International Journal of Scientific Research in Engineering (IJSREM)
Volume: 04 Issue: 03 | Mar -2020 ISSN: 2582-3930

Systematizing and Forecasting Students Interpretation Using Enhanced Decision Tree C4.5 In Higher Education Institutes

Nagaraju Hemanth Kumar[1], J.S. Ananda Kumar[2]
[1]PG Scholar, Dept. of MCA, Sietk, Puttur,
[2]Assistant Professor, Dept. of MCA, Sietk, Puttur, A.P.

Abstract:

Students' information in higher education institutions increases yearly. It is hard for them to extract meaningful information from the huge amount of data manually. Such information can support academic staff to stop students from dropping out at the end of courses. This can be done by evaluating the students' performance for the course and also by predicting their performance in the final exam early by using classification algorithms. Four classification algorithms, which are Decision Tree C4.5, Random Forest, Support Vector Machine (SVM) and Naive Bayes, were used in this research in order to classify and predict the students' performance. Furthermore, this research aimed at improving the Decision Tree C4.5 algorithm by adding a grid search function in order to improve prediction accuracy in classifying and predicting the students' performance. Also, the features of this evaluation have been extracted through the interviews with academic staff of three universities (University of Zakho, Duhok Polytechnic University and University of Duhok), in Duhok province, Kurdistan Region, Iraq and through the review of the literature. A new prototype has been proposed as a tool to classify and predict the students' performance by using Accrod.Net library. Three datasets were utilized in this research in order to test the improved Decision Tree C4.5 with the traditional C4.5 and three other selected algorithms. The results showed that the improved Decision Tree C4.5 outperformed the traditional C4.5 and also performed better when compared to C4.5 (J48) in Weka tool and other algorithms used in this research.

Keywords: Educational Data Mining, Classification Algorithms, Improved Decision Tree, Evaluation Methods,Accord.Net.

Introduction

The concept of Data mining is to draw out the concealed pattern as well as discover the relation among all features in a very large amount of data. Data mining techniques are used in a lot of areas, such as; education, marketing, engineering, finance, sport and medicine (Abdulla et al., 2015). Recently the assurance on data mining and educational systems is increased, this made the educational data mining a new area for research.

Educational Data Mining (EDM) is a very - attractive area which contains many research domains that can handle the improvement of methods to use the data produced by educational context. Computational methodologies are used by EDM to evaluate the educational data.

Additionally, educational data mining is a developed system which concentrates on developing the methods for extracting the relationship from a large amount of data that produced the domain of the education. These methods help in understanding how students learn along with understanding their behavior (Monjurul Alom and Courtney, 2018). In educational institutions, the students' performance is the most important issue in any educational process (Sivakumar et al., 2018). Hence, classifying and predicting student's' performance by using educational data mining techniques takes more attention by researchers according to the large amount of data that is generated yearly in educational institutions. Furthermore, the EDM is used to extract new hypotheses and discoveries about the performance of the students (Sivakumar et al., 2018).

EDM is continuously developing along with many data mining techniques used in educational environments. Therefore, the process of EDM techniques starts with discovering the relation among data by using classification, clustering, regression, association rule mining and others. Then, the trained model can be validated theoretically and after that, the trained model can be used to predict the output for new unseen data. Thus, the results of the prediction can be used to help teachers and students make the right decisions (Baker, 2010). The aforementioned techniques are responsible for exploring data from educational institutions which provide quality education regarding the students' performance as a main purpose (Baradwaj and Pal, 2012) Indeed, many researchers used EDM techniques in order to classify and predict the students' performance (Costa et al., 2017; Rao et al., 2017; Hussain et al., 2018; Kiu, 2018; Sivakumar et al., 2018).

Classification is one of the data mining techniques that is widely used in educational environments for classifying and predicting the students' performance in educational institutions (Baker, 2010). Classification is a supervised learning method used to put data into groups in order to aim at choosing classes. There are two models in the classification method; descriptive model and a predictive model. The descriptive model is responsible for distinguishing the records of different classes, while the predictive model is responsible for predicting classes for new unseen records (Pang-Ning et al., 2006). Many researchers have done their research in classifying and predicting the students' performance using classification algorithms. Based on the literature review, the most popular algorithms used by many researchers are Decision Tree C4.5, Random Forest, SVM and Naïve Bayes (Costa et al., 2017; Sivakumar et al., 2018).

