Volume 3, Issue 12, December 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Review: Mining Educational Data to Forecast Failure of Engineering Students

Komal S. Sahedani, Prof. B Supriya Reddy
C.E. Department, R.K. University
India

Abstract:- Now a days there is an increasing interest in data mining and educational systems, make educational data mining as a new growing research community. The goal of institutions is to give quality education to its students. One way to achieve highest level of quality in higher education system is by discovering knowledge for prediction regarding enrolment of students in a particular course In Our Data driven data mining model, knowledge is originally existed in data, but just not understandable for human. Data mining is taken as a process of transform knowledge into some human understandable format like rule, formula, theorem, etc. This article provides a Review of the available literature on Educational Data mining, Classification method and different feature selection techniques that we should apply on Student dataset. The knowledge is hidden among the educational data set and it is extractable through data mining techniques.

Keywords: Data Mining, Education data mining, Knowledge discovery from data (KDD), Decision Tree, Classification techniques, Attribute Selection techniques.

I. INTRODUCTION

Data mining is the The iterative and interactive process of discovering valid, novel, useful, and understandable knowledge (patterns, models, rules etc.) in Massive databases
The main term that data mining support for data is

Valid: generalize to the future
Novel: what we don't know
Useful: be able to take some action
Understandable: leading to insight
Iterative: takes multiple passes
Interactive: human in the loop

Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging. [1]

Over past few years, many numbers of engineering institutes have opened rapidly in India. This causes a cut throat competition for attracting the student to get them enroll in their campus. Most of the institutes are opened in self-finance mode, so all the time they feel short hand in expenditure. Quality education is one of the most promising responsibilities of any University/ Institutions to their students. Quality education does not mean high level of knowledge produced. But it means that education is produced to students in efficient manner so that they learn without any problem. For this purpose quality education includes features like methodology of teaching, continuous evaluation, categorization of student into similar type, so that students have similar objectives, demographic, educational background etc. [2]

Engineering degrees are mostly offered in different curriculum structures. Engineering students are to fulfill strict requirements in order to graduate and hold a degree in engineering profession. Engineering students' accounts for numbers of departments mainly civil, electrical, mechanical, computer, electronics, communication, information technology, chemical, mining, metallurgical, textile, environment etc., Most of the engineering institutes' first five/six major courses.

This education is residential and at the beginning, student affects due to various factors related to their academic path. Most of the core courses are usually same for all the students in first year. They comprise essentially Mathematics, Physics and chemistry courses. These course are the prerequisites of almost all major courses, students are exposed to the fundamental and basic concepts required to pursue specialized theories on their further studies. Core courses play a decisive role in the student performance and enrolled in this study.

So Due to a greater number of students and institutions, higher education institutions (HEIs) are becoming more oriented to performances and their measurement and are accordingly setting goals and developing strategies for their achievements[5]

