ISSN:0975-9646

Graduate Prediction Using Ontology Approach

Tirth Mehta, Tanmay Kshirsagar, Azeem Merchant, Professor Sindhu Nair

Computer Engineering Department, Mumbai University
Dwarkadas J. Sanghvi College of Engineering
Plot No. U-15, J.V.P.D. Scheme, Bhaktivedanta Swami Marg, Vile Parle (West), Mumbai-400056, India.

***Abstract*—The paper seeks to predict the completion of graduation of students studying in a University. The filtration of the redundant features will be done by Ranker method. The use of SVM method will also be taken into consideration for filtration of attributes. The decision tree algorithms C4.5/J48 and ID3 are used for generating the decision tree which will serve as a basis for the generation of an ontology. Moreover, the classification rules are extended by applying a semantic based approach for creating classification tree ontology. The ontology represents the classification rules that are used to enable machines to interpret and identify learner factors in process of prediction. The explanation of how ontological representation plays a role in classifying students to predictive target class is done. Semantic Web Rule Language is used for the specification of classification rules.**

***Keywords*— Decision Tree, Graduate Prediction, Ontology, Semantic Web, SWRL**

I. INTRODUCTION

University of Minho, a university in Portugal, admits students under various programs from different educational backgrounds, schools, places and various educational metrics in entrance tests. Every year, the number of unsuccessful students who have not completed their bachelor's degree has increased. The survey reports about 15-18% failure of students to successfully complete their academic year. The main aim is to identify all the essential and crucial parameters needed for the successful completion of their graduation degree. Classification data mining techniques ID3 and C4.5 are used to generate the predictive model.

The Semantic Web is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web, most fundamentally the Resource Description Framework (RDF) [1]. Ontologies are generally defined as a representation of a shared conceptualization of a specific domain. It is the main part of the semantic web. Recently many works have been applied in the field of semantic web for several aspects of their domain, such as task ontology, user model ontology, domain ontology and others [2].

Semantic web technologies are used to represent the classification rules. They are further used to identify various important parameters for the completion of graduation. SWRL can be used to express rules and logic. It is a combination of OWL DL and Rule Markup Language. It has been used for determination of the classification rules. The rest of the paper is organized as follows. The section II presents a proposed concept and methodology for the implementation. Section III provides the explanation of the architecture. Section IV gives the in-depth view of the flow chart. Finally, the conclusion of this paper comes up with giving a summary of its main contribution and looking forward to future research work.

II. CONCEPT AND METHODOLOGY

A. Training Dataset and Data Pre-processing:

For generating a classification model, the preparation of training datasets consisting of information about students is performed [3]. This data has details of each student such as full name, primary school, GAT score, GPA of each semester, the parent's background, etc. 32 features are used and they are stored in a training database. The scanning of the dataset for managing the incomplete or inconsistent data is carried out. Next, the data will be coded into defined feature's label. Data transformation process is required to transform continuous data into interval-data. Finally, the creation of the input of training dataset into ARFF file format has been done and then it is imported to WEKA

B. Feature Selection:

In this process, removal of the irrelevant features from the dataset by using a Greedy Stepwise feature selection, Ranker Method and SVM is carried out. The Ranker method classifies all the attributes by associating a rank to each attribute. The rank specifies the contribution of that attribute towards prediction analysis of the graduate students. By removing the attributes with the lowest ranks, the aim is to improve the predictive percentage of the system. For example, the name of the student plays no role in predicting whether, the student will complete the graduation or not. Hence, this attribute will be associated with a low rank. This attribute therefore can be removed from the dataset and excluded during the creation of the Ontology using protégé. On the other hand, attributes like entrance exam score and GPA play a very important role in predicting the completion of the graduation of a student. This will therefore be associated with a very high rank. The data will be filtered according to a set of rules based on the above mentioned methods. This has to be done in-order to increase the accuracy percentage of prediction.

C. Model Construction and Evaluation:

In WEKA, the classification algorithms are used to create decision trees. For evaluating the model effectiveness, a binary (positive and negative) classification is used. There are four possible outcomes of classifier prediction: True Negative (TN), True Positive (TP), False Negative (FN) and False Positive (TP). A two-plan confusion matrix is explained in the table (below).

TABLE I
CONFUSION MATRIX

	Correct Result	
Obtained Result	**+(Success)**	**-(Fail)**
+(Success)	TP	FP
-(Fail)	FN	TN

The four values TN, TP, FN and FP provided by the confusion matrix that are well known and are typically used for determining, where N represents the number of instances in a given set. Mean Absolute Error (MAE) is the average of the difference between predicted and actual value in all test cases; it suggests that the error rate is very small, which can be considered as a measure of effectiveness of the model. It can be calculated by the following equation

$$MAE = (|a_1-c_1|+|a_2-c_2|+|a_3-c_3|+...+|a_n-c_n|)/N$$

Where N represents the number of records in a dataset. This value is computed by taking the average of the absolute differences between each computed (c_i) and its corresponding correct value (a_i).

Then, ontology approach to model the context of the decision tree rules and the student model has been used. The explanation of this approach will be described in next subsection.

D. Student Model Ontology

Based on the decision tree created using WEKA [4] tool, it is first needed to identify the classes which will be included to create the ontology. Next, the object properties associated with the above mentioned classes will be the ones shown in the decision trees. This is the way the entire process of mapping the decision tree to the Ontology will be successfully taken into action. Thus the final step is to create the ontology with all the data properties of the finally considered data set. The entire ontology is implemented using the Protégé 4.3 alpha tool[5].