In educational institutions, the data of the students increase yearly. This makes the teachers and researchers pay more attention to students' performance. Classifying and predicting the students' performance becomes a big duty and it is hard for them to do that manually (Rao et al., 2017). Furthermore, the issue of students' failure in courses becomes a very popular area nowadays for teachers and researchers (Costa et al., 2017). On

the other hand, many researchers applied different classification algorithms in order to classify and predict the students' performance during their academic year. These algorithms were applied on datasets which contain features, such as: Academic features, socio-economic and demographic features. Academic features include Marks, Quizzes, assignments, homework and so on (Mahboob et al., 2016; Kadambande et al., 2017; Rao et al., 2017).

Socio-economic and demographic features include gender, marital status, dormitory, family size and so on (Kostopoulos and Lipitakis, 2017; Rao et al., 2017). Moreover, some of these algorithms still need improvement as (Sivakumar et al., 2018) said "we improved the decision tree to get more effective accuracy results from the attribute values". This study tries to improve the Decision Tree C4.5 algorithm. This improvement can have more accuracy regarding the classification and prediction of the students' performance in the higher education institutes. In addition, researchers in this study used three other classification algorithms which are Random Forest, SVM and Naïve Bayes. These algorithms were used in order to see the comparison of the performance of these algorithms. However, there is no specific algorithm which performs better than others in solving problems (Tiago and Cheplygina,2014). Henceforth, we have chosen the most popular algorithms in such area based on our review. These algorithms are used in classifying the students' performance during the course and also used to early predict their performance in the final exam.

Literature Review

The term "Educational Data Mining" appeared for the first time in 2005, this happened during the 2005 annual conference of the Association for the Advancement of Artificial Intelligence in Pittsburgh, USA, in a workshop on Educational Data Mining. After the first workshop, EDM became popular and widely used in higher education, also many researchers have been working on this area in order to improve the quality of education (Romero and Ventura, 2007).
The data mining techniques presented by (Hussain et al., 2018), these techniques were used in higher education institutes for improving the students' academic performance to prevent them from dropping out. The aforementioned researchers collected data from three different colleges. These data consisted of demographic, socio-economic as well as academic information of three hundred students. Four classification algorithms; J48, Bayes Net, PART and random forest, were used in order to predict the student performance during their academic year. The results showed that Random Forest performed better than the others according to the accuracy of 99%.

Online learning has a social impact on students' performance (Kiu, 2018). Also, the aforementioned study showed that there are many problems which make the experience of students destitute, such as: Less face-to-face interaction,

low understanding, lack of concentration on learning activities, difficulty in performing teamwork activities and high social isolation. Furthermore, the level of understanding of students'learning can be understood and concentrated through knowledge discovered from students' event in their online learning. This knowledge can be used in order to predict the students' performance on their final degrees. In the previously mentioned study, a classification was done based on activities extracted from students' event logs in online learning, which also depends on their learning process, the above study theory was done by using a supervised leaning algorithm, J48. The results of the mentioned study illustrated that the model could improve the learning process of students and enhance their performance. Additionally, the generated model could give advice to students who were at risk in a timely manner.

Moreover, Sivakumar (2012) showed that activities in a learning environment and other activities, such as: Academic extracurricular, Co-Curricular activities, internal examinations and grade obtained in the university examination, affect the students' performance. In the aforementioned work, data were collected using the real-world dataset which was related to students' performance in education in India. This data contained information about the students including CGPA, History of arrears, lab performance and so on. In the same study, different data mining techniques, such as; Decision tree, Naive Bayes, Neural Network, KNN and Improved Decision tree, were used in order to predict the students' performance. The results of this experiment indicated that the improved decision tree performed better than the other classification algorithms.

Costa et al. (2017) claimed that the high rates of students' failure in courses of programming became a very interesting area for teachers and researchers to search for reasons. The strength of Educational Data Mining was presented in the aforementioned study in order to early predict the students' failure in programming courses (for problem statement with ref in overview). Furthermore, in the above study, four Algorithms: Naive Bayes (NB), Neural Network (NK), Decision Tree (DT) and Support Vector Machine (SVM) were used in order to predict student failure. These Techniques were applied on two data sources of Brazilian University, one came from the campus and the other came from distance education.

