Fifth International Conference
Modelling and Development of Intelligent Systems
June 23 - 25, 2017
"Lucian Blaga" University
Sibiu - Romania

Theoretical and practical approaches for documents classification

Livia Sangeorzan, Nicoleta Enache-David

Abstract

Documents classification is a very actual issue and is a continuous challenge; it is based on different techniques of machine learning including Bayesian classification, SVM classifiers (Support Vector Machine), k-NN (k-Nearest-Neighbor) classifier, classification based on association rules, decision trees, etc.

In this paper we use Weka in order to make a comparison of the accuracy and error rates of two Bayesian classifiers, Naïve Bayes and Naïve Bayes Multinomial on a text training dataset.

1 Introduction

The task of classification occurs in a wide range of domains. The notion of *classification* could cover a context in which some decision is made on the basis of currently available information.

In [17] the authors consider that classification has two distinct meanings. Firstly, they consider a set of observations with the aim of establishing the existence of classes or clusters in the data. Secondly, they suppose they have a set of classes, and the scope is to establish a rule whereby they can classify a new observation into one of the existing classes. The first type is known as Unsupervised Learning and the second type as Supervised Learning.

Document classification is based on different techniques of machine learning, like Bayesian classification, Support Vector Machines [9],[10],[11].

Such methods can be applied on complex information systems like the ones presented in [1],[2],[3],[4] and for mathematical models like [6],[7],[8]. Also, in businesses, the decision making systems use decision trees [12],[13],[14].

In this paper we study the performance of two Bayesian classifiers: Naïve Bayes and Naïve Bayes Multinomial. We study also the amount of time taken to build the classification model.

2 Text classification approach

The problem of text classification consists of classifying documents by their content. Text classification is intended to assigning subjects to certain categories. Naive Bayes classifiers are create simple performing models, especially in the field of document classification. They are based on the Bayes' Theorem. [16]

There are several types of Naïve Bayes classifiers: Multinomial Naive Bayes, Binarized Multinomial Naive Bayes and Bernoulli Naive Bayes. Naïve Bayes and multinomial Naïve Bayes model are both supervised learning methods.

Each type of Naïve Bayes classifiers can have as output different results since they use completely different models.

In practice, it is possible to have more than two classes and the naïve Bayesian classifiers estimate the probability of class c_j generating instance d. Generally, the Naïve Bayes attributes have independent distributions. The assumption to have all attributes independent because of the meaning of the word naïve does not fit in real world situations.

We can give a definition for the text classification like the following: we have as input a document d, a fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$ and as output a predicted class $c \in C$ [17].

We denote by X the document space. In text classification, we are given a description $d \in X$ of a document and a fixed set of classes $C = \{c_1, c_2, \ldots, c_n\}$. Classes are called categories or labels.

3 Case study using Naïve Bayes and Naïve Bayes Multiomial classifiers

In our case study we use a text training dataset having 2132 words. We use Weka environment in order to study the accuracy and error rate when applying two bayesian classifiers: Naïve Bayes and Naïve Bayes Multinomial.

Weka is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost any platform. Features of Weka are: machine learning, data mining, preprocessing, classification, regression, clustering, association rules, attribute selection, visualization. [18]

In Weka we have chosen the Filtered Classifier from Meta category, and we made a comparison between Naïve Bayes and Naïve Bayes Multinomial classifiers implemented in Weka environment.

The dataset was tested using two methods for testing the accuracy: percentage split method, where 66% of the data was used as training dataset and 33% as testing dataset and the 10-fold cross validation method. We have obtained the results from the Table 1.

Method	Accuracy		Error Rate	
	Naïve Bayes	Naïve Bayes Multinomial	Naïve Bayes	Naïve Bayes Multinomial
Percentage Split 66%	81.81%	72.72%	18.18%	27.27%
10 Folds Cross Validation	71.87%	75%	28.12%	25%

Table 1. Accuracy and error rate for Naive Bayes and Naive Bayes Multinomial

From Figure 1 we can see that the Naïve Bayes classifier achieved the highest accuracy (81.81%) and the lowest error rate (18.18%) using the percentage split 66% option.