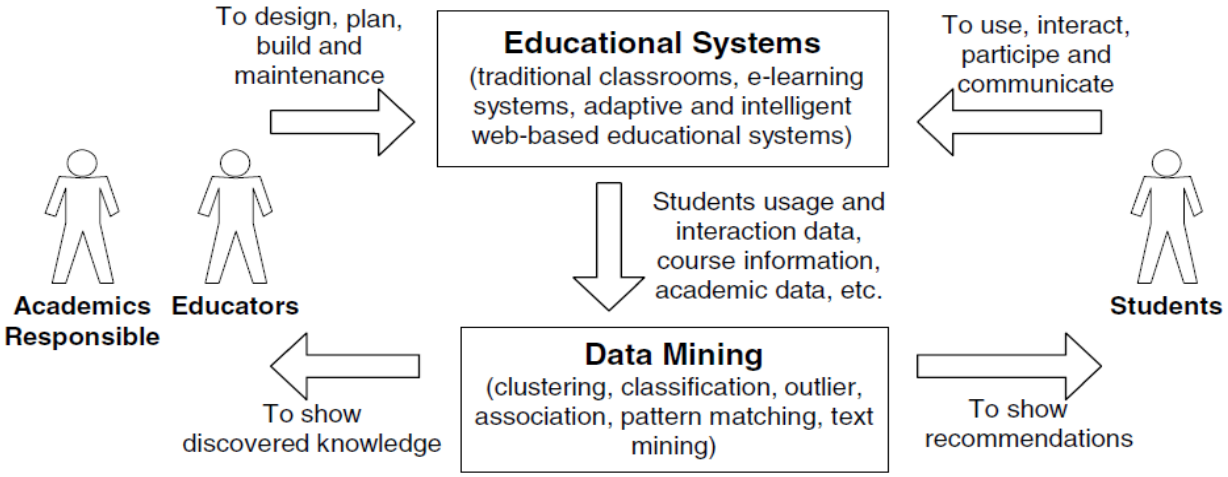

Fig. 1. The cycle of applying data mining in educational systems.

Here as shown in Fig. 1, educators and academics responsible are in charge of designing, planning, building and maintaining the educational systems. Students use and interact with them. Starting from all the available information about courses, students, usage and interaction, different data mining techniques can be applied in order to discover useful knowledge that helps to improve the e-learning process. The discovered knowledge can be used not only by providers (educators) but also by own users (students). So, the application of data mining in educational systems can be oriented to different actors with each particular point of view. [4].

The recent literature related to Educational data mining (EDM) is presented. Educational data mining is an emerging discipline that focuses on applying data mining tools and techniques to educationally related data. Researchers within EDM focus on topics ranging from using data mining to improve institutional effectiveness to applying data mining in improving student learning processes. There is a wide range of topics within educational data mining. So this paper will focus exclusively on ways that data mining is used to improve student success and processes directly related to student learning. For Example, Student success and retention, personalized recommender systems, and evaluation of student learning within course management system(CMS) are all topics within the broad field of educational data mining.

A large number of engineering students got failure during their Engineering Course. The paper is structured as follows: In Section II presents KDD (Knowledge Discovery from Database) process. In Section III different decision Tree Method of classification technique are explained, like ID3, C4.5, CART and ADT. In Section IV presents Different Attribute Selection Techniques for filtering some best attributes from students database.

