Impact of Data Mining Technique in Education Institutions

Mr. Bhushan Bandre, Ms. Rashmi Khalatkar
Research Scholar, Department of Computer Science and Engineering
D.M.I.E.T.R. Wardha Maharashtra India
Bhushanbandre47@gmail.com

Article History

Article Submission
10 February 2015
Revised Submission
27 April 2015
Article Accepted
10 May 2015
Article Published
30 June 2015

Abstract

Major decision making process using large amount of data can be done by various techniques using data mining. In education sectors various data mining techniques are implemented to analyze the student's data from the admission process itself. Due to large number of educational institution in India, excellence becomes a major parameter for the institutions to grow and with stand. Nowadays education institutions use data mining techniques to show their excellence. The main objective of this work to present an analysis of individual semester wise results of engineering college students using different techniques of data mining. Here we used different classification algorithms like decision tree, rule based, function based and Bayesian algorithms to analyze the semester results and comparison is made by considering parameters like accuracy and error rate. Our output shows the most suited algorithm for analyzing data in educational institutions.
Keywords: *Data mining, Error rate, Accuracy, Optimization*

I. Introduction

Secondary and higher education in India has gained significance growth for last few years. Due to more number of institutions and different courses the competition has increased among them in various aspects such as infrastructure, faculty, admissions, results and placements etc. The education regulatory committee stream lined some guidelines with respect to infrastructure, faculty and other required resources. Many new technologies were emerged in order to analyze and work with data, because data management is the key element for the success in any organization including educational organization. By proper data management the results of institution can be improved rigorously .A encouraging field to this objective is the usage of Data Mining technique [1]. Data mining is already well emerged in wholesale and retail business applications, but its entry in secondary and higher education is relatively less. In education sector, data mining is going to prove its best for the solutions in various parameters related to institutions growth [2].

The whole objective is solving all problems in education field and to improve process using various statistical techniques and data mining algorithms [3]. Educational data Mining an excellent practice used for prediction of student performance, analytics ,admissions, student modeling and grouping of students etc.

Educational Data Mining is targeted on various developing methods elaborate the uniqueness of large dataset from educational institutions. By using data mining it provides better environment for the students for their excellence in educating themselves in various things like studies, self-development and job opportunities. Data Mining is the process to converting data from education institutions to required information which can be used by faculty, administration staff, parents and researchers [4]. In education system, the prediction and explanation of performance day by day and monitoring the improvement is very important [5].

ISSN: 2250-0839

II. Related Work

Recent taxonomy encompasses all four types and focus on various applications of educational data mining to internet and web data [6]. The starting point of any study is Statistics and visualization which cannot be formally corrected. The following two categories are normal things which is found to be in data mining projects while the last category is an extension of data text data which is related to web [7].

The data mining in education institutions [8] has four strategies first is to improve students' model second is identify models for the knowledge structure third is pedagogical support by software and scientific methods about learning and learners. Predictive modeling techniques are used to understand the correlations in existing data. Results of above method will help for students associated with new registration and long term education process of students [9]. The well-known techniques such as clustering, decision tree and association is used to improve student's performance in secondary and higher education process. The overall outcome of this process can be explicitly used by faculty, curriculum designer who decides the courses for the students. This method used by educational course planner to have more relevant advance strategies on emerging technologies [10].

Data mining allows the machine learning algorithms for the automated discovery of new patterns and trends in the obtained data. Hence survey concludes data mining in education helps in identify students, behavior, and understanding learning outcomes of the engineering students.

III. Proposed Data path Designs

Initially before starting any process in data mining techniques, reliable data should be available. In order to collect data we are using spread sheet in Google. Even though lot methods are implemented to collect data, but the easiest and cost efficient method is Google spread sheet. An expert team is involved to prepare the required questionnaires to easily collect the data from the students. All the questionnaires are sent to the engineering students from a rural based engineering college in Tamil Nadu. The basic framed questions will cover data such as secondary percentage, semester wise percentage, training undergone, online course attended, skill development modules attended etc. The detailed questions will acquire data such as demographic, skill set and learning behavior.