The results presented that SVM performed better than the other techniques.

Students' data in educational institution increase yearly, this directs researchers' attention more towards the students' performance. Evaluating and predicting student's performance became a big task in the education system and it is hard to do that manually because of the databases which are multidimensional (Rao et al., 2017). For this purpose in this study, Educational Data Mining was utilized (EDM) to extract meaningful information from the huge amount of data. The aforementioned study depends on internal assessment, such as: Quiz, lab mark, class test and

attendance. And also, two more features were added to the others; external assessments and demographics.

Furthermore, in the above study, five data mining algorithms were used; Neural Network, Decision Tree, Support Vector Machine, Naive Bayes and K-Nearest Neighbour. The results showed that neural network performed better than others. The Neural Network accuracy was (98%) followed by Decision Tree which was (91%). Then, Support Vector Machine and K-Nearest Neighbour gave the same accuracy, which was (83%). Lastly, the method that had lower prediction accuracy was Naïve Bayes by (76%).

Kadambande et al. (2017) found that the students' education level is getting lower day by day which has become a large issue and therefore student performance prediction is important. For this purpose, in their study, two data mining techniques; semantic rules and Support Vector Machine (SVM), were used. The first technique (semantic rules) could deliver learning to students and also, could improve the quality content of education. This work helped the students who were at risk to avoid failure and help the good students improve their interest regarding education.

Abdullahet al. (2017) claimed that it has become easier to search data and get valuable knowledge from it because of the emergence of new technology which is enhanced by using data mining. In order to predict students' performance using their previous academic experience, data analytic techniques are applied to real case studies used in the previously mentioned research. A new hybrid classification technique was used which was based on fuzzy multi-criteria and decision tree classification. This technique is used in order to predict students' performance based on numerous principles such as school, age, family size, address, evaluation in previous grades and activities. In order to check the correctness of the model, the proposed method was compared with other famous classifiers. According to (Abdallah et al., 2017) study, this method is an assuring classification tool. Kostopoulos and Lipitakis (2017) worked on predicting the student's performance (fail and pass) in the final exam of a distance learning undergraduate course in HOU.

The effectiveness of active learning algorithms, J48 decision tree, JRip, Logistic Regression (LR), Multilayer Perceptron (MLPs), representative of Neural Networks, Naïve Bayes (NB) and Sequential Minimal Optimization (SMO), a very effective SVM algorithm, were examined in the aforementioned study. Early realization of the low performance of the students could lead to developing the personalized learning strategies accordingly with the students' need which also helps in enhancing the students' academic performance. The results illustrated that MLP algorithm outperformed the others with the accuracy of 81.09%.

Mahboob et al. (2016) Demonstrated that predicting students' performance is one of the most significant efforts in educational data mining. The study mentioned above indicated that it is possible to predict the success rate of the students, by using algorithms, such as Naïve Bayes, Random Forest

and Decision Tree J48. The research mentioned above aimed to find the most useful features. These features effects on students' performance during their academic years which lead them to drop out. The result illustrated that the model was beneficial for predicting the students' performance. Furthermore, the Random Forest algorithm gave an accuracy with 100%, J48 gave an accuracy with 93.3% and Naive Bayes gave an accuracy with 86.6%. Agaoglu (2016) showed that one of the most used tools in solving and understanding educational and administrative problems are data mining techniques in higher education. Research in educational data mining, focuses on building models which predict the students' performance. One of the most used tools in such models is the course evaluation. In this research, four classification tools; Support Vector Machine (SVM), Decision tree, discriminant analysis and Artificial Neural Network, were employed for evaluation and prediction. The results showed that the decision tree outperformed the others in terms of accuracy with 92.3%. The benefits of this research illustrate the expressiveness and effectiveness of such models in higher education course evaluation, also, these benefits may be utilized to enhance the performance of the students and teachers.

Mueen (2016) employed three data mining classification techniques; Naïve Bayes, Decision Tree and Multilayer Perception, in order to predict and analyze the students' academic performance in a specific course. All these algorithms were applied on students' information, which were collected from two semesters of graduate students. The results showed that the Naïve Bayes algorithm performed better than the others in predicting with an accuracy of 86%. The aforementioned study helps teachers find out the students who were expected to fail in the course and it also helps the instructors give more attention to students who are at risk and need to improve their performance.

