International Journal of Mathematics and
Computer Science, **15**(2020), no. 1, 389–394

Data mining classification algorithms

Mohamed Saouabi, Abdellah Ezzati

LAVETE Laboratory
University Hassan the 1^{st}, FST
Settat, Morocco

email: mohamed.saouabi@gmail.com, abdezzati@gmail.com

(Received May 30, 2019, Accepted November 20, 2019)

Abstract

Data mining nowadays became more and more important; it is used to extract knowledge from a large amount of data to help decision makers to make good decisions. Based on the type of the data and the type of the variable we want to predict, we choose the suitable algorithm, classification is one of the different algorithms used in data mining. In this survey paper, we first presented a brief overview about data mining and the data mining classification algorithms; then, we present classification algorithms used in data mining for prediction, and we present previous work and experiments using classification algorithms, such as Decision tree, Logistic regression and Naive Bayes. The aim is to give a clear insight of the classification algorithms and how they are used in several domains.

1 Introduction

Data mining is about using data analysis tools to find out new unknown knowledge, hidden relationships between the large data set we have on hands. These tools can handle mathematical algorithms, statistical models and machine learning methods. Data mining is not just about collection and managing the data but it includes, as well, data analysis and prediction. There is a lot of algorithms we can use with data mining in order to solve a particular

Key words and phrases: Data mining, Classification, Knowledge, Decision tree, Logistic regression, Naive Bayes.
AMS (MOS) Subject Classifications: 68T01.
ISSN 1814-0432, 2020, http://ijmcs.future-in-tech.net

problem, based on the type of data we have, we choose which method we will apply for modeling. We will be focusing on classification algorithms, also known as supervised classification, which can be used to categorize the data into classes, and predicts the class for the new data.

2 Data Mining

There is a huge amount of data available in the information industry. This data is of no use until it is converted into useful information. It is necessary to analyze this huge amount of data and extract useful information from it. Extraction of information is not the only process we need to perform; data mining [1] involves other processes such as Problem Understanding, Data Exploration, Data Preparation, Modeling, Evaluation and Deployment. Once all these processes are over, we would be able to use this information in order to improve a particular domain.

3 Classification Algorithms

Classification [2] consists of predicting a certain outcome based on a given input. In order to predict the outcome, the algorithm processes a training set containing a set of attributes and the respective outcome, usually called goal or prediction attribute. The algorithm tries to discover relationships between the attributes that would make it possible to predict the outcome. Next the algorithm is given a data set not seen before, called prediction set, which contains the same set of attributes, except for the prediction attribute - not yet known. The algorithm analyses the input and produces a prediction. The prediction accuracy defines how " good" the algorithm is. For example, in a medical database the training set would have relevant patient information recorded previously, where the prediction attribute is whether or not the patient had a heart problem.

3.1 Decision Trees

A decision tree algorithm [3] is used in classification by dividing the data into classes; it is one of the most effective algorithms for data mining. It is used in several disciplines due to its simplicity and robustness in presence of missing values. A decision tree can be used for both prediction and description. Decision trees, often times, are used for classification systems to attribute

type information, and also with predictive systems, where the predictions are based on a past experience of historical data and it will help to drive the structure of the decision tree and the output.

3.2 Logistic regression

Logistic regression [4] is a predictive algorithm we can use when the variable we want to predict is categorical, which means having two categories, and they can be either numerical or categorical. It estimate also the probability that an event will occurs for a random selected observations versus the probability that an event does not occurs. Logistic regression classifies observations by estimating the probability that an observation is in particular category.

3.3 Naive Bayes

Naive Bayes (NB) [5] or Bayesian prediction followed more human thought patterns than classical statistical analysis or even machine learning algorithms. The major disadvantage of this fact is that two humans can (and often disagree) in the decisions they make as a result of this reflection. But there are many other situations in which the Bayesian approach to truth is much more appropriate and may even be better when we have to deal with the need to classify an entity in the world around us.

4 Previous Experiments Using Classification Algorithms

In this part, we will list previous experimentation working with data mining classification algorithms such us Decision tree, Logistic regression and Naive Bayes.

In the first experiment, Hetal Bhavsar and Amit Ganatra [6] presented a comparison of five classification algorithms which are Decision tree, Multilayer perceptron, K-nearest neighbor, Naive Bayes and Support Vector Machine, they compared the performance of each algorithm using different metrics such us accuracy, error rates, building time of classifier and other statistical measures on WEKA tool, the result they came up with is there is no universal classification algorithm which works better for all the dataset.

In the second experiment, Parneet Kaura, Manpreet Singhb and Gurpreet Singh Josanc [7] presented an experimentation study using classification algorithms, in order to identify the slow learners among students, by using a real world data set collected from high school using WEKA as an open Source Tool. They compared five classifiers which are Multilayer Perception, Naive Bayes, SMO, J48 and REPTree, and the result have shown that MLP proves to be potentially effective and efficient classifier algorithm for the data set they used.

In the third experiment, Ahmed Mohamed Ahmed, Ahmet Rizanerc and Ali Hakan Ulusoyc [8] presented an experimentation to predict the instructor performance and investigates the factors that affect the students' achievements and to improve the education system quality. They compared four classifiers to find the best performing classification algorithm, which are Decision Tree, Multilayer Perception, Naive Bayes and sequential Minimal Optimization. The results show that using the attribute evaluation method on the dataset increases the prediction of the performance accuracy.