II. KDD PROCESS

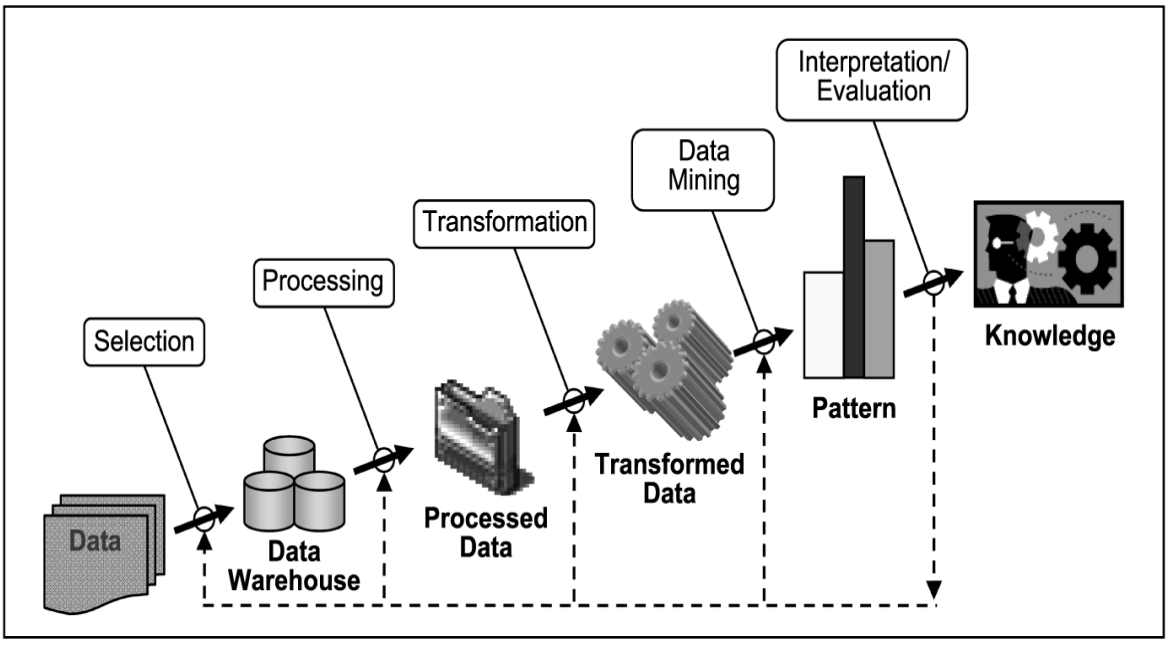

Figure 2 KDD Process

Knowledge discovery as a process is depicted in Figure 2 KDD have iterative sequence of the following steps:[16]

1. **Develop an understanding for the application domain and identify the goal.**
2. **Create a target dataset**

Selecting a dataset or focusing on a subset of samples or variables on which to make discoveries

3. **Data cleaning and preprocessing**

removing of noise and outliers from collecting necessary information to model or account for noise handling of missing data accounting for time sequence information.

4. **Data reduction and projection**

Finding useful features to represent the data relative to the goal dimensionality reduction/transformation ==> *reduce number of variables* identification of invariant representations

5. **Selection of appropriate data-mining task**

Summarization, classification, regression, clustering, etc.

6. **Selection of data-mining algorithm(s)**

Methods to search for patterns decision of which models and parameters may be appropriate match method to goal of KDD process

7. **Data-mining**

Searching for patterns of interest in one or more representational forms

8. **Interpretation and visualization**

Interpretation of mined patterns visualization of extracted patterns and models visualization of the data given the extracted models

Data mining includes fitting models to or determining patterns from observed data. The fitted models play the role of brings knowledge. Deciding whether the model reflects useful knowledge or not is a part of the overall KDD process for which subjective human judgment is usually required.

The more common Techniques in current data mining practice include the following.

1) **Classification**: classifies a data item into some of several predefined categorical classes.
2) **Regression**: maps a data item to a real valued prediction variable.
3) **Clustering**: Clustering is maximization of similarity and minimization of dissimilarity between categorical classes.
4) **Rule generation**: extracts different classification rules from the data.
5) **Discovering association rules**: describes association relationship among different attributes.
6) **Summarization**: provides a compact description for a subset of data.
7) **Dependency modeling**: describes relating dependencies among variables.

III. DECISION TREE

Decision trees are often used in classification and prediction. It is simple yet a powerful way of knowledge representation.

The decision tree classifier has two phases [6]:

a) Growth phase or Build phase.

b) Pruning phase.

The tree is built in the first phase by recursively splitting the training set based on local optimal criteria until all or most of the records belonging to each of the partitions bearing the same class label. The tree may over fit the data. The pruning phase handles the problem of over fitting the data in the decision tree. The prune phase generalizes the tree by removing the noise and outliers. The accuracy of the classification increases in the pruning phase. Pruning phase accesses only the fully grown tree. The growth phase requires multiple passes over the training data. The time needed for pruning the decision

Tree is very less compared to build the decision tree.

A. ID3 (Iterative Dichotomies 3)

This is a decision tree algorithm introduced in1986 by Quinlan Ross [8]. ID3 is based on Hunts algorithm. The tree is constructed in 2 phases. The two phases are tree building and pruning. ID3 uses information gain measure to choose the splitting attribute. It only approves categorical attributes in building a tree model. It does not give accurate result when there is noise. To remove the noise preprocessing technique has to be used.

To build decision tree, information gain is calculated for every single attribute and select the attribute with the greatest information gain to designate as a root node. Label the attribute as a root node and the possible values of the attribute are represented as arcs. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the instances are falling under the same class, the node is represented with single class name, otherwise choose the splitting attribute to classify the instances.

