Higher Education Students' Enrolment Forecasting System Using Data Mining Application in Ethiopia

Mahlet Mulugeta
HiLCoE, Computer Science Programme, Ethiopia
Ministry of Education, Center of Educational ICT
mahimulgeta@yahoo.com

Berhanu Borena
HiLCoE, Ethiopia
PhD Candidate, Addis Ababa University, Ethiopia
berhanuborena@gmail.com

Abstract

Predictive data modeling for enrolment planning is an innovative methodology that can be utilized by higher educations. Accurate and reliable student's enrolment prediction at the higher education is crucial in order to minimize inefficient utilization of resources and funds at the universities. The objective of this research is to develop predictive model, which determines the number of higher education students' enrolment at department level ahead of time using data mining approaches. Three different dataset are used for this purpose: one year, two year and three year projection. The data set included selected attribute of historical data from government and private sample universities such as Addis Ababa University, Mekele University, Bahirdar University, Jimma University, Harmaya University, Unity University and St. Mary University using purposive sampling technique. For each of the universities a maximum of five departments are taken as a sample. These are Medicine, Civil Engineering, Chemistry, Management, Law and Information Science. In order to develop a model for each of the experiments, the data mining algorithms: Decision tree (J48 Classifier), Bayesian Classifier (Naïve Bayes) and Neural Network (Multilayer Perceptron) are used. 10-fold cross validation is employed for all the selected attributes at each dataset. In this research, the higher education student enrolment at department level is predicted using the Weka software. The algorithms are finally evaluated and compared using model comparison technique such as confusion matrix, classification rate and Area under the ROC curve at each of the experiment in order to select the best algorithm for higher education student's enrolment forecasting at each of the departments. The Neural Network (multilayer perceptron) performed best in this research study since the classification rate of Multilayer Perception of one year, two year and three year of the experiments were 91.9%, 91.4% and 92.6% respectively. The ROC, TP Rate, F-measure and TN Rate of neural network model of each of the three experiments were also higher than decision tree and Naïve Bayes. Furthermore, a student enrolment prediction system prototype is developed for one year ahead prediction. Finally, this study found that institution can use historical data of high school and department enrolment trend to predict student's enrolment at each of the department.

Keywords: J48 Classifier; Multilayer Perceptron; Naïve Bayes; Predictive Data Mining

1. Introduction

There has been a steady increase in the number of undergraduate students' enrollment in higher education in Ethiopia in the last 10 years – rising from 34,589 students in 2000/01 to 494,110 in 2011/12 [1]. A growing majority of undergraduate students were enrolled full-time or regular. There were 23,320 full time students in 2000/01; by 2011/12, the undergraduate student body had more than doubled to 269,862 [2]. Increasing enrollment will require additional space, equipment, faculty and human resource especially professional teachers. There is also a need to monitor the possible quality tradeoffs associated with such expansion.

Nowadays, the challenges that higher education institutes are facing is to improve the quality of managerial decision, quality of education, efficient resource utilization, and lack of a sufficient number of

qualified teachers and so on [3]. According to Naeimeh & Mohammad the managerial decision making process became more complex as the complexity of educational entity increases [4]. Higher educational institutes seek more efficient technology to better manage resources and support decision making procedure. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students and resources required in a particular course, alienation of traditional classroom teaching model, detection of unfair means used in online examination, prediction about students' performance and so on [1]. Prediction is the most widely used technique in Data Mining and in recent research data mining in education is at its peak [2]. Data mining and forecasting abilities benefit organizations across a variety of fields and in many application areas of education [5]. Different data mining techniques perform better than other depending on the information that organizations hope to capture from the data as well as the types and quantities of the data itself.

The main contribution of this paper will be to develop a forecasting system for higher education student enrolment at each program level or department level using data mining technology in the context of Ethiopia using time series data. This means that the system will predict number of students at each program (Computer Science, Engineering, etc.) at each specific university within each specific program. The system helps transform disparate corporate educational data into information and knowledge, and helps managers/officers make better decisions in resource sharing. The purpose of projection modeling is to reduce complex institutional problems to simpler proportions, so that the human skills of decision makers can most effectively be brought to bear on the issues to be resolved [6]. This system is based on objective baseline data of National Educational Examination and Assessment Agency, MOE (Ministry of Education) and Higher Educations as they benefited from this application.