In the fourth experiment, Mirpouya Mirmozaffari, Alireza Alinezhad, and Azadeh Gilanpour [9] presented an experiments using classification algorithms in order to predict heart disease. Several classification algorithms have been implemented in order to choose the most efficient algorithm for the data set used. The results have shown that Random tree model accuracy is 97.6077% and they considered that it's the highest performance algorithm.

In the fifth experiments, Jitendra Jain and Parashu Ram Pal [10] presented an experiment using classification algorithm which is Naive Bayes to detect unknown computer worms for the computers protection. They succeeded in building classifiers that classify an unknown tuple to a particular virus.

In the sixth experiments, Monire Norouzi, Alireza Souri, Majid Samad Zamini [11] presented a data mining classification approach to detect malware behavior. They proposed different classification methods in order to detect malware based on the feature and behavior of each malware. A dynamic analysis method has been presented for identifying the malware features using Weka as a data mining tool. The results have shown the availability of the proposed data mining approach. Also their proposed data mining approach is more efficient for detecting malware and behavioral classification of malware can be useful to detect malware in a behavioral antivirus.

In the seventh experiments, Puspita Kencana Sari, Andry Alamsyah, Sulistyo Wibowo [12] presented an experiment using the Naive Bayes classification algorithm in order to chart the level of service quality according

to customer perception, the classified the reviews into positive and negative sentiment for five dimensions of electronic service quality. The results have shown that personalization and reliability dimension required more attention because of its high negative sentiment. And trust and web design dimension have high positive sentiments. And the responsiveness dimension has balance sentiment positive and negative.

In future works, we want to present an employability model prediction, in the first place, we will apply various data mining classification algorithms, after that we will choose the most efficient and best suited algorithm for the employability data, with the highest accuracy model, using different metrics to evaluate the models. Afterwards, we'll present the model in details and discuss the results. We are working in a Big Data environment, using Hadoop. In the next work, we want to propose a system, in a Hadoop environment, dedicated for employability prediction using data mining techniques.

5 Conclusion

Data mining offers many techniques in order to discover hidden patters between the data. These hidden patterns can be used to predict future behavior. In this paper, we presented a survey about classification algorithms, such as Decision Tree, Logistic regression and Naive Bayes, we defined the classification algorithms, how it works and when we need this kind of algorithms, after that, we presented several experiments using classification algorithms comparing between these algorithms using different metrics, and the results they came up with. Each algorithm has its own specificity, and can be used for a particular type of data set, and the choice of the algorithm depends on how accurate the model is.

References

[1] Tariq O. Fadl Elsid, Mirghani. A. Eltahir, Data Mining: Classification Techniques of Students' Database A Case Study of the Nile Valley University, North Sudan, International Journal of Computer Trends and Technology IJCTT, **16,** no. 5, (2014), 192–203.

[2] Akhtar, Syed Muhammad Fahad, Big Data Architect's Handbook, Packt Publishing, 2018, ISBN: 978-1-78883-582-4.

[3] Alexandru Toprceanu, Gabriela Grosseck, Decision tree learning used for the classification of student archetypes in online courses, Procedia Computer Science, Volume 112, (2017), 51–60.

[4] Shaoyan Zhang, Christos Tjortjis, Xiaojun Zeng, Hong Qiao, Iain Buchan, John Keane, Comparing data mining methods with logistic regression in childhood obesity prediction, Information Systems Frontiers, **11,** no. 4, (2009), 449–460.

[5] A. Pisote, V. Bhuyar, REVIEW ARTICLE ON OPINION MINING USING NAIVE BAYES CLASSIFIER, Advances in Computational Research, **7,** no. 1, (2015), 259–261.

[6] Hetal Bhavsar, Amit Ganatra, An Empirical Evaluation of Data Mining Classification Algorithms, International Journal of Computer Science and Information Security (IJCSIS), **14,** no. 5, (2016), 142–150.

[7] Parneet Kaura, Manpreet Singhb, Gurpreet Singh Josanc, Classification and Prediction Based Data Mining Algorithms to Predict Slow Learners in Education Sector, Elsevier Procedia Computer Science, **57,** (2015), 500–508.

[8] Ahmed Mohamed Ahmed, Ahmet Rizanerc, Ali Hakan Ulusoyc, Using data Mining to Predict Instructor Performance, Procedia Computer Science, **102,** (2016), 137–142.

[9] Mirpouya Mirmozaffari, Alireza Alinezhad, and Azadeh Gilanpour, Data Mining Classification Algorithms for Heart Disease Prediction, International Journal of Computing Communications & Instrumentation Engg (IJCCIE), **4,** no. 1, (2017), 11–15.

[10] Jitendra Jain, Parashu Ram Pal, Detecting Worms Based on Data Mining Classification Technique, International Journal of Engineering Science and Computing, **7,** no. 5, (2017), 11388-11391.

[11] Monire Norouzi, Alireza Souri, Majid Samad Zamini, A Data Mining Classification Approach for Behavioral Malware Detection, Journal of Computer Networks and Communications, Volume 2016, 1–9.

[12] Puspita Kencana Sari, Andry Alamsyah, Sulistyo Wibowo, Measuring e-Commerce Service Quality from Online Customer Review using Sentiment Analysis (Case Study: Tokopedia), Journal of Physics: Conference Series, Volume 971, (2018), 1–6.