Research Paper Computer Science

Analysis of Students Enrollment in Government And Private Schools in India Using Classification Mining

Samiksha H. Zaveri Research Scholar, Parul University, Vadodara, Gujarat,India

Divya R. Jariwala Research Scholar, Shree JJT University, Chudela-Dist., Jhunjhunu, Rajasthan, India

ABSTRACT

Educational organizations are one of the important parts of our society and playing a vital role for growth and development of any nation. Data Mining is an emerging technique with the help of this one can efficiently learn with historical data and uses that knowledge for predicting future behavior of concern areas. Based on the information analyzed, the data mining approaches regarding government and private schools are helpful for the development and the solutions of higher education in India. An emerging interdisciplinary research field known as educational data mining (EDM) is concerned with developing methods and applying data mining techniques for exploring the unique types of data that come from educational system. Its goal is to better understand the ratio of primary government and private schools to improve educational outcomes state wise. Growth of current education system is surely enhanced if data mining has been adopted as a futuristic strategic management tool. In this paper State wise primary government and private schools data has been taken and various classification approaches have been performed. In this research work Linear Regression is established as a best classifier with maximum accuracy and minimum root mean square error (RMSE). We present the results achieved with WEKA tool.

KEYWORDS : Education Data Mining ,WEKA, Classification : LinearRegression, KSTAR, LWL,S MO, ZeroR, REPTree

1 INTRODUCTION:

1.1 Introduction of Data Mining:

Data Mining is an effective tool to extract meaningful and interesting patterns from the current and historical data stored in data repositories which may be analyzed to predict future trends. Seeking knowledge from massive data is one of the most desired attributes of Data Mining. Data could be large in two senses: in terms of size & in terms of dimensionality. With the help of data mining tools and emerging research trends in this field, the data miner may extract knowledge from the large data marts very efficiently and quickly which may be used for the betterment of the organizations and the society (S. Lakshmi Prabha and Dr.A.R.Mohamed Shanavas,2015). Data mining is useful whenever a system is dealing with large data sets. In Data Mining classification, clustering and regression are the three key approaches into identified classes (A. Merceron and K. Yacef, 2005). Classification rules may be identified Classification is a supervised learning approach in which they are grouped from a part of data known as training data and further it may be tested for rest of the data (I. H. Witten and E. Frank, 2005).The effectiveness of classification approach may be evaluated in terms of reliability of the rule with test data set (Sonali Agarwal, G. N. Pandey, and M. D. Tiwari, 2012).

1.2 Role of Data Mining in Education:

Data mining had touched many fields including bioinformatics, ecommerce, fraud detection and now in the field of education as well. The data mining in the field of educational research is known as Educational Data Mining (EDM) (S. Lakshmi Prabha and Dr.A.R.Mohamed Shanavas, 2015).School education in India is a two-tier system; the first tier has ten years of study covering basic education followed by the second tier which has two years of secondary education. The basic education acts as a bridge to the secondary education. To improve the quality of education in India, data mining techniques can be utilized to improve the traditional process .Data mining consists of a set of techniques that can be used to extract relevant and interesting knowledge from huge amount of data. It is a technique that can be used to analyze dataset. Data mining technique fall into three methods; which are association rule mining, classification and prediction, and clustering (Han J., Kamber M., and Pie J., 2011 and Qasem A. Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-Shawakfa, 2011).

Numerous studies were performed on the evaluation of the educational systems using educational data mining. Data mining techniques are very much useful in the development and finding out the solutions of basic education in India. Data mining aided to identify some of the most important factors in evaluation of the educational system (Manisha Sahane, Sanjay Sirsat, Razaullah Khan and Balaji Aglave, 2014 and Haisan, A.-A., and Bresfelean, V. 2013). Education has always been important but perhaps never more so in man's history than today, Research suggests that primary education is very important for the development of young children before they enter formal school (Kaul, 2002). Primary education is considered to be very important for the child as it is the first step towards entering the world of knowledge as well as a healthy and purposeful life. Primary education helps children become more independent and confident as well as promoting the all-round development of the children (Ramachandran et al.,2003). Availability of quality primary education will promote inclusive education and meaningful access to school education by increasing enrolment and reducing the vulnerability of children (Manmohan Singh and Dr. Anjali Sant, 2014).

2. Objectives

The main objectives of this study are:

- To examine State wise primary government and private schools in India for the year
- 2010-11.
- To examine findings for critical factors upon which to provide recommendations for Intervention and further investigations.
- To mine the clean dataset using data mining techniques.
- To Utilizes Classification algorithm from the training data set of primary state wise schools and then test the result.
- To find out the percentage of children who are enrolled in the primary government or private school of India.

