World Academy of Science, Engineering and Technology
International Journal of Information and Communication Engineering
Vol:9, No:7, 2015

Using Data Mining Technique for Scholarship Disbursement

J. K. Alhassan, S. A. Lawal

***Abstract*—**This work is on decision tree-based classification for the disbursement of scholarship. Tree-based data mining classification technique is used in other to determine the generic rule to be used to disburse the scholarship. The system based on the defined rules from the tree is able to determine the class (status) to which an applicant shall belong whether Granted or Not Granted. The applicants that fall to the class of granted denote a successful acquirement of scholarship while those in not granted class are unsuccessful in the scheme. An algorithm that can be used to classify the applicants based on the rules from tree-based classification was also developed. The tree-based classification is adopted because of its efficiency, effectiveness, and easy to comprehend features. The system was tested with the data of National Information Technology Development Agency (NITDA) Abuja, a Parastatal of Federal Ministry of Communication Technology that is mandated to develop and regulate information technology in Nigeria. The system was found working according to the specification. It is therefore recommended for all scholarship disbursement organizations.

***Keywords*—**Decision tree, classification, data mining, scholarship.

I. Introduction

DATA mining is the computational procedure of discovering patterns in large data sets connecting methods at the intersection of artificial intelligence, machine learning, statistics and database systems. The main goal of the data mining process is to mine information from a data set and convert it into an understandable structure for further use. Market based analysis, sales promotion exercise using association, sales promotion analysis with regression analysis, fraud detection with classification, managing customer relationship with clustering, stock level management with time series, among others are example of application of data mining technique. The importance of choosing eligible candidates in disbursement of scarce resources for the advancement learning brought about this research work. This work used the data mining technique for disbursement of scholarship for postgraduate studies**.**

A decision tree is a simple inductive learning formation, which is one of the data mining techniques. Given an instance of an entity, which is precise by a set of properties, the tree returns a "yes" or "no" decision about that occurrence. Each inner node in the tree stands for a test on one of those entities, and the branches from the node are categorized with the likely outcomes of the test. Each leaf node is a Boolean classifier for the input instance. In these tree arrangements, leaves stand for classifications and branches stand for conjunctions of features that guide to those classifications. The machine learning technique for inducing a decision tree from data is referred to decision tree learning, or (colloquially) decision trees [1]. Another technique used for data mining is the artificial neural network (ANN); it is a mathematical form or computational model based on biological neural networks. It is made up of an interrelated group of artificial neurons and processes information by means of a connectionist approach to computation. In most cases an ANN is an adaptive method that changes its arrangement based on external or internal information that flows in the course of the network for the period of the learning phase [1]. Bayesian Learning is a probabilistic approach to learning and deduction, which is as well used for data mining. It is based on the supposition that the quantities of interest are controlled by probability distributions. It is good-looking because in theory it can turn up at best decisions. It provides a quantitative approach to weighing the substantiation supporting option hypotheses. Machine Learning domain ANN approaches is measured to be a baseline technique for data-driven modeling. The higher Machine Learning techniques comprise of Support Vector Machines. They are autonomous of the dimensionality of the input attribute space and permit development of robust non-linear models with good generalization abilities. These methods were freshly introduced in the scope of Statistical Learning Theory. According to the inductive learning principles of the Statistical Learning Theory, the optimal predictive model for a given data modeling task has to be built by finding the trade-off among the model complexity and its fit to training data. It gives rise to outstanding generalization abilities of the Support Vector Machines [1]. Instance based learning methods differ from other approaches to function roughly because they delay processing of training examples until they have to label a new query instance. Advantages of instance based way include the ability to model complex target functions by a compilation of less complex approximation and thus preventing the information available in training instances be lost. K-nearest neighbour algorithm (k-NN) is a technique for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is delayed until classification. It can also be used for regression [1]. Data mining is the act of discovering meaningful patterns, data turns into information. Information or patterns that are novel, valid and potentially

J. K. Alhassan is with Department of Computer Science, Federal University of Technology, Minna, Nigeria phone: +2348035961620; e-mail: jkalhassan@ futminna.edu.ng).