E. Inference Method

The inference itself is a task of the used rule engine which can have different algorithms implemented for it. The inference method that decision tree algorithms are based on will be described next, namely successful graduate predictive rules. The predictive rule ontology, student ontology and the part of representation has been implemented using backward chaining reference method. Next, the implementation of the rules in Protégé is done with the help of SWRL tab.

III. ARCHITECTURE

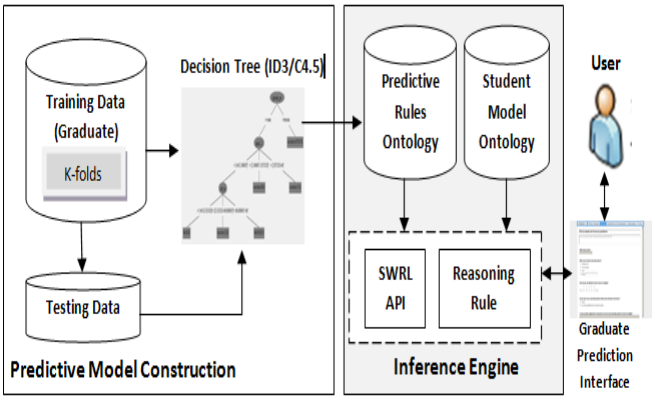

Fig 1. Architecture of Graduate Prediction System

The basic process can be divided into 2 stages as follows:

Stage 1:

Firstly, start with creating a dataset. The next step is to filter the redundant attributes and apply the classification algorithms to create the decision tree along with the confusion matrix. This process is done with the help of the tool WEKA. The output of the stage 1 is the decision tree.

Stage 2:

This includes mapping the output of stage 1 (decision tree) to the input of stage 2 (ontology development). This takes into consideration, the identification of classes which will make the ontology and then further adding the object properties as observed in the decision tree. These object properties are the nodes of the decision tree that act as necessary parameters for graduate prediction. After filling the data in the ontology, SWRL [7] is used as an inference layer to specify the classification rules. This concludes the final stage 2 as well as the entire process. In the end, the prediction of the final completion of graduation of the students from the University can be successfully carried out.

IV. FLOWCHART FOR THE GRADUATE PREDICTION PROCESS

The flowchart shows the entire process for the graduate predictive analysis of the students in the University. The process mentioned in the flowchart is also divided into 2 sections. The part mentioned up to the 'Visualization of the Decision tree' aims at obtaining the logic. The second section, up to the end, is the execution of the logic in the form of an ontology. The flowchart is provided below with all the details:

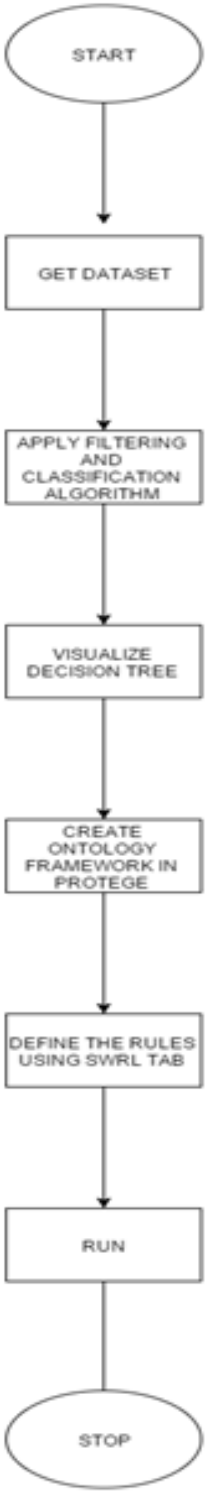

Fig 2. Flowchart for Graduate Prediction System

V. CONCLUSION AND FUTURE WORK

This paper has described the system which is used to predict the results of completion of the graduation degree of students. The classification algorithms in data mining are used to generate the rules. Graduation factors are identified by an experiment result. The use of SWRL creates a set of rules for the final result which is used to predict the success or failure for each individual student.

In the future, the implementation of the process will be taken into consideration by training the prediction parameters with the help of including new training sets. A Learning Management System (LMS) will also be used to provide a decision support module for a current student.

REFERENCES

[1] https://www.w3.org/RDF/

[2] OWL Web Ontology Language Reference, http://www.w3.org/TR/owr-ref

(Accessed: June 13, 2013)

[3] Using data mining to predict secondary school performance, Paulo Cortez.

[4] WEKA: http://www.cs.waikato.**a**c.nz/ml/weka

[5] Protégé Ontology and Knowledge Base Framework http:///protege.standford.edu. (Accessed: May 25, 2013)

[6] A Semantic based approach for representing successful graduate predictive rules, Thaksin University.

[7] Wikipedia SWRL: A Semantic Web Rule Language- Combining OWL and Rule ML, https://en.wikipedia.org/wiki/Semantic_Web

[8] M. Kantardzic, Data Mining: Pearson Education Inc, 2003.

[9] A. A. Freitas, Data Mining Evolutionary Algorithms, Naturu York, 2002.

[10] R. J. Roiger and M. W. Geatz, Pearson Education Inc., 2003.

[11] J. Han and M Kamber, Data Mining, Concept and Techniques, Elsevier, 2006.