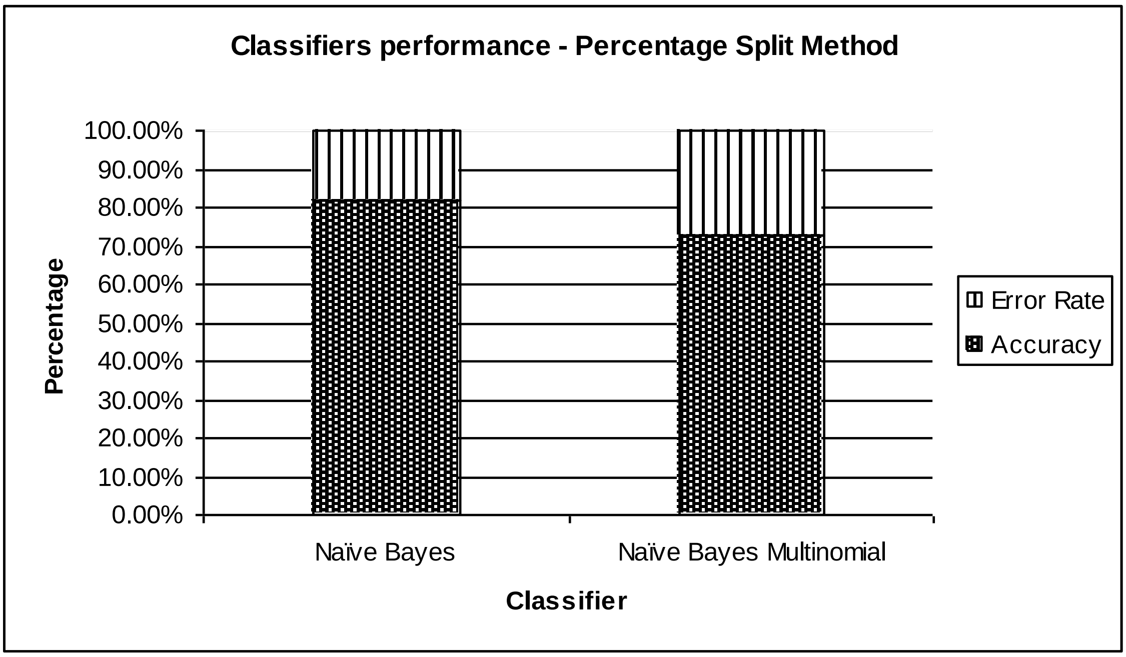

Fig. 1. Accuracy and Error Rate using percentage split method

From Figure 2 we can wee that Naïve Bayes Multinomial classifier achieved the highest accuracy (75%) and the lowest error rate (25%) using 10 folds cross validation method.