In our analysis we used SPSS and WEKA online tools. WEKA is an education data mining analysis tool and is used to analyze different classification algorithms and it can also compare the performance of these algorithms based on various parameters. SPSS is well known statistical tool to identify the parameters of student's performance enhancement. Here different classification algorithm is implemented in student's data base by using various parameters of algorithms for efficient analysis. To improve the accuracy data is tested in 20 fold cross validation. The various algorithms which were implemented on our WEKA tools are NBDT, Decision Tree (J48), J Rip, One R ,Random Forest, Random Tree, REP Tree, , Simple Logistic and Zero R.

ISSN: 2250-0839

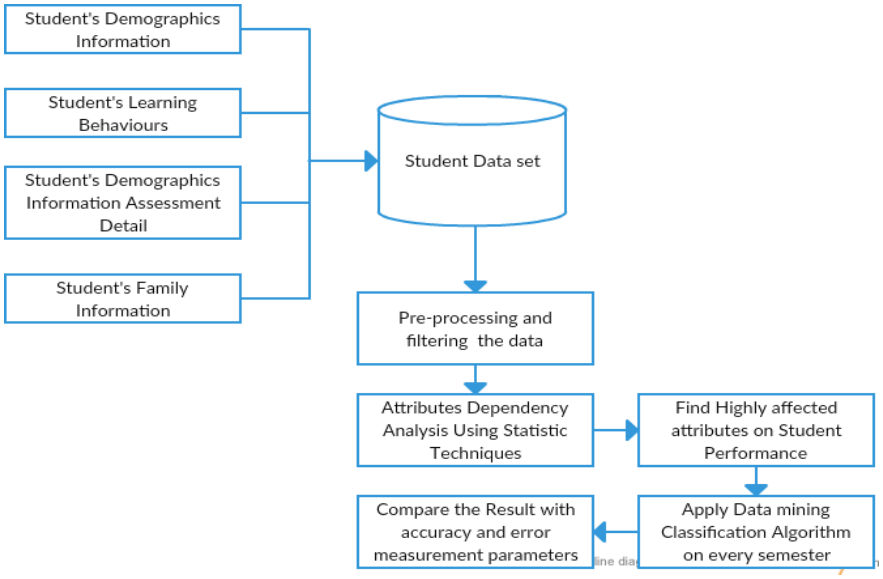

Fig 1: Proposed Research model

IV. Implementation and Analysis

The student parameters are analyzed based on the following stages:

Stage1: Collecting the information of the students.

Stage 2: preprocessing and optimization of the data.

Stage 3: Statistical techniques are applied to find students affected parameters.

Stage 4: Data mining techniques like Classification, Clustering, and Association is applied on student's data set.

Stage 5: Finding out the optimized data set.

TABLE 1: Student Parameters

PARAMETERS	DATA	VALUES WITHOUT ERROR
General	Nominal	Male or female
% in HSC	Nominal	Poor, Good, Excellent
Overall attendance	Nominal	Poor, Good, Excellent
Hour attendance	Nominal	Poor, Good, Excellent
Day attendance	Nominal	Poor, Good, Excellent
Week attendance	Nominal	Poor, Good, Excellent
Month attendance	Nominal	Poor, Good, Excellent
College Internal marks	Nominal	Excellent ,Good, Average ,Fail
College Assignment marks	Nominal	Excellent ,Good, Average, Fail
Extracurricular activities	Nominal	Nil, Average, Good, Excellent
Practical Marks	Nominal	Excellent, Average, Good
Theoretical marks	Nominal	Fail, Pass
Project Marks	Nominal	Fail, Pass

ISSN: 2250-0839

Previous semester %	Nominal	All clear, Backlogs
Name of the subject	Nominal	Compiler Design
Affiliated Theory Marks	Nominal	Pass, Fail
Affiliated Practical Marks	Nominal	Pass, Fail
Attendance	Nominal	Poor, Average, Good
Faculty Performance %	Nominal	Poor, Average, Good
Result %	Nominal	Poor, Good, Excellent
Semester result %	Nominal	Poor, Average, Good

Th multiple regressions were run to find seventh semester results from independent variables. These variables statistically predicted results of seventh semester, $F(51, 5217) = 736.946$, $p < .0002$, In order to retain all variables level is < 0.0003 and discard data which is > 0.0002 from the model.

TABLE 2: SPSS analysis of student's performance

.

% in HSC	Assignment Marks
% in first year	Subject Practical Knowledge
% in second year	Subject Theory Marks
Attendance	Project
Internal Marks	Previous Semester Marks
Extra Curriculum	Conferences

Here we used various classification algorithms like J48, Bayes Net, Multi-layer perception, Naïve Bayes, One R, Rep Tree, Decision stump, Logistic Regression, and sequential minimal optimization. Then results are compared with algorithms using WEKA tool. The comprative result is postulated.

