A Better Prediction for Higher Education Performance using the Decision Tree

Anwer Mustafa Mohamedsalih Hilal[1, 3,*], Abu Sarwar Zamani[1], Muhammad Shahid Ghulam Farid[2], Mohammed Rizwanullah[1]

a.hilal@psau.edu.sa, a.zamani@psau.edu.sa, mfarid@su.edu.sa, r.mohammed@psau.edu.sa

[1]Department of Computer and Self Development, Preparatory Year Deanship, Prince Sattam bin Abdulaziz University, AlKharj, Saudi Arabia

[2] Computer Science Department, Huraymila College of Science and Humanities, Shaqra University, Saudi Arabia

[3]Faculty of Computer Science and Information Technology, Omdurman Islamic University, Omdurman, Sudan

*Corresponding Author: Anwar Hilal, Email Address: a.hilal@psau.edu.sa

Abstract

Data mining is the application of specific algorithms for extracting patterns from data and KDD is the automated or convenient extraction of patterns representing knowledge implicitly stored or captured in large databases, data warehouses, the Web, other massive information repositories or data streams. Data mining can be used for decision making in educational system. But educational institution does not use any knowledge discovery process approach on these data; this knowledge can be used to increase the quality of education. The problem was happening in the educational management system, but to make education system more flexible and discover knowledge from it huge data, we will use data mining techniques to solve problem.

Key words:

Data Mining, KDD, Algorithm, Extracting Patterns.

1. Introduction

Education is an essential element for the betterment and progress of a country. It makes the people of a country civilized and well mannered. As we know, large amount of data is stored in educational database; data mining is a process of discovering valuable information from large amount of data stored in databases, data warehouses, or other information repositories. Data mining is an essential part of knowledge Discovery Database (KDD), KDD stands for as the non-trivial process of identifying valid, novel, potentially useful information and finally understandable pattern in data [1].

Data mining including classification rules or trees, regression, clustering, sequence modeling, dependency, and so forth approaches, in this research I focused in classification task as data mining techniques.

As data mining proficiency, a decision tress is one of the most commonly used methods acting for knowledge discovery. A decision tree is used to describe rules and relations by consistently damage and rubricating the information restrained in data. The characteristics of the decision tree are easy to understand and a simple top-down tree structure where decisions are made at apiece node. As a result, the nodes under the tree provide the result. [2]

Data mining and KDD is forcefully valuable in virtually any industrial and business sectors where database and information technology are used. Data mining applied in many applications such as fraud spotting, investment analysis, portfolio trading, and marketing and sales data analysis. [3]. Nowadays, data-mining techniques are used in education system to enhance learning management system. In this direction, some models have been proposed and implemented. The authors of [9] have proposed a model to represent how data mining can be used in a higher educational system to improve the efficiency and effectiveness of the traditional processes. In the model, several processes are proposed to be enhanced through data mining functions. The

Model is also presented as a guideline for higher educational system to improve the decision-making processes.

2. Literature view

In this way, some models have been proposed and implemented. The research by [4] has used decision tree as an assortment approach to generated rules that allows students to predict the final grade in which falls under the study and courses, evaluated the students data using WEKA toolkit data are collected from students those studying in University. The research by [5] they proposed models for quislings' management to characterize similar behavior groups in unstructured quisling's spaces, they mining students data using clustering to discover patterns reflecting user behavior.

The research by [6] apply data mining clustering for web learning to promote group-based collaboration learning and to provide incremental learner diagnosis, they

Manuscript received April 5, 2021
Manuscript revised April 20, 2021
https://doi.org/10.22937/IJCSNS.2021.21.4.25

find clusters of students with similar learning characteristics base on the sequence and the contents of the pages they visited. The research by [7] Implement decision tree C5.0 algorithm and data cube technology from web log portfolios for managing classroom processes. The induction analysis finds out potential student’s groups that have similar symptoms and reactions to a specific educational strategy. In [2] Hsu, P.L. et al, propose a prototype system based AGA (Association-based Genetic Algorithm) approach is developed to predict the learning performance of college students. Application data, including student learning profiles and related course information, was obtained from a reputed University.

The research by [8] the proposed model has been used by writing code of assignment to extract all the information of the students' and in order to increase students' performance, writing some code is used to find statistical pattern or predictions. The data were gathered as well as studied in a concurrent version system (CVS) by several students working on a small project in a final year undergraduate computer science students. The main use of data mining has been done in this research for the purpose of education system. The classification task used at the secondary level to estimate students' future discipline using a decision tree. The data is used in this research is restricted to those students in Shaqra University of Saudi Arabia in 2019.

3. The Proposed Model

To up building a classification model, data mining process that consist of several steps: firstly, collecting data (features) from problem under study, then data that collected in the later step needs to prepared by selecting the relevant features and so, then finally using such models to discipline students in the future such as building a classification model, as well as using one of the assessment methods. In this step, I designed questionnaire to collect data from students in the first year in college, that College I was distributed questionnaire over them existing in Shaqra University, KSA. The questionnaire has questions about relevant features that effect in selecting the students that discipline in college. Initially 19 attributes have been collected about subject, and then irrelevant attributes manually eliminated to get more accuracy about problem under study. Finally, only 9 attributes and one class remained to using in classification model, the attributes and their description and possible value for each attributes are presented in Table 1. The class attributes are the student discipline (Discipline).