Based on the previous works done, the most features used in their studies are illustrated in Table1. In this research, the authors used EDM algorithms, such as: Decision Tree, Random Forest, SVM and Naïve Bayes, in order to evaluate the students' performance in a specific course, during the semester or academic year and also predict their performance in the final exam. In addition, this research proposed a system as a new tool for classification and prediction of students' performance in the higher educational institutes. Next, this research added a function to Decision tree C4.5 in order to get more accuracy regarding to the classification and prediction of the students' performance.

Table 1: The most common features used by previous studies Features

Exams Marks			Children	
Lab Work			Marital status	
Assignment			Projects	
Gender			Seminars	
Online Resources		Work		
Quiz				Age
Dormitory			Participation	
Attendance			Family size	

Classification Algorithms

In the decision tree, Iterative Dichotomiser 3 (ID3) algorithm was created by Ross Quinlan in order to build a decision tree (Quinlan, 1986). Ross Quinlan Developed ID3 algorithm to C4.5 algorithm. Many researchers confirm that C4.5 is one of the most powerful algorithms in classification and outperforms others (Agaoglu, 2016; Hussain et al., 2018; Kiu, 2018). A recursive partitioning was used for building a decision tree. Divide and conquer is a general name for this process because it utilizes the values of the features to divide the data into smaller subsets of the same class. Furthermore, C4.5 can deal with discrete and continuous data, missing values, noise data and pruning tree after construction. C4.5 depends on entropy and gain ration in order to choose the best attribute as a root and then go on further splitting features.

Random Forest

Breiman (2001) has developed the Random Forest algorithm; both classification and regression can be done by Random forest. Random Forest can construct several decision trees in order to enhance the rate of classification and also to overcome the overfitting problem (Mahboob et al., 2016).

Random Forest is a data mining technique which utilizes the decision tree for classification. For creating a K number of not pruned trees, Random Forest each time chose a different part of the dataset. In Ran-dom Forest, the test data applied to all constructed trees and the most frequent output will be assigned to the tested data as a label (Mishra et al., 2014).

Random Forest has the idea of the real forest which says the more trees in the forest will have more robust. Also, Random forest will give the best accuracy if it has a higher number of trees in the forest. Random Forest does not pay attention to the number of trees in the forest, can handle missing values, never overfit and also can deal with categorical data. For measuring the purity and impurity of the features, Random Forest utilizes the Gini Index measurement.

Support Vector Machine (SVM)

Support Vector Machine is a supervised machine learning algorithm that analyzes data for classification and regression. SVM assigns the input values to one of the classified class based on previous training data. In SVM, the classes are separated by a plane or hyperplane which maximizes the margins between data and minimizes the classification error (Agaoglu, 2016). There are three types of SVM; Hard margin SVM (Linear), Soft margin SVM (Linear with outliers) and Kernel SVM (Non-Linear) (Boser et al., 1992). Many studies used SVM in order to predict the students' performance (Costa et al., 2017; Kadambande et al., 2017; Rao et al., 2017), in which some of them the algorithm outperformed the others in terms of classification (Costa et al., 2017; Kadambande et al., 2017).

Naive Bayes

Naive Bayes techniques are a probabilistic classifier's simple family which is based on Bayes theorem with very powerful (Naive) independence assumption among all features. In machine learning, Naive Bayes uses the theory of

probability that considers the effect of the value of the parameter of a given class (Pandey and Pal, 2011). Instead of a deterministic relationship where not always the same attribute values have the same output label, in Bayesian classifier the probabilistic relationship is considered among all classes and attributes. Attributes can be classified depending on its values, "can be expressed as the probability of a record of being from the class Y, given that the record has a set of attributes X. That is, P(Y = y |X = x). A record is assigned to the class with the largest probability" (Reason, 2009). Additionally, Naive Bayes is highly scalable, fast, easily trained on a small dataset and can be used for both binary and multiclass classification. This technique was used in many studies, such as: (Costa et al., 2017; Rao et al., 2017; Sivakumar et al., 2018) and performed better than others in terms of classification in (Lopez Guarin et al., 2015).