What initiated this research is minimizing human power, cost and time of both for university's department and MOE for enrolment forecasting at department level. And also to increase accuracy, reliability and consistency in their department enrolment projection system so that inefficient utilization of resources at the higher education sectors will be minimized such as professional teachers, classrooms, textbooks, well equipped laboratories, budget and so on at the sectors which have direct impact on the quality of education. According to Jing [7] data mining is a powerful analytical tool that enables educational institutions to better allocate resources and staff, proactively manage student outcomes, and improve the effectiveness of alumni development. In this research, the possible applications of data mining technique are used to forecast higher education student enrollment at each department based on the underlying assumption. Thus, the result of this data mining process could be used for top and middle level managers at the universities of each department as well as planning department at MOE for better decision making, better resource utilization and so on. To address these challenges the following research questions were addressed in this research:

- What are the appropriate variables to forecast higher education student enrollment at each department?
- What are the important rules (patterns) for prediction of higher education enrolment at different departments?
- What is the appropriate data mining technique to forecast higher education student enrollment at each department and to evaluate the forecasting model?

CRISP-DM (Cross Industry Standard Process for Data Mining) is used subsequently to conduct systematic data mining analysis. All of its stages are fully organized, structured and defined; allowing that a project to be understood or revised. It also clearly defines the business understanding perspective.

The methodological root of this research that are used in the six phases of CRISP – DM processes such as Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Prototype development. The three data sets such as for one year ahead, two year ahead and three year ahead enrolment predictions are first introduced, as well as the preliminary diagnoses done on each data set to gain an insight into their properties. The data is preprocessed to manage outliers, missing values, etc. and reduce level of dispersion between the variable in the data set. At this stage, the data sets are divided into two parts as the "training set" and "test set" using 10-fold Cross Validation. For each data set, a predictive data mining model is built using three data mining techniques: Decision tree, Neural Network and Bayesian Classifier. Four model adequacy criteria are used at this stage to measure the performance and adequacy of the prediction model.

2. Related Work

One of the purposes of the research conducted by Brown [8] was to test a new method to forecast University enrollments using Ratio Smoothing Model. Ratio and Ratio Trend Smoothing modeling process were used to project ratios for all grade classification for three semesters into the future. The forecasted ratios were translated into actual enrolments by multiplying an enrollment figure by the ratio to arrive at a new projected enrollment. The underlining assumption to develop the ratio smoothing model was based upon the causal factors, either unknown or not quantifiable, which influence each transition ratio. Such factors include the marketability of program skills, the student sprit of the times (liberal or conservative), administrative decisions to augment or curtail programs, rate of promotion of students and the retaining power of Universities. The most recent ratio is given a weight, the next most recent a lesser weight, and so forth into the past.

A research conducted in [9] presents a prediction model capable of forecasting undergraduate unit enrolments at the Faculty of Science, Macquarie University. The research investigated four different models with three different approaches to predict university enrolments. These include linear regression, logistic regression, and rule based prediction. The variables used in the paper are the population count of the city, the strength of the economy, the value of education and Demographic variables of students such as age, gender and students' financial situation. With all types of prediction models, there are three data sets that are used. The training data is used to calculate any correlations and build the model. The validation data is used to validate the model to ensure the correct variables and constants were used. Lastly, the test data is used to test the success rate of the model. The result shows that since most enrolment figures were on the decline for the past six years and other units showed inconsistent behavior, linear regression could not be applied. Instead, they looked at how unit enrolments changed from year to year. Therefore they predicted if a unit enrolment should rise or fall, and the magnitude, based on the change of its dependent unit.

The research conducted by Padmapriya [10] applied data mining techniques to predict higher education admissibility. In this research, real data of about 690 undergraduate students from public arts college were used. The research focused on the development of data mining models for predicting the students likely to go for higher studies, based on their personal (such as Annual Income, Parents educational qualification, Age, Number of Children in the Family, Native place), precollege (Place of School Education, Profile School Education and Score of School Education) and graduate performance (Branch Name, Percentage of marks and Interest for higher studies) characteristics. The dataset used for the research includes data about students admitted to the college in three consecutive years. Several well-known data mining classification algorithms, including a decision tree classifier and Naive Bayesian classifier, are applied on the dataset. The performance of these algorithms is analyzed and compared.

