www.ijecs.in
International Journal Of Engineering And Computer Science ISSN:2319-7242
Volume 6 Issue 4 April 2017, Page No. 21034-21041
Index Copernicus value (2015): 58.10 DOI: 10.18535/ijecs/v6i4.41

Survey on Educational Data Mining Techniques

[1,2]P V V Satya Eswara Rao,[2,3]S K Sankar
[2]Assistant Professor, Department of Information Technology, Sasi Institute of Technology and Engineering
[3]Assistant Professor, Department of Computer Science and Engineering , Sasi Institute of Technology and Engineering

Abstract
In the recent years, data mining is the most important domain in the real world aspects. By using data mining Techniques, we can identify the knowledge of different areas and get the best patterns. Educational institutions and universities are facing problems in terms of student employability. It became a big task for the educational institutions. In this regard, we identified the techniques to bring co-relation between the student academics and faculty responsibilities, where co-relation pattern means reading the knowledge from the educational student performance data. Using this data, we are applying different Data mining Techniques to find out the useful patterns and fill the gap between the Student Academics and Employability. This paper includes survey on different prediction algorithms like Classification, decision tree algorithm, C4.5, Feature with Graph structure, Bayesian, RIPPER, SVM, and compares the best performances on different aspects.

Index terms –Classification, Decision Tree, Neural network and Machine Learning Algorithms.

I. Introduction

A. Educational Data Mining

Data mining is data analysis methodology used to identify hidden patterns in a large data set. It has been successfully used in different areas including the educational environment. Educational data mining is an interesting research area which extracts useful, previously unknown patterns from educational database [1] for better understanding, improved educational performance and assessment of the student learning process. Evaluating students' performance is a complex issue, which can't be restricted for the grading. Reasons of good or bad performances belong to the main interests of teachers, because they can plan and customize their teaching program, based on the feedback. Data mining is one of the approaches, which can provide an effective assistance in revealing complex relationships behind the grades [1].

B. Educational Data Mining Process

Figure 1 explains different stages of methodology for Mining information on educational data for academic decision making.

DOI: 10.18535/ijecs/v6i4.41

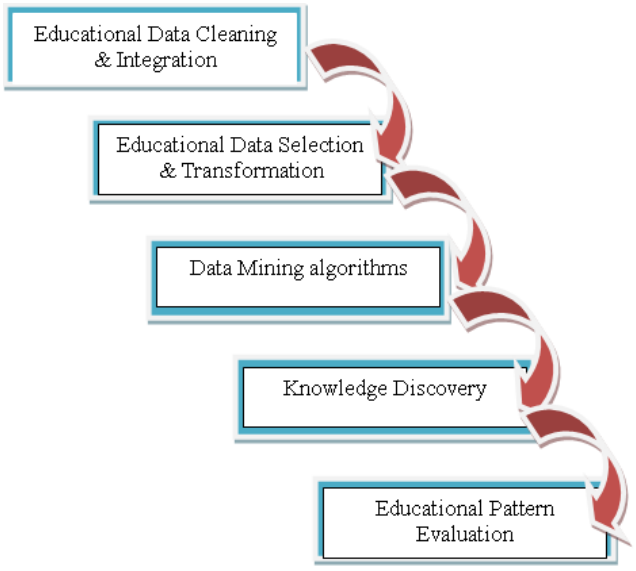


Fig 1: Educational Data mining Process

In the following section II discuss about the Literature survey on different classification algorithms And Section III Talks about the Algorithm and working process, IV Discuss about the Algorithms Performance factors

II.Literature Survey On Educational SYSTEMS USING Data Mining Algorithms

A. Survey on Different EDM Techniques and Process and Data Collection Process

Educational data mining is the most important aspect for the to identify the hidden patterns in the data. But, Collecting Student data not only collect from Engineering colleges and universities there are different trends student will interact with the ICT, MOODELS, MOOCS and student will interact with the courses evaluated through the different online systems .By using this data more number of authors are applying data mining Techniques are used make it as decision making system for the stake holders, Faculty for Engineering Institutes and universities. By using this decision making system are used to identify the hidden patterns, Knowledge using to make them as decision by using student patterns. So in this section we are study about the different Educational Data Mining Techniques different student and how to they make as a decision system in Educational System Make them use. In the section we are survey about the different educational data mining predictive algorithms papers from 2013 to 2015 following