Table 1: the symbolic attribute description

Attribute	***Description***	***Possible values***
Gender	Student Gender	Male, Female
St_Fath	Level of Study for Father	nrw, Elementary, Secondary, University
St_Moth	Level of Study for Mother	nrw, Elementary, Secondary, University
Income *	Income of Family	Low, Medium, High
Family-size **	Size of Family	Small, Large
Avg-8-9-10 ***	Average of Students in 8,9,10 grades	A, B, C, D
Family-select	Family Select	Scientific, Anthology, H-edu, MIS
Your-Select	Student Select	{Scientific, Anthology, H-edu, MIS, Kids-edu, Industry}
Who-Select	Who Select Discipline	I, Family, Friends, other
1.1.1.1.1 Discipline	The Discipline of Student in college level (The Class).	Scientific, Anthology, H-edu, MIS

Notes: *Income: Low= (Less than 200 JD), Medium= (between 200 and 350),
High= (More than 350 JD).
**Family-Size: Small (No. Of Family <=7), Large (No. Of Family >7).
***Avg-8-9-10: A= 90-100, B= 80-89, C= 70-79, D= 50-69.

The collected data in the previous step we need to converted it to suitable form to using next by data mining techniques, the most important attributes list have the following attributes: Your-Select, Who-Select, Family-Select, St_Moth.

Classification is one of the most frequently studied problems by data mining researchers. It consists of predicting the value of a categorical attribute based on the value of other attributes. Classification methods like decision trees, rule mining, Bayesian network etc. can be applied on the educational data for predicting the students behavior and performance.

The decision tree method comprehends a number of specific algorithms, including Classification and regression tree, Chi-squared Automatic Interaction Detection (CHAID), C4.5 and C5.0 [3]. The decision tree using ways to get information to select the most useful (Best) attribute at each node in the tree. The information gain depends on entropy measure. [1]

The best attribute to this study is Your-Select (The Student Select). This feature considered the root node of the tree and then to the next level we need to calculate the most useful attribute until complete the decision tree, the set of classification rules generated, following all the paths of the tree where the generated tree has created 18 classification rules. The generated rules are given in the Table 2.

Table 2: The Generated Classification Rule

Rule #	Rule	# Obj	# Attrib
1	If your-select=scientific then discipline=scientific	61	2
2	If your-select=h-edu and family-select=h-edu then discipline=h-edu	33	3
3	If your-select=anthology then discipline=anthology	29	2
4	If your-select=mis and who-select=I then discipline=mis	29	3
5	If your-select=industry then discipline=anthology	5	2
6	If your-select=h-edu and family-select=scientific and st_moth= elementary then discipline=h-edu	4	4
7	If your-select=h-edu and family-select=scientific and st_moth= secondary then discipline=scientific	4	4
8	If your-select=mis and who-select=family and family-select=scientific then discipline=scientific	3	4
9	If your-select=mis and who-select=other then discipline=anthology	3	3
10	If your-select=kids-edu then discipline=anthology	3	2
11	If your-select=h-edu and family-select=anthology then discipline=scientific	2	3
12	If your-select=mis and who-select=family and family-select=anthology then discipline=anthology	2	4
13	If your-select= mis and who-select=family and family-select=h-edu then discipline=h-edu	2	4
14	If your-select=mis and who-select=friends then discipline=anthology	2	3
15	If your-select= h-edu and family-select=scientific and st_moth= university then discipline=mis	1	4
16	If your-select=h-edu and family-select=mis then discipline=mis	1	3
17	If your-select=mis and who-select=family and family-select=mis then discipline=mis	1	4
18	If your-select=h-edu and family-select=scientific and st_moth= nrw then discipline=scientific	0	4

In Table 2, the first column represents the number of rule, the second column are represent the generated rules, the number of students how satisfy the rules is given in the third column, and the last column is given the number of attributes contained in the rules. The table shows the rules in descending order based on the number of students who successfully complete the rule. This order is important to indicate the most significant rule and then indicate the most significant attributes.

4. Experiments and Evaluation

As described in [1], using training data to derive a classifier and then estimate the accuracy of the classifier, in this study we have 185 instances as a data set. Holdout and cross-validation are two common techniques for assessing classifier accuracy. The WEKA toolkit used along with a classification model, to achieve accuracy. To acting of many classification methods are used to test accuracy of classifier as shown in Table 3.

Table 3. Classification Accuracy of different algorithms

Algorithm	Hold out 70%	10 CV
ID3	80.3571%	78.9189%
C4.5	82.1429%	83.2432%
Naïve Bayes	85.7143%	83.7838%
IBK	80.3571%	81.6216%
Kstar	83.9286%	82.1622%

LBR	85.7143%	83.2432%
RepTree	85.7143%	81.6216%
Decision Table	82.1429%	81.0811%
VFI	82.1429%	84.3243%

The results that came, we have to see that the accuracy of different classification methods is high and analogy.