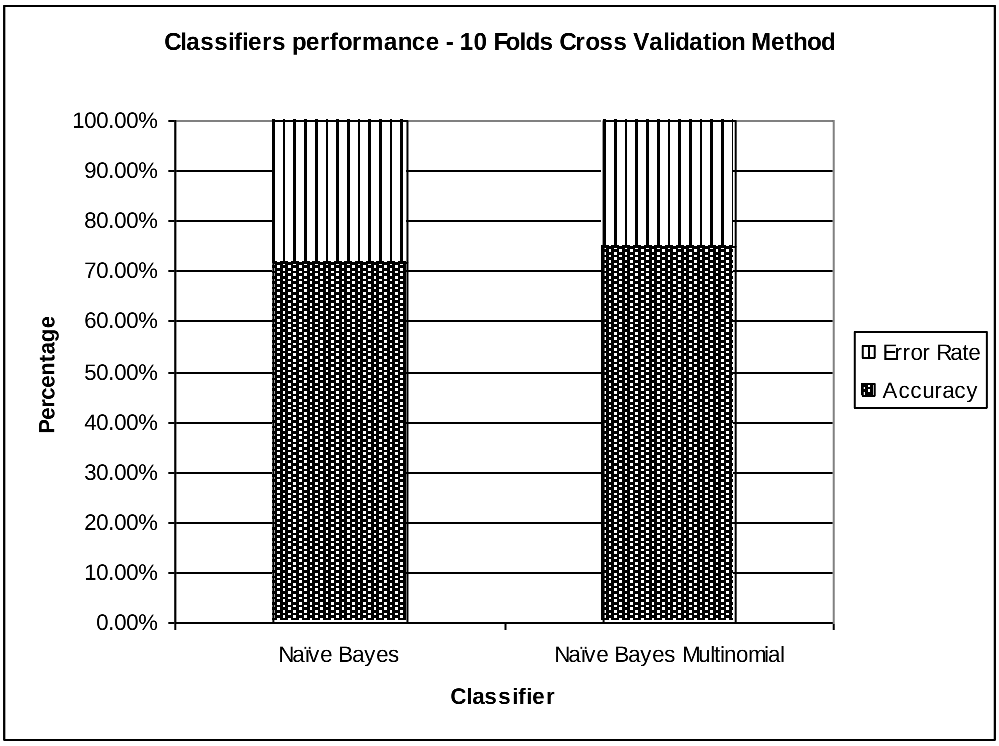

Fig. 2. Accuracy and Error Rate using 10 folds cross validation method

We have studied also the time needed for building the model for the two aforementioned classifiers. In table 2 we present the results.

Method	Time (sec)	
	Naïve Bayes	**Naïve Bayes Multinomial**
Percentage Split 66%	0.07	0.02
10 Folds Cross Validation	0.02	0.01

Table 2. Time to build the model for Naïve Bayes and Naïve Bayes Multinomial

The results show that the best amount of time was achieved by the Naïve Bayes Multinomial classifier with 10 folds cross validation method (0.01 sec), while Naïve Bayes classifier achieved 0.07 sec with the percentage split 66% method.

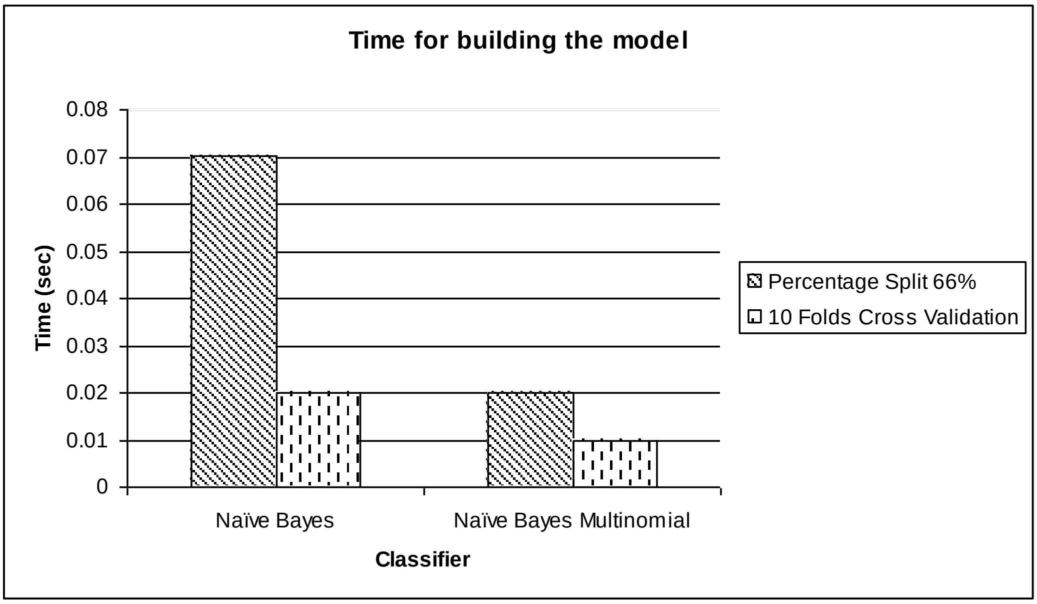

Fig. 3 Amount of time to build the model

4 Conclusion

In this paper we made a comparison regarding the performance of two types of Naïve Bayes classifiers on a document classification problem.