[September 2013] Predicting Students' Performance Using Id3 And C4.5 Classification Algorithms creators Kalpesh Adhatrao et al India are proposed the desion tree calculations like ID3 and C4.5 calculation how to foresee the understudy exhibitions furthermore examine about the An instructive establishment needs a rough earlier information of enlisted understudies to anticipate their execution in future scholastics. This helps them to distinguish promising understudies furthermore gives them a chance to pay consideration on and enhance the individuals who might presumably get lower evaluations. As an answer, we have built up a framework which can anticipate the execution of understudies from their past exhibitions utilizing ideas of information mining strategies under Classification. We have investigated the information set containing data about understudies, for example, sexual orientation, marks scored in the board examinations of classes X and XII, stamps and rank in placement tests and results in first year of the past clump of understudies. By applying the ID3 (Iterative Dichotomiser 3) and C4.5 arrangement calculations on this information, we have anticipated the general and individual execution of crisply conceded understudies in future examinations. [2014] A Classification Model on Graduate Employability Using Bayesian Approaches: A Comparison creators Bangsuk Jantawan1, Cheng-Fa Tsai are examine about the study introduces a graduate employability display that uses Bayesian techniques to look the most vital element of graduate employability, and to think about the exactness of every calculation under Bayesian strategies including Naïve Bayesian Simple, Naïve Bayesian, Averaged One-Dependence Estimators, Averaged One-Dependence

Estimators with sub assumption determination, Bayesian systems, and Naïve Bayesian Updateable. The outcomes demonstrate that 3 components with an immediate impact on employability are the work territory, occupation sort, and times look for some kind of employment[4]. [2014] Assessment of Robust Learning with Educational Data Mining creators Ryan S. Dough puncher, Albert T. Corbett, from Carnegie Mellon University and examine about the Many college pioneers and staff have the objective of advancing discovering that associates crosswise over areas and gets ready understudies with aptitudes for their entire lives. In any case, as evaluation rises in advanced education, numerous appraisals concentrate on information and aptitudes that are particular to a solitary space. Improving appraisal in advanced education to concentrate on more vigorous learning is an essential step towards making evaluation coordinates the objectives of the setting where it is being connected. Specifically, appraisal ought to concentrate on whether learning is vigorous (Koedinger, Corbett, and Perfetti, 2012), whether learning happens in a way that exchanges, gets ready understudies for future learning, and is held after some time; furthermore on aptitudes and meta–competencies that sum up crosswise over areas. Thusly, we can gauge the results that we as instructors need to make, and build the chance that our appraisals offer us to enhance the results we some assistance with wishing to make. In this article, we examine and look at both customary test–based techniques for surveying hearty learning, and better approaches for deducing strength of learning while the learning itself is happening, contrasting the strategies inside of the area of school hereditary qualities[5].

[2015] Towards Automatic Prediction of Student Performance in STEM Undergraduate Degree Programs creators Laci Mary Barbosa Manhães,Sérgio Manuel Serra da Cruz,Seropédica, Brazil structure Universidad Federal do are told about the STEM is characterized as learning in the fields of Science, Technology, Engineering and Mathematics. In Brazil, numerous understudies leave the instructive framework before accomplishing a tertiary degree in these fields. Poor scholarly execution in STEM college classes is an issue confronted by numerous colleges, both in created and developing nations. In spite of the fact that these colleges store a lot of information, there are few learns about instructive information mining (EDM) programming apparatuses intended to help instructive chiefs in examining understudy learning and enhancing the nature of college degree programs. Our methodology might help administrators in managing understudies toward the end of every scholarly term, hence empowering them to distinguish the understudies in trouble of satisfying the scholastic prerequisites toward a degree. This paper demonstrates quantitative test contemplates utilizing an expansive dataset of genuine information from five conventional STEM college classes of one of the biggest open Brazilian colleges. At long last, the outcomes demonstrate that information mining calculations can set up successful forecast models from existing understudy information.