TABLE 3: Semester wise data to build the model by various classifiers

Semesters	**J48**	**LS**	**MLP**	**BN**	**DS**	**ONE R**	**NB**	**SMO**	**LR**
I	0.12	0.78	0.9	1.21	2.13	0.012	0.126	0.156	0.245
II	0.0402	0.102	0.8202	2.3302	55.4502	0.023	0.154	0.125	0.235
III	0.0525	0.125	0.8325	2.1425	55.4625	0.032	0.185	0.124	0.214
IV	0.043	0.119	0.8299	2.3399	55.4599	0.12	0.121	0.145	0.265
V	0.042	0.103	0.8273	2.0373	55.4573	0.1235	0.112	0.165	0.222
VI	0.02	0.12	0.9	2.41	56.43	0.225	0.1325	0.144	0.213
VII	0.023	0.01	0.08	2.04	55.23	0.225	0.164	0.178	0.214
VIII	0.01	0.03	0.02	2.10	55.21	0.54	0.125	0.111	0.298
Mean Value	0.0326	0.1234	0.8183	2.3565	55.881	0.129	0.156	0.173	0.242

ISSN: 2250-0839

TABLE 4: Semester wise corrected classified instance by various classifiers

Semesters	J48	LS	MLP	BN	DS	ONE R	NB	SMO	LR
I	89.42	81.27	50.64	87.89	87.07	87.7	68.9	87.1	77.81
II	89.12	86.45	59.64	87.41	82.26	87.4	72.9	85.8	82.7
III	89.29	87.65	61.64	87.41	83.76	88.4	74.9	86.8	83.7
IV	89.95	87.65	61.64	88.01	83.16	88.4	74.9	87.8	84.7
V	89.05	87.65	62.04	88.41	84.66	88.4	74.5	87.8	84.7
VI	89.17	88.27	49.94	8.49	79.27	88.7	69.9	88.1	79.1
VII	89.01	88.00	45.02	97.98	88.23	88.00	85.20	88.32	95.02
VIII	89.02	87.01	59.32	97.56	89.32	98.12	84.02	86.25	96.25
Mean Value	89.24	86.95	58.47	87.42	81.54	88.15	72.30	86.12	81.60

TABLE 5: Semester wise kappa statistical rate

Semesters	J48	LS	MLP	BN	DS	ONE R	NB	SMO	LR
I	0.042	0.103	0.8273	2.0373	55.45	0.123	0.112	0.165	0.222
II	0.02	0.12	0.9	2.41	56.43	0.225	0.1325	0.144	0.213
III	0.023	0.01	0.08	2.04	55.23	0.225	0.164	0.178	0.214
IV	0.01	0.03	0.02	2.10	55.21	0.54	0.125	0.111	0.298
V	0.02	0.78	0.9	1.41	2.43	0.012	0.126	0.156	0.245
VI	0.0402	0.102	0.8202	2.3302	55.45	0.023	0.154	0.125	0.235
VII	0.0525	0.125	0.8325	2.1425	55.46	0.032	0.185	0.124	0.214
VIII	0.043	0.119	0.8299	2.3399	55.45	0.12	0.121	0.145	0.265
Mean Value	0.0126	0.2134	0.7983	2.2465	55.58	0.139	0.166	0.153	0.212

TABLE 6: Semester wise Mean Absolute Error classifiers

Semesters	J48	LS	MLP	BN	DS	ONE R	NB	SMO	LR
I	0.0634	0.060	0.067	0.023	0.154	0.125	0.103	0.8273	2.0373
II	0.0934	0.070	0.097	0.032	0.185	0.124	0.12	0.9	2.41
III	0.1034	0.080	0.107	0.12	0.121	0.145	0.01	0.08	2.04
IV	0.1134	0.090	0.117	0.1235	0.112	0.165	0.03	0.02	2.10
V	0.1234	0.100	0.127	0.225	0.1325	0.144	0.78	0.9	1.41
VI	0.0734	0.060	0.077	0.225	0.164	0.178	0.102	0.8202	2.3302
VII	0.098	0.07	0.054	0.54	0.125	0.111	0.125	0.8325	2.1425
VIII	0.09	0.07	0.043	0.129	0.156	0.173	0.119	0.8299	2.3399
Mean Value	0.0950	0.077	0.098	0.023	0.154	0.125	0.2134	0.7983	2.2465