5. Conclusion and Future Work

This research completed to other research before it, and the goals of this research to increase using of data mining techniques in this field (Learning Management System), to facilities students to indicate the best choice in the future, and to use this system from other decision maker to make decisions more flexibilities and more accuracy. When generated the classification rules are facilitates to the students to predict the discipline in future in secondary stage. An attractive task in the future is to collect large and real student database of any university and apply as well as implement using these kind of model and make more powerful and get high accuracy, and applied this data in another data mining technique. Moreover, several other classification methods can also be applied to test the most suitable method that suit the structure of the student data and give a better classification accuracy. Data mining tools are typically designed more for power and flexibility for simplicity. Most current data mining tools are too complex for educators to use and their features can go beyond the scope of what an educator wants to do. Therefore, these tools should be more comfortable and easy use interfaces with interface-free data mining algorithms to make configuration and execution easier, and with good visualization features to create their results meaningful for teachers and e-learning designers.

Acknowledgments

The authors would like to thank the Deanship of Scientific Research at Prince Sattam Bin Abdulaziz University, Alkharj, Saudi Arabia for the assistance.

References

[1] Al-Radaideh, Q.A., Al-Shawakfa, E.M., and Al-Najjar, M.I. (2006) “Mining Students Data Using Decision Trees”, The 2006 International Arab Conference on Information Technology.

[2] Chen, G., Liu, C., Ou, K., and Liu, B. (2000) Discovering Decision Knowledge from Web Log Portfolio managing Classroom process by applying decision tree and data cube technology. Journal of Education Computing Research, 23(3), pp. 305-332

[3] Han, J., Kamber, M. (2001) Data mining: Concepts and Techniques. USA: Morgan Kaudmann Publishers.

[4] Hsu, P.L., Lai, R., Chiu, C.C., and Hsu, C.I (2003) “The hybrid of association rule algorithms and genetic algorithms for tree induction: an example of predicting the student course performance”, Expert System with Applications, 25, pp. 51-62

[5] Mierle K., Laven K., Roweis S., and Wilson G. (2005) Mining Student CVS Repositories for Performance Indicators.

[6] Romero, C., Ventura, S. (2007) “Educational data mining: A survey from 1995 to 2005”, Expert Systems with Applications, 33, pp. 135-146

[7] Tang, C., Yin, H., Li, T., Lau, R., Li, Q., and Kilis, D. (2000) Personalized courseware construction based on web data mining. In Proceeding of the first international conference on web information system engineering, Washington, DC, USA, pp. 204-211

[8] Zhang, C., Zhang, S. (2002) Association Rule mining. Springer-Verlag Berlin Heidelberg.

[9] Delavari N, Beikzadeh M. R. A New Model for Using Data Mining in Higher Educational System, 5th International Conference on Information Technology based Higher Education and Training: ITEHT ’04, Istanbul, Turkey, 31st May-2nd Jun 2004.

Anwer Mustafa Hilal is an Assistant Professor of Computer Science in the Department of Computer and Self Development at Prince Sattam bin Abdulaziz University. He obtained his PhD degree in 2017 from Omdurman Islamic University, Khartoum, Sudan, for his thesis entitled A Semantic Data Mining Model for Exploring the Holy Quran. His research interests include data mining, text mining and mobile and Web development.

Abu SARWAR Zamani started his onerous career from his native place Bihar, India. Currently working as an Assistant Professor in Prince Sattam Bin Andulaziz University, Kingdom of Saudi Arabia (KSA). Worked as a Senior lecturer in College of Science & Humanity in Shaqra University, KSA and as a lecturer in Computer Sc. Department in King Saud University, KSA.

Before Onset of teaching, he have onus as a software engineer for two years in a Multinational Software Company, India. He persuaded his PhD in 2019 from Pacific University, Udaipur, India. He was honored with

Master of Philosophy in Computer Science from Vinayak Mission University, Chennai, India in 2009 with aloof mode. Moreover, received his Master of Computer Science in 2007 from Jamia Hamdard (Hamdard University), New Delhi, India). He has assiduously attended and published various research papers in national as well international.

Muhammad Shahid is a Lecturer of Computer Science in computer science department and self-Development in Huraymila College of Science and Humanities at Shaqra University. He obtained his Master of sciences in Information Technology in 2009 with Distinction from Punjab University of College of Information Technology, Pakistan. His Research interest in Network security, Internet of Things, cloud computing, programming skills, image processing and Database security systems.

Mohammed Rizwanullah is a Lecturer of Computer Science in the Department of Computer and self Development in DPY at Prince Sattam Bin AbdulAziz university. He obtained his Master of Technology in 2010 with Distinction from Jawaharlal Nehru Technological University, Hyderabad, India. His Research interest in Network security, Internet of Things, cloud computing, image processing and Database security systems.