[January 2015] Performance Analysis and Prediction in Educational Data Mining: creators Pooja Thakar,Anil Mehta,Manisha from Banasthali University Jaipur The consistent mission is on to discover better approaches to make it more compelling and productive for understudies. These days, heaps of information is gathered in instructive databases, yet it remains unutilized. Keeping in mind the end goal to get required advantages from such a major information, intense apparatuses are required. Information mining is a rising capable instrument for investigation and expectation. It is effectively connected in the region of misrepresentation discovery, publicizing, promoting, advance evaluation and expectation. In any case, it is in incipient stage in the field of training. Extensive measure of work is done in this course, yet at the same time there are numerous untouched ranges. In addition, there is no brought together approach among these scrutinizes. This paper shows an extensive overview, a traveling (2002-2014) towards instructive information mining and its degree in future[6].

Here we are study about the how the data prepare and how to apply on those student data on different predictive algorithms get the best results by using this survey.

III. Study on predictive analytics algorithms

By using above survey we are identify the different predictive algorithms on different areas and these algorithms are how to fit for the in educational system and to predicted student performances and faculty performances on different aspects . By using above study the predictive algorithms are Classifier, decision

tree algorithm, C4.5, Feature with Graph structure, Bayesian, RIPPER, SVM, algorithms we are study in detail in the following section.

Feature with Graph structure

Educational data mining applications, knowledge in features and form as a group structures. In each factor may have several levels, and it can be denoted as a group of dummy features .When performing feature selection ,we tend to select or not select features in the same group simeltuniously.Group lasso ,driving all coefficients in one group to zero together and thus resulting in group selection attract more and more attention .

Assume the features from k disjoint groups G={G1,G2,…..Gk}and there is no overlap between any two groups .With the group structure ,we can rewrite w into the block form as w={w1,w2,…..,wk} where wi corresponds to the vector of all coefficients of features in the ith group Gi .Then the group Lasso performs the lq,1-norm regularization on the model parameter as

Step 1

Lasso does not consider the group structure and selects a subset of features among all groups.

Penalty (W,G)=Sigmai=1Khi||wGi||q----------(I)

Step2

Group Lasso can perform group selection and select a subset of groups. once the group is selected, all futures in this group are selected.

Penalty (W,G)=alpha||w||1+(1-alpha)Sigmai=1Khi||wGi||q (II)

Step 3

Sparse group Lasso can select groups and features in the selected group at the same time

Advantages

Groups may overlap - one protein/gene may belong to multiple groups. In these situations, group Lasso does not correctly handles overlapping groups and a given coefficient only belongs to one group.

Disadvantages

Algorithms investigating overlapping groups. General overlapping group Lasso regularization is similar to that for group Lasso regularization in

Penalty (W,G)=alpha||w||1+(1-alpha)Sigma i=1 Khi||wGi||q

However, groups for overlapping group Lasso regularization may overlap, while groups in group Lasso are disjoint[7].

A. ***Naive Bayes classifiers***

It is known to be the simplest Bayesian classifier and it has become an important probabilistic important model and has been remark bull successful in practice despite its strong independence assumption .it has proven effective text classification, medical diagnosis and computer performance management, among other application. We will describe the model of naive Bayes classifier .Important task is to train a classifier that will output the posterior probability p(Y|X) for possible values of Y.According to Bayes thermo ,p(Y=Ck|X=x) can be represented as

P(Y=Ck|X=x)=p(X=x|Y=Ck)p(Y=Ck))

$$P(X=x)=\frac{p(X1=x1,X2=x2,\ldots.,Xd=xd|Y=Ck)p(Y=Ck)}{P(X1=x1,X2=x2,...,Xd=xd)}$$

