

PERFORMANCE EVALUATION OF THE DATA MINING CLASSIFICATION METHODS

CRISTINA OPREA

UNIV. ASSISTANT, PHD. STUDENT,
PETROLEUM-GAS UNIVERSITY, PLOIEȘTI
oprea_cris2005@yahoo.com

Abstract

The paper aims to analyze how the performance evaluation of different classification models from data mining process. Classification is the most widely used data mining technique of supervised learning. This is the process of identifying a set of features and templates that describe the data classes or concepts. We applied various classification algorithms on different data sets to streamline and improve the algorithm performance.

Keywords: classification, mining techniques, algorithms, cost-sensitive, ROC curve.

JEL: C38, C8, C13

1. Introduction and context of the study

Data mining is the science that uses computational techniques from statistics, machine learning and pattern recognition to analyze large data sets or databases. This type of analysis has two objectives: predictive and descriptive. Thus distinguishes two types of data mining methods, as follows:

Predictive methods that use some variables to predict one or more other variables. Aceste metode sunt: clasificarea, regresia, detectia anomalilor.

Descriptive method that identifies patterns that describe the data. These methods are: clustering, association rules discovery and sequential patterns discovery.

In this paper we will focus on the method of classification, performance evaluation of different classification models. Performance evaluation of classification model is important for understanding the quality of the model, to refine the model, and for choosing the adequate model. The performance evaluation criteria used in classification models are: confusion matrix and receiver operating curves (ROC).

A study [3] based on a similar approach was presented in the specialized literature where the performance of fourteen classification algorithm is evaluated, using three distinct datasets. It has been shown that in terms of accuracy, any of the classifiers do not exceed on the others when they are applied on the three datasets. Most algorithms have performed better on data sets larger.

Another study [1] evaluates the performance of data mining algorithms, J48, Naive Bayes, Multilayer Perceptron, IBK and Bayes Net using Mc Nemar test. The results showed that the Multilayer Perceptron produced better results than the other algorithms while the Bayes algorithm has recorded the lowest performances. It was also found that the Mc Nemar test results largely coincide with Kappa statistics and RMSE as a confirmation of the integrity of the method used.

2. Data mining classification

Classification technique, as predictive method is a machine learning technique, supervised, assuming the existence of a group of labeled instances for each class of objects. The classification process is characterized by [4]:

- *Input*, a set of attributes with instances, including a class attribute, predictable;
- *Classifier* is used to predict the class of new instances;
- *Output*, a pattern classifier that classifies the instance in a certain category based on the other attributes.

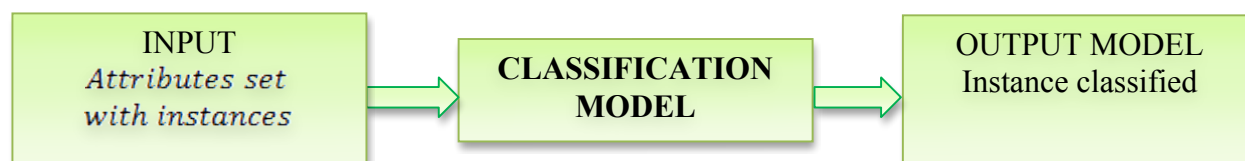


Fig. no. 1 The process of classification

In the table 1 lists the classification methods presented in the literature and their corresponding algorithms. Among these algorithms, in the study presented in this article we have chosen to analyze ID3, Naive Bayes, Multy-Layer Perceptron and K-NN.

Tabel no. 1 Classification methods and algorithms

Classification methods	Algorithms
– Classification and decision trees – Bayesian classifiers – Artificial neural network – K-nearest neighbor classifier – Regresion – Classifiers based on association rules – Rough set	– ID3, C4.5 și C5.0, CART, SPRINT, THAID, CHAID – Naive Bayes, BayesNet – Single-Layer Perceptron, Multy-Layer Perceptron, RBFNetwork, SVM – K-NN, PEBLS – Linear Regresion, SimpleLogistic – RIPPER, CN2, Holte’s 1R, C4.5

ID3 algorithm (Iterative Dichotomizer 3), predecessor of algorithm C 4.5, is an inductive technique of artificial intelligence that generates classification trees. These trees are similar to those used in simple expert systems. Unlike many other existing statistical techniques, classification tree model is a plausible inductive representation of human knowledge as it is on the same line of matching the cognitive constraints of the human brain [7].

