cs.uregina.ca/˜dbd/cs831/notes/ROC/ROC.html

weighting factors that allow us to define the relative cost of improper classification, if such data are available.

4. Conclusions

This research discusses the various classification models that can be used in the problem of classification. This research could help in future works, such as the implementation of an adequate classification model in different classification problems. There are many questions and issues that remain to be addressed and that we intend to investigate in future work. These conclusions and recommendations will be used in classification problems in the near future.

Acknowledgements. The authors are grateful to the Ministry of Science and Technological Development of the Republic of Serbia for the support (projects: TR 34009 and TR35026).

References

Fawcett, T. (2003). ROC graphs: Notes and practical considerations for data mining researchers. Technical Report HPL-2003-4. Hewlett Packard, Palo Alto, CA.

Kohavi, R. and F. Provost (1998). *Glossary of terms*. Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process.

Kubat, M., R. Holte and S. Matwin (1998). Machine learning for the detection of oil spills in satellite radar images. *Machine Learning* **30**, 195–215.

Marzban, C. (2004). The ROC curve and the area under it as performance measures. *Wea. Forecasting* **19**(6), 1106–1114.

Provost, F. and T. Fawcett (1997). Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In: *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining (KDD-97)*.

Swets, J. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285–1293.

Vardhan, R. V., S. Pundir and G. Sameera (2012). Estimation of area under the ROC curve using exponential and weibull distributions. *Bonfring International Journal of Data Mining* **2**(2), 52–56.