One way to learn p(Y|X)is to use the training data to estimate p(X|Y) and p(Y).We can then use these estimates, together with the Bayes theorem, to determine p(Y|X=x(i)) for any new instances x(i).
Advantages
The Naive Bayes technique is very fast, highly scalable model building and scoring. It balances linearly with the number of predictors and rows. The build process for Naive Bayes is parallelized.
Disadvantages No variable dependency, which is much unjustified for the real-life data Best[7]

B. C45 Decision tree algorithm

C4.5 makes binary splits for numerical attributes and k-way splits for categorical attributes.ID3 used the information Gain criterion.C4. Normally uses the gain ratio, with the caveat that the chosen splitting rule must also have an information gain that is stronger than the average information gain. Numerical attributes are first sorted..However instead of selecting the mid points,C4.5 consider each of the values themselves as the split points. If the sorted values(x1,x2,x3,….,xn), then the condition rules are {x>x1.x>x2…..,x>xn-1}.optionally, instead of splitting categorically attributes in to k branches, one branch for each different attributes values, they can be split into b branches, where b is a user defined number .To implement this C4.5 first perform the k-way split and then merges the most similar children until there are b children remaining

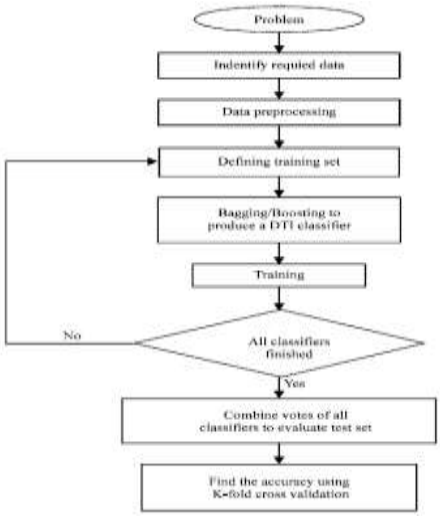


Fig 2:C4.5 Decision Tree Algorithm Process flow diagram

Advantages
C4.5 produces the accurate result, It takes the less memory to large program execution, takes less model build time, It has short searching time.
Disadvantages
C4.5 has Empty branches, and insignificant branches, Over fitting problems

C. Ripper Classification for ordered Classes

Ripper learn all rules for each class individually, in particularly, only after rule learning for one class is completed it moves on to the next classes as such all rules for each class appear to gather in rule decision list. The sequence rules for each individual class is not important, but one rules the rules for the least frequent class first, then second minority class and so, on. This process ensured that some rules are learn for rear or minority classes ,otherwise they maybe denoted by frequent or majority classes and we will end with no rules for minority classes. the ripper rule induction algorithms shown[10]

Ripper algorithm and its Variation (D, C)

1. Rule List<-Null;

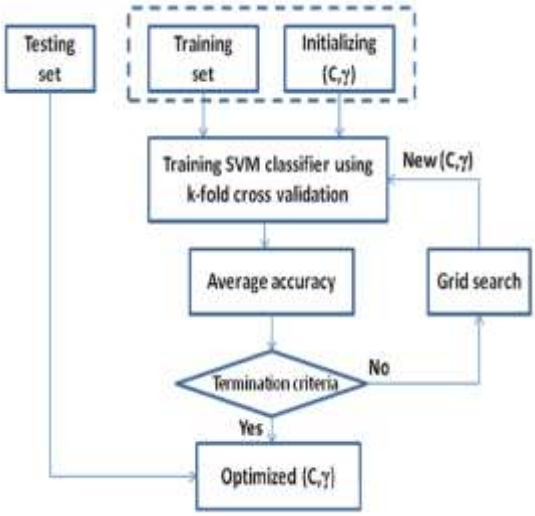
2. For each class c Belongs to C do
3. Prepare data(Pos,Neg) where Pos contains all the examples of class c from D, and Neg contains the rest of examples in D;
4. **While** Pos!= Null do
5. Rule<-learn –one –rule(pos,neg,c)
6. **If** rule is NULL then
7. Exit-While-Loop
8. **Else** Rule list<- insert Rule at the end of Rule List;
9. Remove examples covered by Rule from(Pos,Neg);
10. **End If**
11. **End While**
12. **End For**
13. **Output Rule list**

Advantages

Ripper is abbreviated as a Repeated Incremental Pruning to Produce Error Reduction .It is a classification algorithm deliberate to calculate rules set openly from the training dataset. The name is drawn from the reality that the rules are learned incrementally. A new rule related with a class value will cover a variety of attributes of that class .The algorithm was deliberate to be fast and successful when dealing with large and noisy datasets compared to decision trees.