The research conducted by Numan *et al.* [11] in Bangladesh Open University (BOU) presents the analysis of students' enrollment trend at different programs of BOU as well as shows the projection of enrollment trend. The University has 20 formal programs. Students' enrollment pattern in different programs of BOU is analyzed using linear regression. The study showed that every year students' enrollment is rising for only four programs and declining for the rest of the programs. The study reveals that some programs already have touched the borderline of zero students' enrolment. At the same time, projection shows that if these trends continue at these rates, then students' enrollment of six of these programs would reach to nil in the next few years. So, it is time for the policy makers and the academics to rethink about those programs and take effective strategies to revive those programs. In the end, the authors also pointed out some measures to be considered in offering effective and more popular educational programs through BOU. A total of 781,904 students have been enrolled in different programs at BOU up to 2006. Among them 76.1% are males and 23.9% are females. Up to 2005, there were 131,068 students awarded certificate or degree from BOU. The year wise students' enrollment pattern in different programs of BOU has been analyzed using linear regression analysis.

According to Florida Postsecondary Education Planning Commission [12] the projection model relies upon a series of calculations, such as average annual increase, graduation and retention rates, econometric relationship of enrolment to college age population, high school graduate and returning adult. These methods fall in to five categories: Rule of Thumb, Average Annual Increase, Cohort Survival, and regression analysis. The simplest method of estimating future enrollment is the "Rule of Thumb". Basically increase in high school graduates have a direct effect on increase in higher education enrollment. In Average Annual Increase it is appropriate to use three years or five years or even twenty years using a rolling average, and each produced substantially different results and it does not take into account external factors, such as population, high school graduates, retention, etc. There is no rule to determine the correct period that should be used.

The third type of model which is Cohort Survival most closely mirrors present policies and is very amenable to modeling variations of policies. This technique requires a rather extensive database, including retention and graduation rates by institution, by level and cohort for at least a 10 year period. In regression method the regression equation is developed and used several different variables to predict enrollment. Enrolment planning is informed by a mix of internal and external considerations, including student demand for programs, institutional capacity, government funding and policy, as well as a determined size and mix of programs that are aligned with institutional priorities and goals [13].

As mentioned earlier this research focused on prediction of student enrolment at department level at higher education. The main difference or feature of this study is that its prediction is mainly focused at department level which is not done in any of the literatures before. The three commonly used data mining algorithms are selected and used in order to develop the prediction model.

3. The Proposed Solution

In order to minimize the gap of the existing system at the MOE and Higher Education, in this research an attempt was made to develop an operational application prototype named Higher Education Student Enrolment Prediction System (HESEPS) that uses three different data mining algorithms, namely, J48 Classifier, Naïve Bayes and Neural Network to develop the prediction model. This is done by converting the output of modeling to prototype using Java programming. The prototype predicts students' enrolment at the department level not only using the best algorithm, but using all of the three data mining algorithms, which means the experts can have different alternatives.

The prototype is developed using the output of the models generated from the three different data mining algorithms, namely, J48 Classifier, Naïve Bayes Classifier and Multilayer Perceptron. During experimentation and modeling the ARFF, CSV and MDL format of the dataset of each of the predictions were stored in a file for each of the modeling techniques. The built models of the three algorithms were manually saved in order to use it in the software implementation phase.

The programming language used for writing the code is Java, NetBeans. The reason for selecting Java is because it is object oriented programming language, which follows modern programming methodology. It also provides greater security, speed, reliability, improves development productivity, standardizes the platform, and ensures portability of developed applications and support [14].

In the implementation phase, the first user interface is built using two J-penal which is one for the text area and the other contains buttons. This interface involves three kinds of predictions such as one year, two year and three year ahead. The experts interact using this user interface and can select one of the prediction types (see Figure 1). The prediction was done by interacting with the stored files using appropriate function calls. The data of each of the factors are collected from the user interface provided by the expert.