3. Related Work

A review suggests that majority of the work is based on data classification by using only a specific approach or classifiers. The aim should be not only to find out a solution for the problem specified but there should be some work for identifying the best approach. Here in the present research work few classifiers are taken together and applied to the dataset in order to select best classifier for the indentified problem.

3.1 WEKA as Tool

Researchers select WEKA (Waikato Environment for Knowledge Analysis) software that was developed at the University of Waikato in New Zealand. WEKA is open source software issued under the GNU General Public License. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. It is portable & platform independent because it is fully implemented in the

Java programming language and thus runs on almost any modern computing platform and is now used in many different application areas, in particular for education & research (S. Lakshmi Prabha and Dr.A.R.Mohamed Shanavas, 2015). WEKA can efficiently work with limited data. It takes data from excel file in Comma Separated Values (CSV) format, which is a very common application software to be used in each school for initial collection of data.

3.2 Data set Used

The data has been accessed by the site of DISE and NUEPA where the data is collected on annual basis. The schools have to submit the data on a yearly basis and the data is processed by the editing and processing department of the data banks. The researchers used 637 instances and 4 attributes in WEKA as Data Mining Tool.

3.3 Preprocessing Tool

Data processing is required to make dataset appropriate for various classification algorithms. Here numerical data sets are converted as nominal datasets. It is essential to have a suitable Data Mining tool for the purpose of carrying out Data Mining analysis of the available data

4. Proposed Methodology

Classification techniques are supervised learning techniques that classify data item into predefined class label. It is one of the most useful techniques in data mining to build classification models from an input data set. The used classification techniques commonly build models that are used to predict future data trends. There are several algorithms for data classification; one of them is the decision tree classification technique. Generally, this paper is a preliminary attempt to use data mining concepts; particularly classification, to help in supporting the quality of the educational system by evaluating state wise primary government and private schools data to study the main attributes that may affect the classification in basic education (Qasem A. Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-Shawakfa, 2011).

Classification analysis are applied for real data and results show even if there is lack of attributes, one may still apply certain data mining algorithms over school data to gain knowledge on the mainstream flow (Ahmedi, L., Bytyci, E., Rexha, B., and Raca, V., 2012). Suitable prediction techniques using data mining tool WEKA to help in enhancing the quality of the educational system. (Manisha Sahane, Sanjay Sirsat, Razaullah Khan and Balaji Aglave, 2014). This study investigates the accuracy of some classification techniques for analyzing performance of a student. It is indicated that the Data mining techniques are widely used in primary educational system in order to increase the effectiveness and efficiency of the traditional method and as a guideline to improve their decision making processes (Manmohan Singh and Anjali Sant, (2014).

Linear regression	The simplest form of regression is simple linear regression that just contains one predictor and a prediction. The relationship between the two can be mapped on a two dimensional space and the records plotted for the prediction values along the Y axis and the predictor values along the X axis. The simple linear regression model then could be viewed as the line that minimized the error rate between the actual prediction value and the point on the line (the prediction from the model) (http://www.thearling.com/text/dmtechniques/dmtechniques.htm).
SMOreg	SMOreg implements the support vector machine for regression. The parameters can be learned using various algorithms. The algorithm is selected by setting the RegOptimizer(https://www.knime.org/files/nodedetails/weka_classifiers_functions_SMOreg.html).
Kstar	K* algorithm as an instance based learner which uses entropy as a distance measure. The benefits are that it provides a consistent approach to handling of real valued attributes, symbolic attributes and missing values. K* is a simple, instance based classifier (S. Vijayarani and M. Muthulakshmi, 2013).
LWL	Locally weighted learning. Uses an instance-based algorithm to assign instance weights which are then used by a specified Weighted Instances Handler (https://www.knime.org/files/nodedetails/weka_classifiers_lazy_LWL.html.).
ZeroR	ZeroRis the simplest classification method which depends on the target and ignores all predictors. ZeroR classifier basically predicts the majority category (class).
REPTree	RepTreeuses the regression tree logic and creates multiple trees in different iterations. After that it selects best one from all generated trees. That will be considered as the representative. REPTree is a fast decision tree learner. Builds a decision/regression tree using entropy as impurity measure and prunes it using reduced-error pruning (Sushilkumar Kalmegh, 2015).