TABLE 7: Semester Root Mean Squared Error Rate classifiers

Semesters	J48	LS	MLP	BN	DS	ONE R	NB	SMO	LR
I	0.126	0.156	0.245	0.042	0.103	0.2	0.060	0.067	0.125
II	0.154	0.125	0.235	0.02	0.12	0.22	0.070	0.097	0.154
III	0.185	0.124	0.214	0.023	0.01	0.21	0.080	0.107	0.165
IV	0.121	0.145	0.265	0.01	0.03	0.23	0.090	0.117	0.134
V	0.112	0.165	0.222	0.02	0.78	0.28	0.100	0.127	0.125
VI	0.1325	0.144	0.213	0.0402	0.102	0.22	0.060	0.077	0.155
VII	0.164	0.178	0.214	0.0525	0.125	0.21	0.07	0.054	0.141
VIII	0.125	0.111	0.298	0.043	0.119	0.27	0.07	0.043	0.161
Mean Value	0.156	0.173	0.242	0.0126	0.2134	0.24	0.077	0.098	0.151

V. CONCLUSION

In this work various algorithms were applied into the WEKA tool like J48, BN, NB, One R, RT, DS, LR, MLP and SMO optimization for getting the semester wise performance of the students. The algorithms concentrated on the accuracy and error measured parameters. In our analysis, we considered accuracy parameters like time taken to build the model, correctly classified instances and used error measurement parameters like Root mean square error, kappa statistics and mean absolute error. The analysis results in the best algorithm based on accuracy, lowest error rate and time required to build a model. From the above tables it is proved that J48 algorithm is best among the considered algorithms.

References

[1] Rakesh Mora. A Survey from on data mining in education, JATIT, Vol. 32, 2010, 125-131.

[2] Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. In: Paper presented at International conference on knowledge discovery and data mining. 1996

[3] Han J, Kamber M. Data mining: concepts and techniques. Massachusetts: Morgan Kaufmann Publishers; 2001.

[4] Ravi Shankar. Institutional Research Using Data Mining, IJACSE, Vol. 132, 2015, 69-81.

[5] Ranny Tigery. Education emphasis in data mining, Journal of Neural Networks,2015,Vol.17,315-323

[6] Berkhin P. A survey of clustering data mining technique. In: Kogan J, Nicholas C, Teboulle M, editors. Grouping multidimensional data. Berlin: Springer; 2006. p. 25–72.

[7] DeLong, C., P. Radclie, L. Gorny. Recruiting for Retention: Using Data Mining and Machine Learning to Leverage the Admissions Process for Improved Freshman Retention. – In: Proc. of the Nat. Symposium on Student Retention, 2007.

[8] B.M. Patil, Hybrid prediction model. Expert Systems with Applications 37 (2010) 8102-8108.

[9] M. A. Khan, W. Gharibi and S. K. Pradhan, "Data mining techniques for business intelligence in educational system: A case mining," 2014 World Congress on Computer Applications and Information Systems (WCCAIS), Hammamet, 2014, pp. 1-5, doi: 10.1109/WCCAIS.2014.6916559.

[10] P. Guleria, M. Arora and M. Sood, "Increasing quality of education using educational data mining," 2013 2nd International Conference on Information Management in the Knowledge Economy, Chandigarh, 2013, pp. 118-122.

[11] T. N. Manjunath, Ravindra S. Hegadi, I. M. Umesh and G. K. Ravikumar, "Realistic Analysis of Data Warehousing and Data Mining Application in Education Domain", *International Journal of Machine Learning and Computing*, vol. 2, no. 4, August 2012.

[12] Murtadha M. Hamad and Shumos T. Hammadi, "Quality Assurance Evaluation For Higher Education Institutions Using Statistical Models", *International Journal of Database Management Systems (IJDMS)*, vol. 3, no. 3, August 2011.

[13] G. Satyanarayana Reddy et al., "Data Warehousing Data Mining OLAP And OLTP Technologies Are Essential Elements To Support Decision-Making Process In Industries", *(IJCSE) International Journal on Computer Science and Engineering*, vol. 02, no. 09, pp. 2865-2873, 2010.

ISSN: 2250-0839