Continuous attributes can be handled using the ID3 algorithm by discretizing or directly, by considering the values to find the best split point by taking a threshold on the attribute values. ID3 does not support pruning.

B. C4.5

This algorithm is a successor to ID3 developed by Quinlan Ross [8]. It is also based on Hunt's algorithm.C4.5 handles both categorical and continuous attributes to build a decision tree. In order to handle continuous attributes, C4.5 splits the attribute values into two partitions based on the selected threshold such that all the values above the threshold

as one child and the remaining as another child. It also handles missing attribute values. C4.5 uses Gain Ratio as an attribute selection measure to build a decision tree. It removes the biasness of information gain when there are many outcome values of an attribute. At first, calculate the gain ratio of each attribute. The root node will be the attribute whose gain ratio is maximum. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification.

C. ***CART***

CART stands for Classification and Regression Trees introduced by Breiman [8]. It is also based on Hunt's algorithm. CART handles both categorical and continuous attributes to build a decision tree. It handles missing values.

CART uses Gini Index as an attribute selection measure to build a decision tree .Unlike ID3 and C4.5 algorithms, CART produces binary splits. Hence, it produces binary trees. Gini Index measure does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy.

D. ADT (Alternating Decision Tree)

ADTrees were introduced by Yoav Freund and Llew Mason [9]. It generalizes decision trees and has connections to boosting.

An alternating decision tree consists 2 nodes. One is decision nodes and other is prediction nodes. First nodes specify a predicate condition. Second nodes contain a single number. ADTrees have prediction nodes as both root and leaves also.

IV. ATTRIBUTE SELECTION TECHNIQUES

Weka tool support many attribute selection techniques.
Different Attribute Selection Techniques that would be applied on Educational Database those are explained as below.

1. Cfssubseteval

Synopsis

- Evaluates the worth of a subset of attributes by considering the individual predictive ability of each feature along with the degree of redundancy between them.

Options:

Option	Description
locally Predictive	Identify locally predictive attributes. Iteratively adds attributes with the highest mutual relationship with the class as long as there is not already an attribute in the subset that has a higher correlation with the attribute in question
missingSeperate	Take missing as a separate value.

Capabilities:

Capability	Supported
Class	Missing class values, Numeric class, Nominal class, Date class, Binary class
Attributes	Empty nominal attributes, Nominal attributes, Numeric attributes, Unary attributes, Date attributes, Binary attributes, Missing values
Min # of instances	1

2. Chisquaredattributeeval

Synopsis

Chisquaredattributeeval is evaluates an attribute by computing the value of the chi-squared statistic with respect to the class.

Options

Option	Description
binarizeNumericAttributes	Only binarize numeric attributes instead of properly discretizing them.
missing Merge	Distribute the counts for missing values. Then Counts are distributed across other values in proportion to their frequency. Or else, missing is treated as a separate value.

Capabilities:

Capability	Supported
Class	Binary class, Nominal class, Missing class values
Attributes	Nominal attributes, Missing values, Numeric attributes, Unary attributes, Date attributes, Empty nominal attributes, Binary attributes
Min # of instances	1

3. Consistency-subseteval

Synopsis

Evaluates a subset of attributes when the training instances are projected onto the subset of attributes by the level of consistency in the class values

Capabilities

Capability	Supported
Class	Nominal class, Missing class values, Binary class
Attributes	Date attributes, Empty nominal attributes, Nominal attributes, Numeric attributes, Binary attributes, Missing values, Unary attributes
Min # of instances	1

4. Filteredattributeeval

Synopsis

Class for running an arbitrary attribute evaluator on data that has been passed through an arbitrary filter (note: filters that alter the order or number of attributes are not allowed). Like the evaluator, the structure of the filter is based exclusively on the training data.

Options

Option	Description
attributeEvaluator	The attribute evaluator to be used.
filter	The filter to be used.

Capabilities

Capability	Supported
Class	Nominal class, Binary class
Attributes	Missing values, Date attributes, Unary attributes, Empty nominal attributes, Numeric attributes, Nominal attributes, Binary attributes, Relational attributes, String attributes
Min # of instances	0

5. OneRAttributeEval

Synopsis

By using the OneR classifier evaluates an attribute.