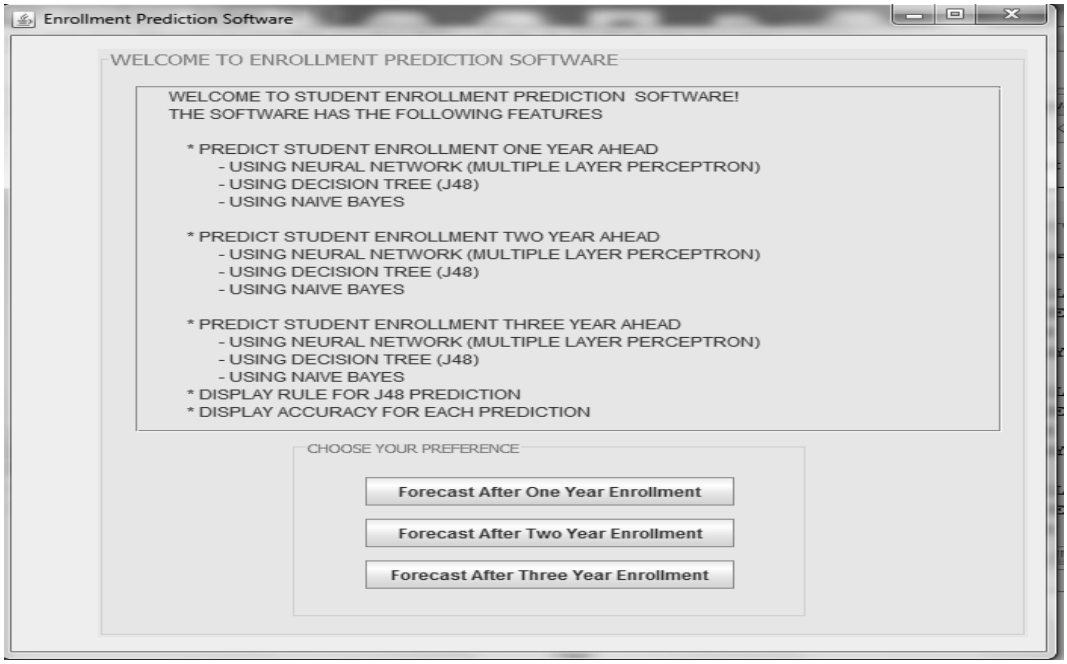

Figure 1: System Prototype User Interface

Figure 1 shows the main page of enrolment prediction software and users can easily select prediction type. The prototype is developed using Java whereas the model is developed using Weka as mentioned earlier. This makes the implementation phase easier and the stored file in the modeling phase can be easily interacted.

4. Result

For J48 classifier, Naïve Bayes and Multilayer Perception at each of the three experiments such as one year ahead, two year ahead and three year ahead separately, Tables 1, 2, and 3 shown the summary result of the three models.

Table 1: Summary Result from One Year Ahead Projection

Model	*Accuracy*	*TP Rate*	*F-Measure*	*ROC Area*	*Time (Sec)*
J48	85.42%	0.854	0.853	0.934	0.03
Naïve Bayes	72.22%	0.722	0.715	0.874	0.02
Multilayer Perceptron	91.9%	0.919	0.919	0.969	32.72

Table 2: Summary Result from Two Year Ahead Projection

Model	*Accuracy*	*TP Rate*	*F-Measure*	*ROC Area*	*Time (Second)*
J48	83.33%	0.083	0.832	0.951	0.03
Naïve Bayes	60.41%	0.604	0.597	0.851	0.03
Neural Network	91.44%	0.914	0.915	0.949	22.25

Table 3: Summary Result from Three Year Ahead Projection

Model	*Accuracy*	*TP Rate*	*F-Measure*	*ROC Area*	*Time (Second)*
J48	85.95%	0.86	0.858	0.949	0.02
Naïve Bayes	61.26%	0.613	0.55	0.815	0.02
Neural network	92.61%	0.93	0.92	0.96	77.25

In general, neural network and Multilayer perception outperformed the other algorithms by achieving the highest accuracy, TP Rate, TN Rate and ROC Area value but it took higher response time than the other two algorithms. The Naïve Bayes classifier has the lowest performance.

These experimental results have shown that neural network and multilayer perception outperformed Naïve Bayes and J48 classifiers in the domain of predicting higher education students' enrolment at the department level.