Nayve Bayes technique is a classification method with potential both predictive and descriptive. It allows the analysis of the relationship between each independent variable and the dependent variable by calculating a conditional probability for each of these relationships. When a new instance is intended to be classified, the prediction is achieved by combining the effects of independent variables on the dependent variable. [6]

A multilayer perceptron is a feedforward artificial neural network consists of one input layer, one or multiple hidden layers and one output layer. Except the input nodes, each of nodes is a neuron with a activation nonlinear function.

K-Nearest Neighbor is a classification method based on learning by analogy, namely the search for solutions obtained for similar problems. This method classifies a new object / case based on the nearest (k) objects / cases in the neighborhood [4]

3. Evaluation of a Classification Model

Evaluation of classification algorithms is one of the key points in any process of data mining. The most commonly tools used in analyzing the results of classification algorithms applied are: confusion matrix, learning curves and receiver operating curves (ROC).

The confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the test data. Confusion matrix for a classifier with two classes True and False is presented in Table 2.

Table no 2 The confusion matrix of a classifier with two classes

		Classes predicted	
Current classes		True class	False class
	True class	True Pozitive	Fals Pozitive
	False class	True Negative	Fals Negative

The number of correctly predicted values relative to the total number of predicted values specified by Precision parameter that takes values between 0 and 1. Precision equal to 0 indicates that the model has no predictive power, it is not conclusive

True positive Rate (TP Rate) is the fraction of positive cases predicted as positive and is equivalent to the Recall.

False positive Rate (FP Rate) is the fraction of negative cases predicted as positive.

True negative Rate (TN Rate) is the fraction of negative cases were correctly classified as negative.

False negative Rate (FN Rate) is the fraction of positive cases that were incorrectly classified as negative.

A measure that combines precision and sensitivity is the harmonic mean of the two parameters, F-measure.[2] It is calculated using the formula:

$$F_Measure = \frac{2 * TP_Rate * Precision}{TP_Rate + Precision} \quad (1)$$

The Accuracy is the proportion of the total number of correct predictions and calculated as the ratio between the number of cases correctly classified and the total number of cases.

$$Accuracy = \frac{cases_on_main_diagonal}{number_of_totals_cases} = \frac{TP\ Rate + FN\ Rate}{number_of_totals_cases} \quad (2)$$

The Error indicates the proportion of cases classified incorrectly.

$$Error = 1 - Accuracy \quad (3)$$

Sensitivity is the proportion of cases classified positive in relation to all cases tested positive.

$$Sensitivity = \frac{true_positive_cases}{true_positive_cases + negative_false_cases} \quad (4)$$

ROC curves (Receiver Operating Characteristic) are widely used in evaluating the results of predictions. These are the cases of false positive rate OX and OY axis rate of true positive cases. These represent on the OX axe the rate of false positive cases and on the OY axe the rate of true positive cases. Each binary classifier is represented by a point on the graph (FP Rate, TP Rate).[8] Thus,

- The point (0,1) of the ROC plot is the perfect, ideally prediction, classify correctly all the positive and negative cases ;
- The point (1,1) represents a classifier that predicts all cases to be positive;
- The point (1,0) represents a classifier that classifies all instances incorrectly.

4. Case study. Evaluation of classification algorithms on datasets

After determining the algorithms to be used and their performance evaluation criteria, we chose three sets of data for the study.

Table 3 presents selected datasets that address different areas: human resources management, marketing and education. This data sets were selected because they have different characteristics, the number of instances range 141 to 343, and the number of attributes 8 to 18.

Table no. 3 Description of the datasets

Domain	The purpose of the study	Number of characteristics	Effect characteristic	Number of instances
Human Resources Management	Employee performance analysis	14	Performance	216
Marketing	House price analysis	18	Price	343
Education	Student performance analysis	8	Performance	141

Human resource management Dataset (HRM dataset) describes the factors that influence employee job performance of non-life insurance firms: age, origin, sex, marital status, training, specialization, type of recruitment, age, class insurance , hours worked, effective customers, the average time allocated to a client, the level reached, the average policy concluded.