S. A. Lawal is with National Information Technology Development Agency, (NITDA), Abuja, Nigeria (e-mail: sarafalawa2011@gmail.com).

useful are not merely information, but knowledge. Data mining have two primary goals prediction and description. Prediction involves predicting unknown or future values using some variables or fields in the data set. Description involves finding patterns describing the data that can be interpreted by humans. Data mining is also an integral part of knowledge discovery in databases (KDDs) which includes several steps; Data cleaning, which is also known as data cleansing; it is a phase in which noise and irrelevant data are removed from the collection. Data integration combines data from multiple and heterogeneous sources into one database. Data selection, which allows the user to obtain a reduced representation of the data set to keep the integrity of the original data, set in a reduced volume. Data transformation, the selected data is transformed into suitable formats. Data mining, analysis tools are applied to discover potentially useful patterns. Pattern evaluation, identifies interesting and useful patterns using given validation measures. Knowledge representation, the final phase of the knowledge-discovery process, discovered knowledge is presented to the users in visual forms. The primary data-mining tasks are described; Classification which is discovery of a predictive learning function that classifies a data item into one of more predefined classes. Regression is discovery of a predictive learning function, which maps a data item to a real value prediction variable. Clustering is identifying a finite set of categories or clusters to describe the data. Summarization is a descriptive task that involves methods for finding a compact description for a set (or subset) of data. Dependency Modeling is finding a local model that describes significant dependencies between variables or between the values of a feature in a data set. Change and deviation detection is discovering the most significant changes in the data set. The most important issue that determines the quality of a data mining technique is the choice of data representation, and selection, reduction or transformation of features. Data reduction techniques can be applied to obtain a reduced representation of the data set that is much smaller in volume and closely maintains the integrity of the original data. The data reduction should be performed prior to applying data mining; the basic operation in data reduction are delete a column, delete a row, and reduce the number of values in a column (smooth a feature). Feature selection (also known as subset selection) is a process commonly used in machine learning, wherein a subset of the features available from the data is selected and irrelevant, weakly relevant or redundant attributes or dimensions may be detected and removed. The best subset contains the least number of dimensions that most contribute to accuracy. Mining on a reduced set of attributes has an additional benefit. It reduces the number of attributes appearing in the discovered patterns, helping to make the patterns easier to understand. Basically, choosing the most relevant features to achieve maximum performance with the minimum measurement and processing effort. A feature-reduction process should result in less data so that the data mining algorithm can learn faster; higher accuracy of a data mining process so that the model can generalize better from data; simple results of a data mining process so that they are easier to understand and use; Fewer feature so that in the next round of data collection, a saving can be made by removing redundant or irrelevant features. A decision tree is created in two phases; Tree building phase which is done in top-down manner, this phase can repeatedly partition the training data until all the examples in each partition belong to one class or the partition is sufficiently small. And Tree Pruning Phase This phase should remove dependency on statistical noise or variation that may be particular only to the training set.

Algorithm: Generate decision tree. Generate a decision tree from the training tuples of data partition D.

Input:

Data partition, D, which is a set of training tuples and their associated class labels; attribute list, the set of candidate attributes;

Attribute selection method, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a splitting attribute and, possibly, either a split point or splitting subset.

Output: A decision tree.

Method:

Create a node N;

if tuples in D are all of the same class, C then return N as a leaf node labeled with the class C;

if attribute list is empty then return N as a leaf node labeled with the majority class in D;

// majority voting

Apply Attribute selection method(D, attribute list) to find the "best" splitting criterion;

label node N with splitting criterion;

if splitting attribute is discrete-valued and multiway splits allowed then // not restricted to binary trees attribute list attribute list splitting attribute; // remove splitting attribute for each outcome j of splitting criterion // partition the tuples and grow sub trees for each partition

Let Dj be the set of data tuples in D satisfying outcome j; // a partition

if Dj is empty then attach a leaf labeled with the majority class in D to node N;

else attach the node returned by Generate decision tree(Dj, attribute list) to node N; end for

Return N;

II. Related Works

Several scholars have explored the application of data mining techniques for series of research analysis and to provide solutions to myriad of problems confronting human race. Chang [1] applied data-mining technique to prediction of enrolment behaviours of applicants at a large state university. In another application of data mining technique [2] used support vector machines and rule-based predictive models to build predictive models for new, continued and returned students, respectively first, and then aggregates their predictive results from which the model for the total headcount is generated. In another work, [3] adopted clustering technique for course scheduling and online course offerings. They used clustering analysis to develop course scheduling and online course offering. Eykamp [4] also used data mining to explore which students use advanced placement to reduce time to degree. Goyal and Vohra [5] proposed the use of data mining techniques to improve the

efficiency of higher education institution. They stated that, if data mining techniques such as clustering, decision tree and association are applied to higher education processes, it would help to improve students' performance, their life cycle management, selection of courses, to measure their retention rate and the grant fund management of an institution. This is an approach to examine the effect of using data mining techniques in higher education. Silwattananusarn and Tuamsuk [6] explored the applications of data mining techniques which have been developed to support knowledge management process. They reviewed journal articles indexed in ScienceDirect Database from 2007 to 2012 and analyzed and classified them, their discussion on the findings is divided into 4 topics- (i) knowledge resource; (ii) knowledge types and/or knowledge datasets; (iii) data mining tasks; and (iv) data mining techniques and applications used in knowledge management. The applications of data mining techniques in the process of knowledge management were summarized and discussed.