The conclusion is that the Naïve Bayes classifier achieved the highest accuracy using the percentage split 66% option, while the best amount of time was achieved by the Naïve Bayes Multinomial classifier with 10 folds cross validation method.

References

[1] C. Carstea, Control and management in complex information systems, *Bulletin of the Transilvania University of Braşov, Series III Mathematics • Informatics • Physics*, pp 73-87, 2013;

[2] C.G.Carstea, Modeling System's Process for Control Of Complex Information Systems, In Proceedings of The 25th International Business Information Management Association Conference (IBIMA), Soliman KS (ed). Amsterdam, Netherlands, May 2015, pp 566-574, 2015.

[3] C. Carstea, Optimization Techniques in Project Controlling, *OVIDIUS UNIVERSITY ANNALS, ECONOMIC SCIENCES SERIES*, Volume XIII Issue 1, pp. 428-432, Ovidius University Press, 2013.

[4] C.G. Carstea, IT Project Management – Cost, Time and Quality, ISSN 2067-5046, *Economy Transdisciplinary Cognition International* , Volume17, Issue 1/2014, pg.28-34, 2014.

[5] N. Enache-David, L. Sangeorzan, An overview on data mining techniques, *Proceedings of the 20th GBU-IC on Control, Development and APPLIED INFORMATICS in BUSINESS and ECONOMICS*, Brasov, Romania, November 1-2, Editura Didactica si Pedagogica Publishing House, Bucharest, Romania, ISSN:2069-7937, 2013.

[6] O. Florea, I. Rosca, The Mechanical Behavior and the Mathematical Modeling of an Intervertebral Disc, *Acta Technica Napocensis Series-Applied Mathematics Mechanics and Engineering*, 58: 213-218, 2015.

[7] O. Florea, I.C Rosca,, Analytic study of a rolling sphere on a rough surface, *AIP ADVANCES*, Volume: 6, Issue: 11, 2016.

[8] O. Florea, I.C. Rosca, Stokes' Second Problem for a Micropolar Fluid with Slip, *PLOS ONE*, Volume: 10, Issue: 7, Article Number: e0131860, 2015.

[9] A. Khashman, N.I. Nwulu, Support vector machines versus back propagation algorithm for oil price prediction, *Advances in Neural Networks–*ISNN 2011, pp 530-538, 2011.

[10] A. Khashman, Blood Cell Identification using Emotional Neural Networks, *J. Inf. Sci. Eng*. 25 (6), pp 1737-1751, 2009.

[11] A. Khashman, IBCIS: Intelligent blood cell identification system, *Progress in Natural Science 18* (10), PP1309-1314, 2008.

[12] L. Mandru, How to Control Risks? Towards A Structure of Enterprise Risk Management Process, *Journal of Public Administration, Finance and Law*, pp 80-92, 2016.

[13] M. Popescu, L. Mandru, Relationship between Quality Planning and Innovation, *Bulletin of the Transilvania University of Braşov*, Series V, Economic Sciences, Vol. 9 (58) No. 2, pp.203-212, 2016.

[14] L. Mandru, *Managementul integrat calitate-risc pentru societăţile comerciale cu profil industrial*, Editura Universităţii Transilvania, Brasov, 2011.

[15] C.D Manning, P. Raghavan, H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press, 2008.

[16] H. Robbins, J. Ryzin, Introduction to statistics, Science research associates Inc., 1975.

[17] D. Michie, D.J. Spiegelhalter, C.C. Taylor, *Machine Learning, Neural and Statistical Classification*, Ellis Horwood Upper Saddle River, NJ, USA, 1994.

[18] http:// http://www.cs.waikato.ac.nz/ml/weka/.

LIVIA SANGEORZAN
Transilvania University
Department of Mathematics and Informatics
Str. Iuliu Maniu 50, Brasov
ROMANIA
E-mail: livia.sangeorzan@gmail.com

NICOLETA ENACHE-DAVID
Transilvania University
Department of Mathematics and Informatics
Str. Iuliu Maniu 50, Brasov
ROMANIA
E-mail: nicoleta.enache@unitbv.ro