Comparison of Classifier Techniques in Data Mining – A Review

Elda A V
M.Tech Student, Department of School of Computer Science, Mahatma Gandhi University, Kottayam, Kerala, India

ABSTRACT: Data mining is the technique of discovery of unknown patterns from a large database. This discovery of patterns can be done using classification methods. Classification is used to find out in which group each data instance is related within a given dataset. It is used for classifying the extracted data into different classes according to some constrains. The purpose of this paper is to provide the review and analysis on classification and regression techniques models. The importance of naïve bayes classifier and the logistic regression method is compared. And by comparing the classification method results that, the Naïve Bayes outperforms the Logistic regression methods. And also the multinomial Naïve Bayes gave smaller error rate and better performance than the other three Naïve Bayes classifier.

KEYWORDS: Data Mining, Naïve bayes Classifier, Logistic regression.

I. INTRODUCTION

Data mining has attracted a great deal of attention in the information industry and in society as a whole in recent years. Wide availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge make the data mining an important task. Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, useful and understandable patterns in large databases. Conventionally, the information that is mined is denoted as a model for the semantic structure of the datasets.

Classification techniques in data mining are capable of processing a large amount of data. It can be used to predict categorical class labels and classifies data based on training set and class labels and it can be used for classifying newly available data. So a classifier is an algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term “classifier” sometimes also refers to the mathematical function, implemented by a classification algorithm, which maps input data to a category[7].

Applications of data mining

- Market analysis
- fraud detection, and customer retention
- production control
- science exploration

Paper is organized as follows. Section II explains various works on the field of classifiers and compares their performances. In this section the three naïve bayes classifiers taken into the discussion. Finally, Section III presents conclusion.

II. RELATED WORK

On 2016, W. Chen et al. explained, whether the public emotions or opinions in the Chinese social media websites could be used to predict the stock market price in China using Chinese emotion mining approach. The public mood states extracted from Chinese social media websites such as Weibo. This paper first presents a topic-based approach to extract public mood states from Weibo. After that a Granger causality analysis is used to investigate the hypothesis that the traced public mood states are predictive of the stock price movement in China. This paper [1] expound that
the public mood states can be indeed tracked using the proposed new Chinese sentiment analysis method on the large-scale Weiibo posts.

H. Saif et al. analysed the sentiment of tweets by the help of SentiCircles, a lexicon based approach for sentiment analysis on Twitter. For tweet level sentiment detection, this approach also gave a better overall result than the state of the art lexicon based approach SentiStrength on two out of three datasets. Also, concluded that the proposed system outperformed other lexicon labelling methods for both entity level and tweet level sentiment detection [2].

Each classification method has its strengths and limitations and that real world problems do not always satisfy the assumptions of a particular method. So Melody Y. Kiang explained the important role of classification in business decision-making and the effects of data characteristics on their performance in a controlled setting on “A comparative assessment of classification methods”. Also it assesses the relative performance of some well-known classification methods based on statistical and AI techniques. In this research, the main focus is on the investigation of two AI techniques—neural networks and a decision tree method, and three statistical methods—linear discriminant analysis (LDA), logistic regression analysis, and kth - nearest - neighbor (kNN) models. The controlled experiments conducted with simulated data show that the classification algorithms are sensitive to changes in data characteristics. Therefore, one recommendation from this study is to build classification systems that employ a number of different classification algorithms. From the considered classification methods Melody Y. Kiang concludes that neural net and logistic regression methods provide the best relative performance under most scenarios [3]. Another important concern brought forth by this paper is the impact of dynamic variations in data on classification performance.

In 2006, Metsis et al. concluded that the most interesting result of this evaluation was the very good performance of the two Naïve Bayes versions that have been used less in spam filtering, i.e., flexible Bayes and the multinomial Naïve Bayes with Boolean attributes; these two versions collectively obtained the best results in the experiments. Taking also into account its lower computational complexity at run time and its smoother trade-off between ham and spam recall, tend to prefer the multinomial Naïve Bayes with Boolean attributes over flexible bayes, but further experiments are needed to be confident. The best results in terms of effectiveness were generally achieved with the largest attribute set (3000 attributes); as one might have expected, but the gain was rather insignificant, compared to smaller and computationally cheaper attribute sets [4].

In 2012, Cary et al. compared the naïve bayes and logistic regression on “Comparison of Naïve Bayes and logistic regression for computer-aided diagnosis of breast masses using ultrasound imaging”. This study shows that two classifiers employing very different learning strategies can provide highly correlated estimates of probability of malignancy for breast masses on ultrasound. Naïve Bayes is more sensitive to feature selection than logistic regression, but selection by entropy-based information gain can improve performance so that the two classifiers are comparable [5].

This paper [6] presents a survey and analysis for existing techniques on both classification and regression models techniques that have been applied for diseases outbreak prediction in datasets. An analysis critique has been presented detailing the weaknesses of the varies classification and regression models where after the establishment of a comprehensive data mining hybrid model are mostly that it will be able to predict disease outbreaks with a maximum accuracy. Also it request for further research to address the various weaknesses and the idea behind combined (Hybrid) prediction models is to derive advantages of individual models best features to obtain the best possible results in a given problem/situation ( ShubpreetKaur a. D., 2015) [6].

Noor Diaa Ahmad Tarmizi (2013) stated that a classification technique, also known as a classifier is a systematic approach in developing a classification model from a set of input data. There are many classifiers such as decision tree (DT), Neural networks (NN), naïve Bayesian, support vector machine (SVM) and rough set theory (RST). Each classifier uses learning algorithms to discover the most appropriate model for the relationship between an attribute set and class labels of input data. The model produced by the learning algorithm should both fit the input data well and correctly predict the class labels of records that it has never seen before [6]. By comparing the classification methods, the Naïve Bayes outperforms the Logistic regression methods. And also the multinomial Naïve Bayes gave smaller error rate and better performance than the other three Naïve Bayes classifier.
III. CONCLUSION

In recent years, data mining using classification method become more important. The classification methods reviewed here are the naïve bayes and the logistic regression method. From the reviews and comparison it can be concluded that the Naïve Bayes outperforms the Logistic regression methods. And also the multinomial Naïve Bayes gave smaller error rate and better performance than the other three Naïve Bayes classifier.

REFERENCES

https://en.wikipedia.org/wiki/
https://machinelearningmastery.com/naive-bayes-for-machine-learning/