Options

Option	Description
evalUsingTrainingData	Use the training data to evaluate attributes rather than cross validation.
folds	Set the number of folds for cross validation.
minimumBucketSize	The minimum number of objects in a bucket (passed to OneR).
seed	Set the seed for use in cross validation.

Capabilities

Capability	Supported
Class	Binary class, Missing class values, Nominal class
Attributes	Date attributes, Nominal attributes, Empty nominal attributes, Missing values, Unary attributes, Numeric attributes, Binary attributes
Min # of instances	1

6. **FilteredSubsetEval**

Synopsis

Class for running an arbitrary subset evaluator on data that has been passed through an arbitrary filter (note: filters that alter the order or number of attributes are not allowed). Like the evaluator, the structure of the filter is based exclusively on the training data.

Options

Option	Description
filter	The filter to be used.
subsetEvaluator	The subset evaluator to be used.

Capabilities

Capability	Supported
Class	Nominal class, Binary class
Attributes	Relational attributes, Nominal attributes, Missing values, Binary attributes, Empty nominal attributes, Unary attributes, Numeric attributes, String attributes, Date attributes
Min # of instances	0

7. GainRatioAttributeEval

Synopsis

Evaluates an attribute by measuring the gain ratio with respect to the class.
GainR (Class, Attribute) = (H (Class) - H (Class | Attribute)) / H (Attribute).

Options

Option	Description
missingMerge	Distribute counts for missing values. Counts are distributed over other values in proportion to their frequency. Or else, missing is treated as a separate value.

Capabilities

Capability	Supported
Class	Missing class values, Nominal class, Binary class
Attributes	Nominal attributes, Date attributes, Binary attributes, Empty nominal attributes, Numeric attributes, Missing values, Unary attributes
Min # of instances	1

8. InfoGainAttributeEval

Synopsis

Evaluates an attribute by measuring the information gain with respect to the class.
Info Gain (Class, Attribute) = H (Class) - H (Class | Attribute).

Options

Option	Description
binarizeNumericAttributes	Just binarize numeric attributes rather than properly discretizing them
missing Merge	Distribute the counts for missing values. Counts are distributed over other values in proportion to their frequency. Or else, missing is treated as a separate value.

Capabilities

Capability	Supported
Class	Binary class, Nominal class, Missing class values
Attributes	Nominal attributes, Missing values, Numeric attributes, Unary attributes, Date attributes, Empty nominal attributes, Binary attributes
Min # of instances	1

9. **ReliefFAttributeEval**

Synopsis

Evaluates the worth of an attribute by repeatedly sampling an instance and considering the value of the given attribute for the nearest instance of the same and different class can operate on both discrete and continuous class data.

Options

Option	Description
numNeighbours	Number of nearest neighbors for attribute estimation.
sampleSize	Number of instances to sample. Default (-1) indicates that all instances will be used for attribute estimation.
seed	Random seed for sampling instances.
sigma	Set influence of nearest neighbors. Used in an exp function to control how quickly weights decrease for more distant instances. Use in conjunction with weight By Distance. Sensible values = 1/5 to 1/10 the number of nearest neighbors.
weightByDistance	Weight nearest neighbors by their distance.

Capabilities

Capability	Supported
Class	Nominal class, Numeric class, Binary class, Missing class values, Date class
Attributes	Missing values, Date attributes, Empty nominal attributes, Nominal attributes, Binary attributes, Unary attributes, Numeric attributes
Min # of instances	1

10. SymmetricalUncertAttributeEval

Synopsis

Evaluates an attribute by measuring the symmetrical uncertainty with respect to the class.
SymmU (Class, Attribute) = 2 * (H (Class) - H (Class | Attribute)) / H (Class) + H (Attribute).

Options

Option	Description
Missing Merge	Distribute counts for missing values. Counts are distributed over other values in proportion to their frequency. Or else, missing is treated as a separate value.