Specific rules are extracted from J48 Classifier from the three experiments. Even if the best algorithm is Neural network for predicting higher education enrolment at department level, the rules generated by J48 classifier that are useful for this specific domain are discussed here. Seven important rules are selected from the three experiments and domain experts.

The rules shown in Table 4 are good and were accepted since the number of high school students, repetition rate, dropout rate, previous enrolment trend and students' result affect enrolment at the department level.

Table 4: Some Selected J48 Classifier Rules

1	IF [(2YPRIV-ENRL<=46.5)&&(G12-RES-Aavg<=9459)&&(REP-RATE<=0.22)&&DOUT-RATE<=0.025)&&UNI-NAME=BAHIRDAR) &&(DEPT=IS)]	THEN (PROJ-ENRL=MODERATE (53, 106]) (27/7)
2	IF [(2YPRIV-ENRL<=91.5)&&(G12-RES-Aavg<=9459)&&(REP-RATE<=0.22)&&DOUT-RATE<=0.025)&&(UNI-NAME=JIMMA) &&(DEPT=MEDICINE)]	THEN (PROJ-ENRL=MODERATE (>= 106.5]) (9/1)
3	IF [(2YPRIV-ENRL=('46.5-91.5]&&(G12-RES-Aavg<=9459)&&(REP-RATE<=0.22)&&DOUT-RATE'(0.025-0.089]'&&(UNI-NAME=JIMMA) &&(DEPT=CIVIL ENG)]	THEN (PROJ-ENRL=HIGH(>=106.5)) (36/6)

The first rule selected from the rules generated by the J48 algorithm from the third experiment (three year ahead projection) gave a correct result for 27 of the 30 cases that it covers; thus its success rate is 90%. This rule is somehow a strong rule for predicting students' enrolment in IS department in the Bahir Dar University. The second rule from the third experiment gave also a correct result for 9 of the 11 cases that it covers; thus its success rate is 81.8%. The rule for predicting students' enrolment in the Department of Medicine of Jimma University is also promising.

In the second experiment (for two year ahead projection) three rules are selected from J48 algorithm. The first rule gave correct result for 36 of the 42 cases that it covers; thus its success rate is 85.7 for predicting students' enrolment in Department of Civil Engineering of Jimma University. The second rule also gave correct result for 44 of the 48 cases that it covers with a success rate of 91.7% for predicting students' enrolment in the Department of Management of Mekele University. The success rate of the third rule is 80% for predicting students' enrolment in the Department of Management of Bahir Dar University.

Two rules are selected from the rule generated by the J48 algorithm from the first experiment (for one year a head projection). The correct result in the first rule for predicting students' enrolment in the

Department of IS of Haramaya University gave a success rate of 97.61%. The second rule of the same experiment has a success rate of 81.25 % for predicting students' enrolment in the Department of Management of Bahir Dar University.

In general, based on the discussion made with domain experts, the rules generated by J48 classifier of the three predictions such as for one year ahead, for two year ahead and for three year ahead prediction are found to be strong for predicting higher education students' enrolment at the specified universities and departments.

5. Conclusion and Recommendation

5.1 Conclusion

Data mining, extracting meaningful patterns and rules from large quantity of data, is clearly useful in any field where there are large quantity of data and something worth learning. In this respect, the education sector is a potential area of data mining. In this research, an attempt was made to assess the potential applicability of data mining for higher education student enrolment forecasting at department level. This research, which employed the commonly used methodological approach in data mining research, made use of three predictive data mining modeling techniques, namely, Decision tree (J48 Classifier), Bayesian classifier (Naïve Bayes) and Neural network (Multilayer Perception), to address the problem.

The experimentation phase of this research was divided into three groups such as for one year ahead projection, for two year ahead projection and for three year ahead projection. Based on the objective stated earlier to address the problem that exists in the current higher education enrolment forecasting technique, appropriate models are selected, factors are selected and specific rules or pattern are extracted. The Neural network model using multilayer perceptron is selected as a working model among the three models specified for each of the three projections. The second algorithm which is selected as the next best model is decision tree, i.e., J48 classifier for one, two and the three years projection.