Dataset marketing (MK dataset) contains information on offers on real estate market in Ploiesti area: number of rooms, partitioning, comfort, area, floor, building year, balcony, telephone, cable, internet, central heating, air conditioning, garage, decorating, improvements, thermal insulation, price.

Dataset Education (EDU dataset) contains information on the performance of students at a college and its characteristics: the origin of students, gender, graduated profile, the number of outstanding, license theme, correlation between license theme and grade to discipline license theme, grade to license.

For the study we used the software tool WEKA Explorer provided by University of Waikato, New Zealand. Weka is a collection of machine learning algorithms for solving data mining applications.

Table no. 4 Performance indicators obtained from the application of classification algorithms

The algorithm used Performance indicators	Datasate	ID3	Nayve Bayes	Multilayer perceptron	K-Nearest Neighbor
Accuracy	MRU dataset	93,39%	78,74%	91,09%	93,39%
	MK dataset	99,29%	82,27%	96,45%	99,29%
	EDU dataset	90,67%	65,30%	90,38%	90,09%
Error	MRU dataset	6,61%	21,26%	8,91%	6,61%
	MK dataset	0,71%	17,73%	3,55%	0,71%
	EDU dataset	9,33%	34,69%	9,62%	9,91%
F-measure	MRU dataset	0,93	0,77	0,90	0,93
	MK dataset	0,99	0,82	0,96	0,99
	EDU dataset	0,89	0,65	0,90	0,89
ROC	MRU dataset	0,99	0,91	0,97	0,95
	MK dataset	1	0,92	0,97	0,99
	EDU dataset	0,99	0,82	0,99	0,91

The results of applying classification algorithms indicate that ID3 decision trees and K-nearest neighbor provides the highest accuracy and correctly classifies 99,29% of instances of 343 marketing model, 93,39% of the instances of 216 human resource management model and 90,67% of the 141 instances of model education. It is observed that classification algorithms perform better on large data sets with a large number of attributes.

5. Conclusions

This study focuses on evaluating the performance of classification algorithms, ID3, Nayve Bayes, Multilayer Perceptron, K-nearest neighbor, on the different datasets as scope, size and number of attributes. It is noted that the performance indicators: the accuracy, the error F-measure and the ROC curve varies according to the size of the dataset used, and the number of attributes used in the data set. It also recommends testing several algorithms to identify suitable algorithm dataset.

6. References

- [1] **Bostanci, B., Bostanci, E.**, Proceedings of Seventh International Conference on Bio-Inspired Computing: Theories and Applications (BIC-TA 2012), Advances in Intelligent Systems and Computing 201, DOI: 10.1007/978-81-322-1038-2_2, Springer, India 2013
- [2] **Chimieski, BF, Fagundes RDR.**, Association and Classification Data Mining Algorithms Comparison over Medical Datasets, J. Health Inform. 2013 Abril-Junho; 5(2):44-51
- [3] **Gopala K., Murthy N., Bharath K. P., Nagaraju O.**, Performance Analysis and Evaluation of Different Data Mining Algorithms used for Cancer Classification, International Journal of Advanced Research in Artificial Intelligence, Vol. 2, No.5, 2013
- [4] **Gorunesu, F.**, Data Mining. Concepte, modele și tehnici, Editura Albastră, 2006, Cluj-Napoca
- [5] **Oprea C., Zaharia, M., Gogonea, M.**, Analysis of student performance to license exam using data mining techniques”, The 14th IBIMA Conference on Global Business Transformation through Innovation and Knowledge Management, Istanbul, 2010, ISBN: 978-0-9821489-3-8
- [6] **Oprea, C., Zaharia, M.**, Using Data Mining Methods in Knowledge Management in Educational Field, ANNALS of the ORADEA UNIVERSITY. Fascicle of Management and Technological Engineering, Volume X (XX, NR1), 2011

- [7] **Schrodt, P. A.**, Predicting Interstate Conflict Outcomes Using a Bootstrapped ID3 Algorithm, Political Analysis, Oxford Journals, 1990
- [8] **Vuk, M. and Curk, T.**, ROC curve, lift chart and calibration plot, Metodoloski zvezki, vol.3, no.1, p. 89, 2006