D. Support Vector Machine Classification Algorithms

Machine learning algorithms have tendency to over fit. It is possible to achieve arbitrary low training error with some complex models, but testing error may be high, but the testing error may be high, because of poor generalization to unseen test instances .This is probeblematic, because the goal of classification is not to obtain good accuracy on known training data, but to predict unseen test instances correctly. SVM is theoretically sound approach for control model complexity. It picks important instances to construct the separating surface between data instances. When the data is not linearly separable, it can either penalize violation with loss terms, or leverage kernel tricks to construct non lingering separating surfaces.SVM can also perform multiclass classification in various ways, either by an ensemble of binary classifiers or by extending margin concepts. the optimization techniques of SVMs are mature[11].



DOI: 10.18535/ijecs/v6i4.41

Fig 3: Support Vector Machine Process Flow chart

Advantages

SVM is now considered a mature machine learning method. It has been widely applied in different applications. Furthermore, SVM has been well studied, both from theoretical and practical perspectives.SVM significantly faster.

IV. Study about Performance on classification algorithms

In this section we are study about different factors about classification algorithms accuracy factors those are nothing but predictive conditions. Predictive conditions are

1.Total Population :Total population is describe about the condition of positive and negative condition so finally total population is find out by

Prevalence=sum (Condition positive)/sum (total population)

2. Predictive Condition Positive

There are two condition True positive, false positive (type I error),so we find out positive predictive value(PPV),

Precision =Sigma (true positive)/Sigma (Test outcome positive)

False discovery rate (FDR) =Sigma (False Positive)/sigma (Test outcome positive) .

3. Predictive Condition Positive:

There are two condition True negative, false negative (type II error),so we find out negative predictive value(NPV).

False Omission Rate (OMR) =Sigma (false negative)/Sigma (Test outcome negative)

Negative Predictive Value (NPV) =Sigma (True Negative)/sigma (Test outcome negative)

4. Accuracy (ACC): Here is accuracy calculate two conditions True positive rate(TPR),False Positive rate

Accuracy (ACC) = Σ True positive + Σ True negative/Σ Total population

True positive rate (TPR), Sensitivity, Recall = Σ True positive/Σ Condition positive

False positive rate (FPR), Fall-out = Σ False positive/Σ Condition negative

True negative rate (TNR), Specificity (SPC) = Σ True negative/Σ Condition negative

Positive likelihood ratio (LR+) = TPR/FPR

Negative likelihood ratio (LR−) = FNR/TNR

Diagnostic odds ratio (DOR) = LR+/LR−

By using these factors to analysis the Different classification algorithms. By using thus predictive algorithms study about the factors which one efficient algorithm by using classification polar analysis to predictive rules[18].

Table 1: Predictive algorithm and performances

Classification Algorithm	**FP Rate**	**Precision**	**Recall**	**F-M**	**MCC**
Naive Bayes	1.000	0.000	1.000	1.000	1.000
C4.5	0.957	0.031	0.957	0.957	0.957
RIPPER	0.77	0.17	0.73	0.77	0.75

Study on Educational data mining predictive algorithms are Naive Bayes, C4.5mRipperAnd SVM algorithms using to find out accurate values from educational data bases and datasets and Naive Bayes, C.5, RIPPER algorithms for find out accuracy on classification of data values from above shown Table 1.