This research has several contributions. First, the planning and resource mobilization experts at the MOE and at higher education institutions can use and further enhance the use of the output of this research as the main support for making correct decision. Second, the current findings from this research will serve as a base for future studies in the education sector. Lastly, since no previous work was found that predicts higher education students' enrolment at department level analytically, the present study provides additional evidence to the research community and experts at the MOE and higher education institutions.

This experimental research shows that Neural Network is applicable for educational enrolment prediction. The students' department enrolment is not constantly changed as independent variables increasing or decreasing. In such situation, Neural Network is a great classifier since it is good to deal with non-linear relationship between predicator and target variable and for prediction in cases where the characteristics of the data are available explicitly and understood.

5.2 Recommendation

It is the strongest conviction of the authors that if higher size sample is considered and other factors are also included, the predication can be made more accurate by predicting the exact number for the years to come. Thus the authors suggest further research that considers such things.

The experiments conducted in this research were implemented almost with their default parameter of the algorithms and factors regarding universities' resources such as budget which are not included due to unavailability of data. So, further investigation should be performed with different parameter settings, by increasing the factors, departments, universities and by using additional algorithms to enhance and expand the capability and accuracy of the prediction model.

Higher education institutions can use data mining application by using time series data from high school or preparatory school information to build a model of higher education students' enrolment projection at department level for each specific projection mentioned earlier. In general, data mining enables educational institutions to better allocate resources and staff with the ability to uncover hidden patterns in large dataset by building a model that predicts student enrolment at higher education.

References

[1] Education Management Information Systems: Ministry of Education, "Education Statistics Annual Abstract", Addis Ababa, Ethiopia, August 2001.

[2] Education Management Information Systems: Ministry of Education, "Education Statistics Annual Abstract", Addis Ababa, Ethiopia, September 2012.

[3] Godswill Chukwugozie Nsofor, "A Comparative Analysis of Predictive Data Mining Techniques", Unpublished MSc Thesis, The University of Tennessee, Knoxville, August, 2006.

[4] Naeimeh Delavari and Mohammad Reza, "Data Mining Application in Higher Learning Institution", Informatics in Education, 2008, Vol. 7, No. 1, pp. 31–54, June 2007.

[5] Monica Goyal and Rajan Vohra, "Application of Data mining in Higher Education", International Journal of Computer Science Issues, Vol. 9, Issue 2, No 1, March 2012.

[6] Elayne Reiss, "Best Practice in Enrollment Modeling: Navigating Methodology and Processes", FACRAO Conference, University of Central Florida, Florida, June 5, 2012.

[7] Luan Jing, "Data Mining and Knowledge Management in Higher Education Potential Application" Annual Forum for the Association for Institutional Research, Canada, June 06, 2002.

[8] Marilou T. Healey and Daniel J. Brown, "Forecasting University Enrollments by Ratio Smoothing ", Higher Education, Vol. 7, pp. 417-429, Scientific Publishing Company, Amsterdam, 1978.

[9] George Gemayel, "Predicting Enrolments in University Units", ITEC No. 810 Final Report, Macquarie University, June 5, 2009.

[10] A. Padmapriya and Karaikudi, Tamil Nadu, "Prediction of Higher Education Admissibility using Classification Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering, Department of Computer Science & Engineering, Alagappa University, India, November 2012.

[11] Numan, A. Islam and A. Sadat, "Analytical Views of Students Enrolment Trend of different Programs of Bangladesh Open University and Projection", Turkish Online Journal of Distance Education, Vol. 8 No. 2 Article 4, Bangladesh, April, 2007.

[12] William Ho, Helen E. Higson, and Prasanta.Dey, "An Integrated Multiple Criteria Decision Making Approach for Resource Allocation in Higher Education", Operation and Information Management Group, Aston University, United Kingdom, June 2007.

[13] Brijesh Kumar Baradwaj, Rajasthan and Saurabh Pal, "Mining Educational Data to Analyze Student Performance", International Journal of Advanced Computer Science and Applications, Vol. 2, No. 6, 2011.

[14] Richard G. Badwin, "Java Programming Tutorial: A Fast-Moving Guide to Java Programming", retrieved from http://courses.coreservlets.com/course-materials/java.html, Last accessed on June 15, 2013.