Capabilities

Capability	Supported
Class	Missing class values, Nominal class, Binary class
Attributes	Nominal attributes, Date attributes, Binary attributes, Empty nominal attributes, Numeric attributes, Missing values, Unary attributes
Min # of instances	1

V. CONCLUSION

Data mining have the ability to uncover hidden patterns in large databases; community colleges and universities can build models that predict with a high degree of accuracy the behavior of population clusters. By acting on these predictive models, educational institutions can effectively address issues ranging from transfers and retention, to marketing and alumni relations.

REFERENCES

[1] U. Fayyad and R. Uthurusamy, "Data mining and knowledge discovery in databases," Commun. ACM, vol. 39, pp. 24–27, 1996.

[2] Shiv Kumar Gupta,Sonal Gupta &Ritu Vijay," prediction of student success that are going to enroll in the Higher technical education", IJCSEITR, ISSN 2249-6831, Vol. 3, Issue 1, Mar 2013, pp. 95-108.

[3] Richard A. Huebner,Norwich University," A survey of educational data- mining research", Research in Higher Education Journal,2012,pp-1-13

[4] C. Romero *, S. Ventura.(2007)"Educational data mining: A survey from 1995 to 2005",ScienceDirect Expert Systems with Applications 33 pp. 135–146,2007.

[5] Zeljko Garaca, Maja Cukusic, Mario jadric (2010), "Student Dropout Analysis with application of data mining methods",Vol 1,pp. 31-46

[6] U. Fayyad and R. Uthurusamy, "Data mining and knowledge discovery in databases," Commun. ACM, vol. 39, pp. 24–27 , 1996.
[7] Jiawei Han and Micheline Kamber, "Data Mining Concepts and Techniques", 2nd Edition, 2000.
[8] J. R. Quinlan, "Introduction of decision tree", Journal of Machine learning", pp. 81-106, 1986.
[9] Yoav Freund and Llew Mason, "The Alternating Decision Tree Algorithm". Proceedings of the 16th International Conference on Machine Learning, pp. 124-133, 1999.
[10] Bernhard Pfahringer, Geoffrey Holmes and Richard Kirkby. "Optimizing the Induction of Alternating Decision Trees". Proceedings of the Fifth Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, pp. 477-487, 2001.
[11] Saurabh Pal." Mining Educational Data to Reduce Dropout Rates of Engineering Students", IJIEEB, April-2012, Vol-2, pp.1-7.
[12] M. Ramaswami and R. Bhaskaran , " A CHAID Based Performance Prediction Model in Educational Data Mining" , IJCSI , Vol. 7 , Issue 1 , No. 1 , January 2010 , pp.10-18
[13] M. Ramaswami and R. Bhaskaran , " A CHAID Based Performance Prediction Model in Educational Data Mining" , IJCSI , Vol. 7 , Issue 1 , No. 1 , January 2010 , pp.10-18
[14] W. Hamalainen and M. Vinni," Classifiers for educational data mining", 2008, pp.1-34
[15] Mohammad Hassan Falakmasir, JafarHabibi," Using Educational Data Mining Methods to Study the Impact of Virtual Classroom in E-Learning", 2010, pp.241-248
[16] HyonamJeong,"How Students' Self-motivation and Learning Strategies Affect Actual Achievement", Department of Computer Science, Indiana University-Purdue University Fort Wayne
[17] Kun Liu, Yan Xing," A Lightweight Solution to the Educational Data Mining Challenge",2010.
[18] A.S. Kavitha,R. Kavitha, J. VijiGripsy," Empirical Evaluation of Feature Selection Technique in Educational Data Mining", ARPN Journal of Science and Technology, VOL. 2, NO. 11, Dec 2012.
[19] Carlos Marquez-Vera , Cristobal Romero Morales , and Sebastian Ventura Soto , "Predicting School Failure and Dropout by Using Data Mining Techniques" , IEEE journal of latin-american learning technologies , vol. 8 , no. 1 , February 2013.