Table 1. Representing algorithm definition used in EDM

5. Result and Analysis

The research work has chosen 7 different classifiers for comparative analysis of performance of classifiers. As given in Table 2. The researchers applied different classification algorithm namely LinearRegression, SMOreg, KStar, LWL, ZeroR, REPTree. Among all this algorithm LinearRegression provide best result compared to other algorithm in which Correlation coefficient is 1, Mean absolute error is 0, Root mean squared error is 0, Relative absolute error is 0% and Root relative squared error is also 0%, which is helpful for establishing the relationship between different attributes.

Used Algorithm	Time taken to build model	Time taken to test model on training data	Correlation coefficient	Mean absolute error	Root mean squared error	Relative absolute error	Root relative squared error
LinearRegression	5.85 seconds	0.09 seconds	1	0	0	0%	0%
SMOreg	3.1 seconds	0.04 seconds	1	1.4697	1.5979	1.0462 %	0.8145 %
KStar	0 seconds	0.83 seconds	0.9982	9.4436	17.4583	6.7224 %	8.8994 %
LWL	0 seconds	0.63 seconds	0.6735	105.1534	147.4674	74.8529%	75.1721 %
ZeroR	0 seconds	0.01 seconds	0	140.48	196.1731	100%	100%
REPTree	0 seconds	0.01 seconds	0	140.48	196.1731	100 %	100 %

Table 2.Classification Techniques applied on Primary Government and Private School Data

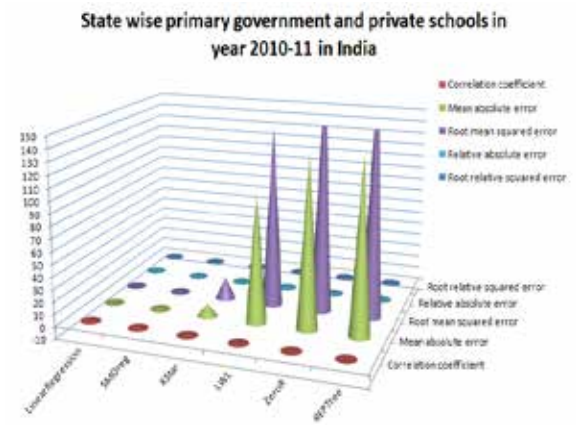


Figure 1. Chart on Classification Techniques applied on dataset

5.1 Histograms

One of the best ways to summarize data is to provide a histogram of the data. Researchers can create a histogram of primary government and primary private by analysis of database. Researchers identify histogram for three different categories state-wise, primary government and primary private schools database. However, for a database of many more records this is a very useful way of getting a high level understanding of the database.

Figure 2. State Wise Histogram

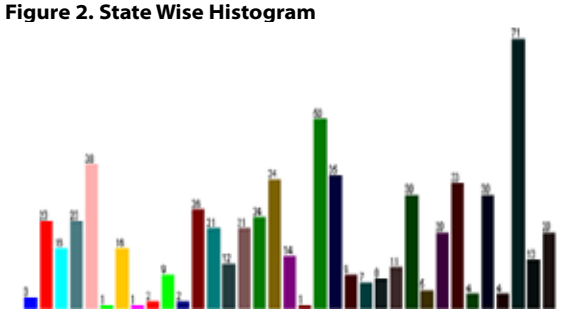


Figure 3. Primary Government Histogram

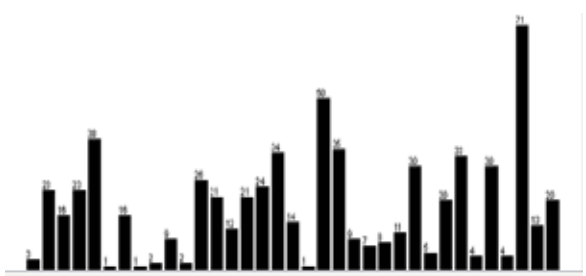
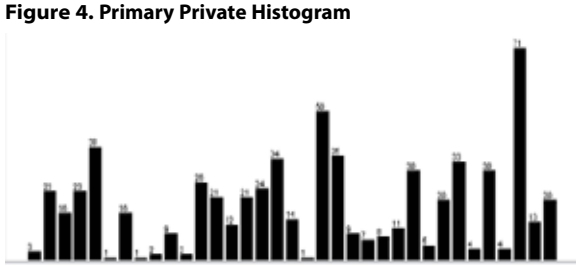


Figure 4. Primary Private Histogram



6. CONCLUSION

Researcher studied and proposed how EDM methods can be applied to educational data, particularly to schools performance. Data Mining could be used to improve business intelligence process including education system to enhance the efficacy and overall efficiency by optimally utilizing the resources available. With specific to education sector it has great demand both teaching and learning prospects. Among this entire algorithm LinearRegression provide best result compared to other algorithm. The management aspects are highly interference by the Information Communication Technology and DM areas. We applied data mining techniques to discover knowledge. This model aims to improve the quality of education where the model is intended to help student to choose the suitable track of their study, by analyzing the experience of previous students with similar academic achievements. Other attributes could also be added to the data set for further enhancing the generated model. Furthermore, some other classification models could be tested in this domain. Other classification models could be tested in this domain.