III. Experimental Design

We designed a decision tree based classification for scholarship disbursement, the dataset that was analyzed consist of about 800 training cases, with one (1) numeric attribute (Score) and nine (9) categorical attributes (Name of Applicants, Qualification, Grade, O'Level, Status, Application, State and Local Government (LG)) and the class (FStatus) to be predicted, as shown in Fig. 1. Each node on the tree corresponds to a test on an attribute, while the leaf node denotes the predicted class**.** We start the test from the root node, if the course is not Information Technology (IT) related then we reject the application and proceed to the right subtree and apply the test. On the second node, if the grade is First or Upper we move to the right subtree node else we reject the application. Again, we test for O'level, if this is complete we move to the right subtree or else we reject the application. We then test for the status of applicant, if the status is shortlisted for the aptitude test we proceed to the next right subtree (application) and test. If the application is MSc., the system check for the state, local government (LG), the LGScore, Final Status (FStatus), and decide whether the applicant is granted scholarship or not.

IV. Rule Based Classification for Scholarship

Given a decision tree T in Fig. 1, each of its leaves corresponds to a rule. The rule can be derived by conjoining the logical tests on each node on the path from the root of the tree to the corresponding leaf. The rule derived from the decision tree of Fig. 1 is highlighted in Fig. 2.

The architecture of the system comprises of software with six user interfaces namely: Input interface, General Applicant list, Shortlisted Candidates, Status form, Applicants Report form, Successful Candidates. Input interface form is used to enter applicants' data into the system. General Applicant Form is used to display the list of all applicants apply for the scholarship together with their data. Shortlisted candidates form is used to display the list of shortlisted candidates for the aptitude test. Report form displays the general report of applicants including the scores, reasons for short listing or rejection of application. The rules derived from the decision tree of Fig. 1 are shown in Fig. 2, while the algorithm is shown in Fig. 3. Status Form is used to check the status of individual applicant, all applicants, Report form, shortlisted and successful candidates. Successful Candidates displays the list of successful candidate for the scholarship scheme. A form is designed with graphical user interface software package which is used to enter input data into the system. The input interface ensures the reliability and validity of the input. The input menu collects data and saves into the database. The input data consists of user identity number, applicants' name, applicants' qualification, grade, Ordinary level, status, state, local government (LG), score, and application. The output interface consists of forms that display the results of processes to the system user. The output forms include shortlisted candidate form, report form, general applicant form, status form, and successful candidates form. The standard query language (SQL) which is at the back end of the system, the query statement is used to query the database for instance, to insert record, delete record, edit record and performs other operation on the database and front-end.

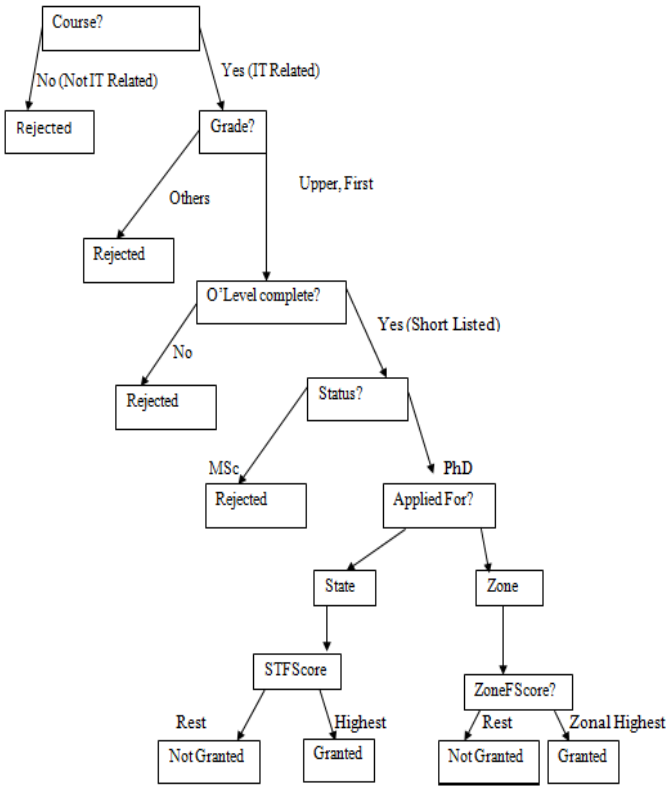

Fig. 1 Decision Tree-based Classification for Scholarship Problem

This is the rule derive from the tree-based decision classification.