DOI: 10.18535/ijecs/v6i4.41

VI. CONCLUSION

In this paper, we studied about, how to analyze and predict rules from Educational Data bases and Educational data sets for student performance improvement and to find out the path for placements. By applying the advanced data mining techniques like Feature with Graph structure, Naive Bayes, C4.5, Ripper and Machine learning algorithm SVM. These are some of the predictive algorithms to apply on educational data to generate the rules and check how many these yields correct results. That is, To find out the positive values and negative values (Poles) to find out fitness and accuracy of the algorithm and analyze the rule to find out the correctly predictive rules and not correctly predictive rules also .

References

1. Mosima Anna Masethe, Hlaudi Daniel Masethe: Prediction of Work Integrated Learning Placement Using Data Mining Algorithms Proceedings of the World Congress on Engineering and Computer Science 2014 Vol I WCECS 2014, 22-24 October, 2014, San Francisco, USA
2. May, P., Ehrlich, H.C., Steinke, T.: ZIB Structure Prediction Pipeline: Composing a Complex Biological Workflow through Web Services. In: Nagel, W.E., Walter, W.V., Lehner, W. (eds.) Euro-Par 2006. LNCS, vol. 4128, pp. 1148--1158. Springer, Heidelberg (2006).
3. D. Jackson, "The contribution of work-integrated learning to undergraduate employability skill outcomes," Asia-Pacific J. Coop.Educ., vol. 14, no. 2, pp. 99–115, 2013.Zafra and S. Ventura, "Predicting Student Grades in Learning Management Systems with Multiple Instance Genetic Programming," no. Mil, pp. 307–314, 2009.
4. K. Pal, "Classification Model of Prediction for Placement of Students," Int. J. Mod. Educ. Comput. Sci., vol. 11, pp. 49–56, 2013.
5. N. Nghe, P. Janecek, and P. Haddawy, "A comparative analysis of techniques for predicting academic performance," in 37th ASEE/IEEE Frontiers In Education Conference, 2007, pp. 7–12.
6. AZIZ, N. ISMAIL, and F. AHMAD, "MINING STUDENTS'ACADEMIC PERFORMANCE.," J. Theor. Appl. Inf. T echnol., vol. 53, no. 3, 2013.
7. Mitchell, T.M.: Machine Learning. McGraw-Hill, New York (1997).
8. USNational Library of Medicine Natonal Institutes of Health:National Library of Medicine: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3936971/
9. Anylatical Vidya: http://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/
10. S. Milinković and M. Maksimović, "USING DECISION TREE CLASSIFIER FOR ANALYZING STUDENTS ' ACTIVITIES," JITA, vol. 3, no. 2, pp. 87–95, 2013.
11. Ghasemian, M. Moallem, and Y. Alipour, "Predicting students ' grades using fuzzy non-parametric regression method and ReliefFbased algorithm," Adv. Comput. Sci. an Int. J., vol. 3, no. 2, pp. 43– 51, 2014.
12. Machine learning in acion by PeterHarrington
13. M'hammed Abdous, W. He, and C.-J. Yen, "Using Data Mining for Predicting Relationships between Online Question Theme and Final Grade," Educ. Technol. Soc., vol. 15, no. 3, pp. 77–88, 2012.
14. S. Sembiring, M. Zarlis, D. Hartama, S. Ramliana, and E. Wani,"PREDICTION OF STUDENT ACADEMIC PERFORMANCE BY AN APPLICATION OF DATA MINING TECHNIQUES.," in International Conference on Management and Artificial Intelligence (IPEDR), 2011, vol. 6, pp. 110–114.
15. K. R. Lakshmi, M. V. Krishna, and S. P. Kumar, "Performance Comparison of Data Mining Techniques for Predicting of Heart Disease Survivability," Int. J. Sci. Res. Publ., vol. 3, no. 6, pp. 1–10, 2013.
16. Badr, E. Din, and I. S. Elaraby, "Data Mining : A prediction for Student ' s Performance Using Classification Method," World J. Comput. Appl. Technol., vol. 2, no. 2, pp. 43–47, 2014.
17. http://www.ftpress.com/articles/article.aspx?p=2248639&seqNum=5
18. http://www.analyticsvidhya.com/blog/2015/09/naive-bayes-explained/
19. http://colobu.com/2015/11/05/common-machine-learning-algorithms/.