Research Methodology

Data Collection

In this study, the data has been collected in two steps. In the first step, the most related works were reviewed by the researcher and find out what are the features used in their studies. Then, extracted the most common features used among them, which could be found in the survey area of this study. In the second step, the researchers interviewed (30) academic staff to get their ideas for selecting the most wanted and effected features on students' performance. These academic staffs were from computer science department in three universities (the University of Zakho, University of Duhok and Duhok Polytechnic University) in the Duhok province, Kurdistan Region, Iraq, as a case study. In the experimental study three datasets were used, two of them taken from academic staff as a case study and another one taken from the study of (Cortez and Silva, 2008) for more proof.

Additionally, a grid search is a process of finding the best suitable parameter for the model based on the type of utilized model. Grid search can be applied to any machine learning algorithm in order to calculate the optimal parameters for an existing model. A grid search will construct a model for each combination point (Evan Lutins, 2017). Thus, in our improved algorithm the grid search function will be used to find the best two parameters: BestJoin and MaxHeight parameters.. This parameter has not been mentioned before by any of the researchers in the literature of this study. For this reason, BestJoin and MaxHeight parameters are used in order to enhance the performance of the algorithm. Indeed, in the improved algorithm of C4.5 each set will be the value of BestJoin and MaxHeight respectively. Figure 1, the explanation of the process of this model is given.

Proposed Prototype

In this research, for classifying and predicating the students' performance, the researchers used a new tool which is called Acoord.Net library in visual studio by using c#. Accord.Net is a scientific computation framework in .NET. The Accord.net framework contains many libraries that include a lot of scientific computing applications, such as: Machine learning, statistical data processing, pattern recognition, computer

audition and computer vision. This framework presented a huge amount of hypothesis tests, support of popular performance measurement techniques, kernel functions and probability distribution.

The Accord.Net library in visual studio was programmed by C#, which was constructed initially in 2010 and stabilized in October 2017 (Souza, 2017). In this proposed prototype (Fig. 2), the user can upload the dataset from excel sheets to the application. After that the user can select the class label and start applying the classification algorithms on that dataset.

Evaluation Methods - Cross-Validation :

A cross-validation is a statistical tool used to evaluate the learning algorithms. In cross-validation, the data are split into two parts. The first part is the training part, which used to train the model. On the other hand, the second part is used to validate the model. The process of cross-validation depends on the cross over the training and validation data in the sequent round. This process used to give each data point a chance to be a validate data within the training data (Mankovskii et al., 2009).

One of the primary types of cross-validation is K-Fold cross-validation. Ten-Fold cross-validation is a form of K-Fold cross-validation and it is the standard method used to measure the error rate of a model. In ten-fold cross-validation, the data is separated into ten equal folds and then nine folds will be used for training the model.