R1: If (Course = Non-IT Relate (No)) = Rejected

R2: If (Grade = Others) = Rejected

R3: If (O'Level = Non-Complete) = Rejected

R4: If (Status = No) = Rejected

R5: If (Course =" IT Related (Yes)", Grade = "First Upper",
O'Level = "Yes", Status = "Shortlisted",
Application = "PhD" Group Zone, Check
Score = "Highest") = Granted

R6: If (Course =" IT Related (Yes)", Grade = "First, Upper",
O'Level = "Yes", Status = "Shortlisted",
Application = "MSc." Group State, Group ST,
Check Score = "First Highest") = Granted

Fig. 2 Rule derived from decision tree of Fig. 1

```
If Course = Non-Related Then Reject the Application
If Grade = Others Then Reject the Application
If O'Level = Non-Completed Then Reject the Application
If Status = No Then Reject the Application
  Else
  If (Course ="IT Related)" And (Grade = "First OR Upper")
   If (O'Level = "Yes") And (Status = "Shortlisted")
   If (Application ="PhD") And (ZoneScore = "Highest") Then
       Scholarship = Granted
    Else
      If (Course = "IT Related)" And (Grade = "First OR Upper")
       If (O'Level = "Yes") And (Status = "Shortlisted")
       If (Application = "MSc.") And (STFScore = "Highest") Then
          Scholarship = Granted
        Else
       Return
        Endif
       Endif
      Endif
     Endif
    Endif
  Endif
 Endif
```

Fig. 3 Algorithm for scholarship disbursement

V. Results and Discussion

This research used a decision tree based classification technique to develop an algorithm. Rules generated from the tree were used to develop software that is applicable in the disbursement of scholarship for postgraduate studies. The system was tested with NITDA data and found effective, efficient and with reduction in other challenges of existing system. Based on the technique of the decision tree based classification, this research was able to predict whether an applicant shall be successful or unsuccessful in his or her application using the applicant data. The tree algorithm used in designing this system is effective and efficient.

References

[1] L. Chang, "Applying data mining to predict college admissions yield: A case Study" New Directions for Institutional Research, 2006, pp.53–68. doi: 10.1002/ir.187.

[2] S. S. Aksenova, D. Zhang, and M. Lu, "Enrollment prediction through data Mining", in Information Reuse and Integration, 2006 IEEE International Conference. Retrieved on 01/13/2009. Available at http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4018543.

[3] J. Luan, and C. Zhao, "Practicing data mining for enrollment management and beyond", in J. Luan & C. Zhao (Eds.), New Direction for Institutional Research, 2006, no.131. San Francisco: Jossey-Bass. W.-K. Chen, *Linear Networks and Systems* (Book style). Belmont, CA: Wadsworth, 1993, pp. 123–135.

[4] P. Eykamp, "Using data mining to explore which students use placement to reduce time to degree", in J. Luan and C. Zhao (Eds.), New Direction for Institutional Research, 2006, no. 131. San Francisco: Jossey-Bass.

[5] M. Goyal and R. Vohra, "Applications of Data Mining in Higher Education", in International Journal of Computer Science Issues (IJCSI), Vol. 9, Issue 2, No 1, March 2012 ISSN (Online): 1694-0814 www.IJCSI.org

[6] T. Silwattananusarn, and K. Tuamsuk, "Data Mining and Its Applications for Knowledge Management : A Literature Review from 2007 to 2012,.International Journal of Data Mining and Knowledge Management Process (IJDKP) Vol.2, No.5, September 2012, Pp 13-24.

Dr. J. K. Alhassan was born at Ganmu-Alhaeri, in Kwara State, Nigeria and obtained Bachelor of Technology in Mathematics/Computer Science, at Federal University of Technology, Minna, Niger State, Nigeria in 2000. Then Master of Science in Computer Science, at University of Ibadan, Nigeria in 2006, and Doctor of Philosophy in Computer Science, at Federal University of Technology, Minna, Niger State, Nigeria in 2014. The major field of study is computer science.

He carried out part of his PhD research at United Institute of Informatics Problems, National Academy of Sciences of Belarus (UIIP NASB) Minsk, Republic of Belarus. He is currently the Ag. Head, at the Department of Cyber Security Science, Federal University of Technology, Minna, Niger State, Nigeria. He has published twelve journal articles and four conference proceedings. His research interest includes Artificial Intelligence, Data Mining, Internet Technology, Database Management System, Software Architecture, Machine Learning, Human Computer Interaction and Computer Security. Dr. Alhassan is a member of Computer Professionals Registration Council of Nigeria (CPN).

S. A. Lawal is a Civil Servant and work at National Information Technology Development Agency, (NITDA), Abuja, Nigeria.