Comparative Study of Data Classifiers Using Rapidminer

Abhishek Kori

Assistant Professor, IT Department, SVVV Indore, India

Abstract--**Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help to focus on the most important information in data warehouses. Classification is the process of organizing data into categories for its most effective and efficient use. A well-planned data classification system makes essential data easy to find and retrieve. In this paper our focus is text classification by Naïve Bayes Classification and KNN and determine accuracy of the classifier using rapidminer tool.**

Keywords:Classification,Naïve Bayes,KNN

I. INTRODUCTION

RapidMiner is a data science software platform developed by the company of the same name that provides an integrated environment for machine learning, deep learning, text mining, and predictive analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the machine learning process including data preparation, results visualization, validation and optimization.[1] RapidMiner is developed on an open core model. The RapidMiner (free) Basic Edition, which is limited to 1 logical processor and 10,000 data rows, is available under the AGPL license.

II.LITERATURE SURVEY

Classification is a learning function that maps a given data item into one of several predefined classes. It is a data analysis technique to extract models describing important data classes and predict future values. Data mining uses classification technique uses with machine learning ,image processing, natural language processing, statistical and visualization techniques to discover and present knowledge in an understandable format. Most of the classification algorithms in literature are memory resident, typically assuming a small data size.

III.CLASSIFICATION ALGORITHM

3.1 Naïve Bayes Classifier

Naive Bayes is a simple technique for constructing classifiers: models that assign class labels to problem instances, represented as vectors of feature values, where the class labels are drawn from some finite set. It is not a single algorithm for training such classifiers, but a family of

algorithms based on a common principle: all naive Bayes classifiers assume that the value of a particular feature is independent of the value of any other feature, given the class variable.

$$p(c/d) = p(c)p(d/c)/\ p(d)$$

Where P(d) plays no role in selecting c.

3.2 KNN Classifier

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The *k*-NN algorithm is among the simplest of all machine learning algorithms.

IV.EXPREMENTATION & RESULTS

In this paper, experiment is carried out using tool Rapidminer 7.4 we are taking into consideration the data set Iris, given as sample data inside the repository panel of the tool. We apply validation tool on the data set which in turn contains training and testing operations. In training column, we took naïve bayes operator and in testing column we took apply model and performance tools respectively. After connecting all operators we execute the tool which in turn shows the accuracy as result. In the paper ,we have consider this process for both naïve bayes and KNN and compare the results given in the figure shows below.

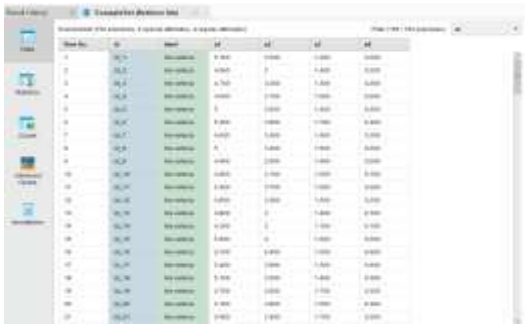

Figure 4.1 Sample Data Set Iris

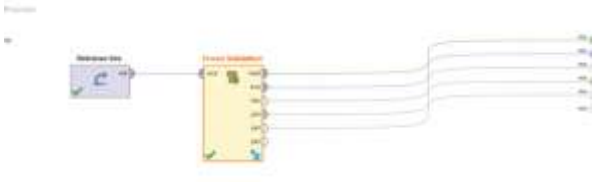

Figure 4.2 Applying Validation Operator

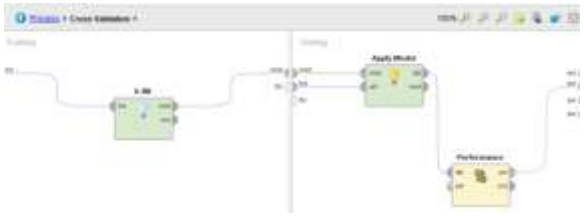

Figure 4.3 Applying KNN Operator

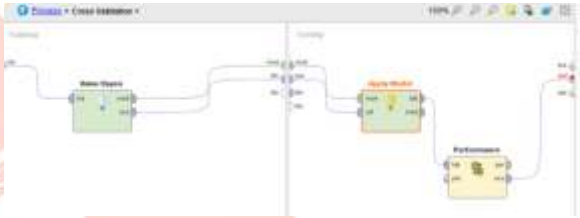

Figure 4.4 Applying Naïve Bayes Operator

Figure 4.5 Performance Evaluation KNN

Figure 4.6 Performance Evaluation Naïve Bayes

V.CONCLUSION

This is because KNN is non-parametric, i.e. it makes no assumption about the data distribution. Contrast this to NB, which assumes that attributes are conditionally independent to each other given the class, and are normally distributed (for real-valued attributes).The experiment carried out shows that accuracy of KNN is greater than Naïve Bayes but this situation is not similar always, as it may vary with different data set.

REFERENCES

[1] Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes
a good classifier for document classification?." *International*
[2]Williamson, Eric R., and Saurabh Chakravarty. "CS5604 Fall
2016 Classification Team Final Report." (2016).
[3].Ikonomakis, M., S. Kotsiantis, and V. Tampakas. "Text
classification using machine learning techni ques." *WSEAS*
transactions on computers 4.8 (2005): 966-974.
[4] Kamruzzaman, S. M., Farhana Haider, and Ahmed Ryadh
Hasan. "Text classification using data mining." *arXiv preprint*
arXiv:1009.4987 (2010).
[5] Joachims, Thorsten. "Text categorization with support vector
machines: Learning with many relevant features." *European*
conference on machine learning. Springer Berlin Heidelberg,
1998.
[6]. Menaka, S., and N. Radha. "Text classification using
keyword extraction technique." *International Journal of*
Advanced Research in Computer Science and Software
Engineering 3.12 (2013).
[7] Williamson, Eric R., and Saurabh Chakravarty. "CS5604 Fall
2016 Classification Team Final Report." (2016).
[8] Dalal, Mita K., and Mukesh A. Zaveri. "Automatic text
classification: a technical review." *International Journal of*
Computer Applications 28.2 (2011): 37-40.
[9] Ting, S. L., W. H. Ip, and Albert HC Tsang. "Is Naive Bayes
a good classifier for document classification?." *International*
Journal of Software Engineering and Its Applications 5.3 (2011):
37-46.
[10] Mahesh Kini M , Saroja Devi H , Prashant G Desai,
Niranjan Chiplunkar.” Text Mining Approach to Classify
Technical Research Documents using Naïve Bayes”
International Journal of Advanced Research in Computer and
Communication Engineering Vol. 4, Issue 7, July 2015
[11]. Gongde Guo, Hui Wang, David Bell, Yaxin Bi and Kieran
Greer, “KNN Model-Based Approach in Classification”, Proc. ODBASE pp- 986 – 996, 2003