References

1. A. Merceron and K. Yacef, "Educational Data Mining: a Case Study (2005)," In C. Looi; G. McCalla; B. Bredeweg; J. Breuker, editor, Proceedings of the 12th international Conference on Artificial Intelligence in Education AIED, 467–474. Amsterdam, IOS Press.
2. I. H. Witten and E. Frank (2005), "Data mining: Practical Machine Learning Tools and Techniques," 2nd ed, Morgan-Kaufman Series of Data Management Systems San Francisco Elsevier.
3. Ahmedi, L., Bytyci, E., Rexha, B., and Raca, V. (2012). Applying data mining to compare predicted and real success of secondary school students. Advances in Applied Information Science, 178-181.
4. Han J., Kamber M., and Pie J. (2011), Data Mining Concepts and Techniques. 3rd edition, Morgan Kaufmann Publishers.
5. Ahmedi, L., Bytyci, E., Rexha, B., and Raca, V. (2012). Applying data mining to compare predicted and real success of secondary school students. Advances in Applied Information Science, 178-181.
6. S. Lakshmi Prabha and Dr.A.R.Mohamed Shanavas (2015) "Application of Educational Data mining techniques in e-Learning- A Case Study" , International Journal of Computer Science and Information Technologies, Vol. 6, No. 5, pp.- 4440-4443.
7. Sonali Agarwal, G. N. Pandey, and M. D. Tiwari (2012) "Data Mining in Education: Data Classification and Decision Tree Approach", International Journal of e-Education, e-Business, e-Management and e-Learning, Vol. 2, No. 2, pp. 140-144.
8. Qasem A. Al-Radaideh, Ahmad Al Ananbeh, and Emad M. Al-Shawakfa (2011), "A CLASSIFICATION MODEL FOR PREDICTING THE SUITABLE STUDY TRACK FOR SCHOOL STUDENTS", IJRRAS Vol. 2, No. 2, pp. 247-252.
9. Manisha Sahane, Sanjay Sirsat, Razaullah Khan and Balaji Aglave (2014), "Prediction of Primary Pupil Enrollment in Government School Using Data Mining Forecasting Technique", International Journal of Advanced Research in Computer Science and Software Engineering Vol. 4, No. 9, pp. 656-661.
10. Haisan, A.-A., and Bresfelean, V. (2013). A Data Mining Survey on Identifying the Factors that Influence Teachers' View of the Romanian Educational System. Advances in Applied Information Science, 7(3), 160-165.
11. "WEKA Data Mining Book" (n.d.) http://www.cs.waikato.ac.nz/~ml/weka/book.html.
12. "WEKA 3: Data Mining Software in Java" (n.d.) Retrieved March 2010 from http://www.cs.waikato.ac.nz/ml/weka/.
13. Manmohan Singh and Dr. Anjali Sant, (2014), "Performance Analysis of Primary School Students Using Data Mining Techniques", International Journal of Research in Advent Technology, Vol.2, No., pp. 100-105.
14. http://www.thearling.com/text/dmtechniques/dmtechniques.htm.
15. https://www.knime.org/files/nodedetails/weka_classifiers_functions_SMOreg.html
16. Ms S. Vijayarani and Ms M. Muthulakshmi (2013), "Comparative Analysis of Bayes and Lazy Classification Algorithms", International Journal of Advanced Research in Computer and Communication Engineerin, Vol. 2, Issue 8, pp. 3118-3124.
17. https://www.knime.org/files/nodedetails/weka_classifiers_lazy_LWL.html.
18. Sushilkumar Kalmegh (2015), "Analysis of WEKA Data Mining Algorithm REPTree, Simple Cart and RandomTree for Classification of Indian News", International Journal of Innovative Science, Engineering & Technology, Vol. 2 Issue 2, pp. 438-446.
19. Osman N. Darcan and Bertan Y. Badur (2012), "Student Profiling on Academic Performance Using Cluster Analysis" IBIMA Publishing Journal of e-Learning & Higher Education Article ID 622480, 8 pages DOI: 10.5171/2012.622480.