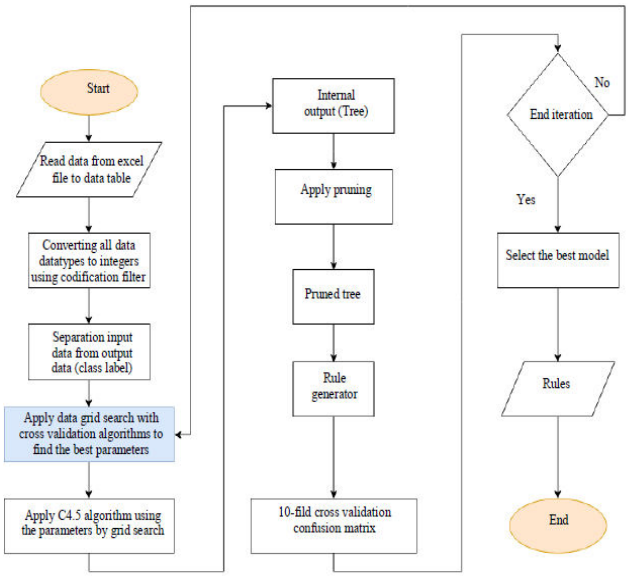

Fig. 1: C4.5 algorithm with data grid search

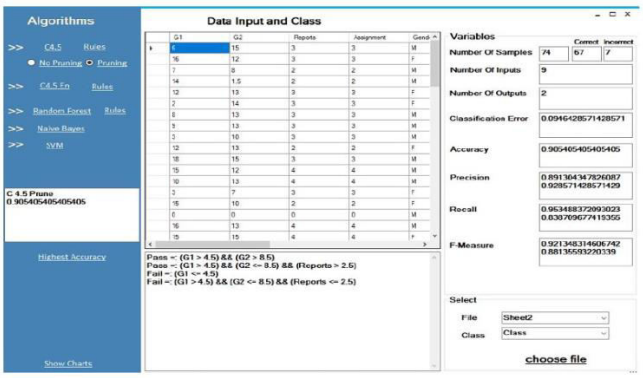

Fig. 2: Proposed prototype

Prediction Process

Prediction is the process of selecting the right class label for new unseen records based on the trained data from the classification models. While the prediction process in Accord.Net Machine learning library is only dependable for predicting one by one record in form of a string. Thus, in this research, the prediction process begins by loading the excel file which contains the data that is wanted for prediction. Then the loaded data datatypes will be converted into integers, this converting is done in order for the system to make predictions for an unlimited number of records in sequence.

Results' Discussion

After applying the classification algorithms on the two datasets (Dataset 1 and Dataset 2), the results showed that the improved Decision Tree C4.5 had performed better than standard C4.5 and also when compared to C4.5 (J48) in Weka tool. Furthermore, the improved Decision Tree C4.5 was applied to the dataset (Cortez and Silva, 2008) study. The results showed that the improved Decision Tree C4.5 had outperformed the Decision Tree C4.5 in the study mentioned in previous chapter. Table 11 gives more details about the comparison of performance of the algorithms used in this work.

Conclusion

The information of students in higher education institutions increases annually. It is difficult for them to take out meaningful information from huge amount of data manually. Such information can support academic staff in helping prevent students from dropping out at the end of courses. This can be achieved by evaluating students' performance for the course, as well as predicting their performance in the final exam early on by using classification algorithms. Four classification algorithms were used in this study; Decision Tree C4.5, Random Forest, Support Vector Machine (SVM) and Naive Bayes, in order to classify and predict the students' performance. Furthermore, this research aimed at improving the Decision Tree C4.5 algorithm by adding a grid search function in order to get more accuracy in classifying and predicting the students' performance. Also, the features of this evaluation were extracted through the interviews with academic staff and through the review of the literature. Three datasets were utilized in this research in order to test the improved Decision Tree C4.5 with traditional C4.5 and others. The results showed that the improved Decision Tree C4.5 outperformed the traditional C4.5 and also performed better when compared to C4.5 (J48) in Weka tool and other algorithms used in this research. Further research and improvements are required to generalize the proposed prototype to be a bedrock system of evaluating a big dataset of students in the educational field by researchers. Improving the system by adding unsupervised learning algorithms to support the unlabeled dataset or make the dataset ready for supervised learning algorithms.

Acknowledgment

This work was supported by the University of Duhok, University of Zakho and Duhok Polytechnic University. I would like to thank these universities for their help, support and contribution during the process of research investigation.

Authors Contribution

- **Mohammed Hikmat Sadiq:** Participated in all works done in this research such as, Improving the decision Tree C4.5 algorithm by adding the grid search function and also helping in organizing question for interview in order to extract the most useful features related to students' performance in our institutions. Comparing results.
- **Nawzat Sadiq Ahmed:** Participated in organizing the interview question for feature selection which related to students'

performance, also he participated in organizing this article according to this journal conditions.

References

I. Abdallah, F.A., T. Anna and D. Babar, 2017.
II. Agaoglu, M., 2016.
III. Baker, R.S.J., 2010.
IV. Hussain, S., N.A. Dahan, F.M. Ba-Alwib and N. Ribata, 2018.
V. Evan Lutins, 2017. Grid searching in machine learning.
VI. Cortez, P. and A. Silva, 2008.
VII. Baker, R.S.J., 2010.
VIII. Boser, B.E., I.M. Guyon and V.N. Vapnik, 1992.
IX. Costa, E.B., B. Fonseca, M.A. Santana, F.F. de Araújo and J. Rego, 2017.

About Authors:

[1]**Mr. Nagaraju Hemanth Kumar** is currently pursuing MCA in Siddharth Institute of Engineering & Technology, Puttur, Andhra Pradesh, India.

[2]**Mr. J.S. Ananda Kumar,**Assistant Professor in Dept. of MCA, Siddharth Institute of Engineering & Technology, Puttur, Andhra Pradesh, India.