

PREDICTION OF DROPOUT: COMPARATIVE ANALYSIS ON ID3 AND C4.5

¹Suganya .S, ²V.Narayani

¹Research Scholar- Computer Science, Bharathiar University
Coimbatore, India

²Director i/c –MCA, Karpagam College of Engineering,
Coimbatore, India

¹nssuganya@gmail.com, ²narayaniv79@rediffmail.com

Abstract: Analyzing data mining method techniques on student's data of various schools have been pursued in order to construct data model for student's dropout. Theoretically, data mining is discovering special patterns among a massive set of fields in a large relational database. Data mining is also used for sorting the enlightening problem and analyzes the techniques to measure the student performance, instructor performance, student dropout, etc. This paper focuses on student dropout performance using classification techniques. The ability to predict dropout performance of the students is very decisive in our current education system. We can use data mining concepts for this purpose. The prominent algorithms are used to generate the decision trees. A decision tree is an effective method used for Classification. For that reason, it is the effortless hierarchical construction for the user to understand and make decisions. More mining algorithms exist for classification of various classifiers and rules. Student's performance is the topmost concern for any institution. A meticulous algorithm cannot be efficient to all applications, owing to the difficulty for suitable data. Consequently, the anthology of the acceptable mining algorithm exists not only for the goal of an application but also for the compatibility of the dataset. The attempt of this paper is to use appropriate data mining algorithms on an educational dataset and also to compare various data mining techniques. Both ID3 and C4.5 sample data are taken and executed through the tools available already. The purpose of this article is to present, compare and highlight the dropout performance of students

Keywords: Classification, C4.5, Data Mining, Educational dataset, ID3, Prediction, Student Dropout.

1. Introduction

From learning and an investigational point of view, EDM can be seen as a repeated process of theory formation, testing, and refinement. The main goal is not just to turn data into knowledge, but also to filter and

get mined knowledge for decision-making about how to modify the educational environment to reduce the dropout. This is a type of formative evaluation of an educational program while it is still in growth, and with the principle of constantly civilizing the education program. The thing is we need to analyze "How students use this system?" It is a pedagogical approach to assess instructional design in a significant mode. It may help the educational designers to improve their learning and teaching ideas for creating a book or other work. For example, EDM techniques have discovered models or patterns that can be used to support didactic designers to establish a scholastic and managerial decision. At hand, there are more of accepted methods inside EDM^[2,3,4,5]. Some of them are widely accredited to be universal across all types of data mining knowledge process. Figure.1 illustrates the process like prediction, clustering, outlier detecting, relationship mining, process mining, and text mining while others have particular eminence within EDM, like the distillation of data for human judgment, discovery with models, knowledge tracing (KT) and factorize non-negative matrix..

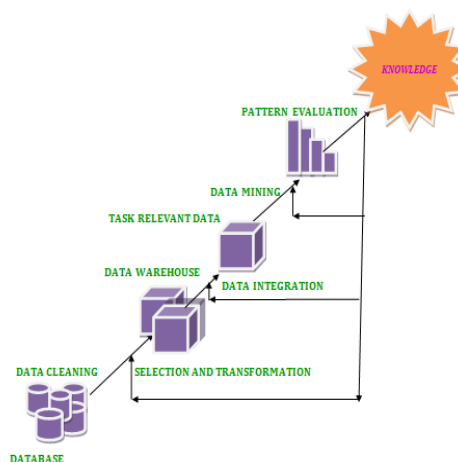


Figure 1. Data Mining Knowledge

The main concept of all the educational institutions is to improve educational excellence and also to decrease dropout. Prediction of the student's dropout in any education foundation is one way to reach the good quality of education culture system. Educational organization team should classify students, who are liable to fall short in the exam, attendance, conduct, family income, etc. But; it is difficult to make out them before time, because of a large number of enrolled students. The student's dropout performance can be exaggerated by countless factors such as personal, social, economic and demographic data, etc. Discovering useful information from a large database is a difficult task^[13]. Data mining helps in extraction of hidden patterns, behaviors and its tools are also used to predict future, allowing all types of business to make practical and to take knowledge driven decisions.

2. Decision-Tree – Classification

A famous branching method that illustrates the possible outcomes of the decision is called a decision tree. It is an important tool which resembles a tree. It is one of the pathways to exhibit an algorithm that only contains the trained and qualified statements that controls the model. The work of the decision tree is to create Classification and Regression model in the form of a tree. It breaks down the huge dataset into smaller units whereas, at the same period, a connected decision tree is accumulatively developed. The ultimate result is a tree with decision and leaf nodes. The Decision tree is one of the easiest and quickest ways to recognize most important variables and the relation between two or more variables. With the help of decision trees, we can create new variables that have superior control to predict target variable. In this article, we have focused on the essentials of the constructing a dataset. Later we presented the ID3 and C4.5 algorithm that respond to numeric and non-numeric data.

3. Id3 Algorithm

Ross Quinlan invented the Iterative Dichotomiser^[11] ID3 algorithm which is used to generate decision tree at the University of Sydney. Constructing a decision tree is all about finding an attribute that returns the Highest Information Gain. We have to calculate the entropy of target variable. The dataset is split into the different attributes. The entropy for each branch is calculated. Then it is added proportionally to get total entropy. The result will be the Information Gain or decrease in entropy. We have to choose the attribute with the largest Information Gain as the decision node. This process is done until last splitting. The ID3 algorithm runs recursively on the non-leaf branches. It employs a top-down, greedy search through the possible branches with no

backtracking. Lastly, the decision tree is converted to decision rules. ID3 is a supervised learning algorithm^[10] build a decision tree from a fixed set of examples. The resulting tree is used to classify future samples. ID3 algorithm builds a tree based on the information (information gain) obtained from the training instances and then uses the same to classify the test data. The ID3 algorithm generally uses nominal attributes for classification with no missing values^[10]. The pseudo-code of this algorithm is very simple. Given a set of attributes not a target variable say C_1, C_2, C_n . C is the target attribute, and a set S for recorded learning^[6].

The ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree. ID3 is a non incremental algorithm; it derives its classes from a fixed set of training instances. The classes created by ID3 are inductive, that is, given a small set of training instances, the particular classes created by ID3 are expected to work for all future instances. The allocation of the unknown instances must be the same as the test cases. Induction classes cannot be proven to work in every case since they may classify an infinite number of instances.

(A) ID3 Algorithm (Data Description)

The ID3 algorithm begins with the original set as the root node. On each iteration of the algorithm, it iterates through every unused attribute of the set and calculates the entropy of that attribute. It then selects the small attribute value as the small entropy value. The set is then split by the selected attribute to produce subsets of the data.

The sample data used by ID3 has certain requirements, which are:

- Attribute value description - the same attributes must be describe each example and have a fixed number of values.
- Predefined classes - an example's attributes must already be defined, that is, they are not learned by ID3.
- Discrete classes - classes must be sharply delineated.
- Sufficient examples - since inductive generalization is used (i.e. not provable) there must be enough test cases to distinguish valid patterns from chance occurrences.

(B) ID3 Algorithm (Attributes Selection)

ID3 decides which attribute will be the best? A statistical property, called Information Gain, is used. Gain measures how well a given attribute separates training examples into targeted classes. The one with the maximum information (information is the most useful for classification) is selected. In order to define Gain, we first scrounge an idea from information theory entropy. Entropy measures the amount of information in an attribute.

4. C4.5 Algorithm

C4.5 is an algorithm used to generate a decision tree developed and it is an extended version of ID3 Algorithm by Ross Quinlan. Many scholars made different kinds of improvement on the decision tree algorithm. But the problem is that these decision tree algorithms need multiple scanning and sorting of data collection several times in the construction process of the decision tree. The processing speed reduced greatly in the case that the data set is so large that can not fit in the memory. At present, the literature is about the improvement of the efficiency of decision tree classification algorithm in the literature^[7].

For example, Wei Zhao, Jamming Su proposed improvements in to the ID3 algorithm, which simplifies the information gain in the use of Taylor's formula. But this improvement is more suitable for a small amount of data, so it's not particularly effective in large data sets. Due to dealing with a large number of data sets, a variety of decision tree classification algorithm has been considered. The advantages of C4.5 algorithm are significant, so it can be chosen. But its efficiency must be improved to meet the dramatic increase in the demand for the large number of data. C4.5 adopt greedy (i.e., no backtracking) approach in which decision trees are constructed in a top-down recursive divide and conquer manner^[14]. Most algorithms for decision tree induction also follow such top-down approach, which starts with a training set of tuples and their associated class labels. The training set is recursively partitioned into smaller subsets as the tree is being built.

(A) C4.5 Algorithm (Data Description)

The sample data used by C4.5 has certain requirements. All the samples in the list belong to the same class. When this happens, it simply creates a leaf node for the decision tree saying to choose that class. None of the features provide any Information Gain. In this case, C4.5 creates a decision node higher up the tree using the expected value of the class. The instance of previously-unseen class encountered. Again, C4.5 creates a decision node higher up the tree using the value.

(B) C4.5 Algorithm (Attributes Selection)

The attribute selection measure provides a ranking for each attribute describing the given training tuples. The attribute having the best score for the measure is chosen as the splitting attribute for the given tuples. If the splitting attributes is continuous values or if we restrict to binary trees then, either a split point or splitting subset must also be determined as a part of splitting criterion.

5. Comparison Of Id3 And C4.5 Algorithm

The ID3 algorithm selects the best attribute based on the concept of entropy and information gain for developing the tree. C4.5 algorithm acts similar to ID3 but improves a few of ID3 behaviors:

- A possibility to use continuous data.
- Using unknown (missing) values
- Ability to use attributes with different weight
- Pruning the tree after being created.
- Pessimistic prediction error
- Sub Tree Raising

6. ID3 And C4.5 Algorithm

Table1. Comparison of ID And C4.5

Characteristic	ID3 Algorithm	C4.5 Algorithm
Splitting Criteria	Information Gain	Gain ratio
Attribute type	Handle only nominal values	Handle both nominal and numeric values
Missing values	Do not handle missing values	Handle missing values
Pruning Strategy	No pruning is done	Error based pruning is done
Outlier Detection	Susceptible on outliers	Susceptible on outliers

7. Evaluation On Training Set (ID3)

Table 2. ID3 Evaluation

Kappa statistic	1
Mean absolute error	0
Root mean squared error	0
Relative absolute error	0%
Root relative squared error	0%
Total Number of Instances	24

8. Evaluation On Training Set (C4.5)

(A) Detailed Accuracy Of ID3

Table3. ID3 Accuracy Rate

Kappa statistic	1
Mean absolute error	0.0148
Root mean squared error	0.0563
Relative absolute error	7.1466 %
Root relative squared error	17.6192 %
Total Number of Instances	24
Ignored Class Unknown Instances	1

(B) Detailed Accuracy Of C4.5

Table 4. C4.5 Evaluation

<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision Call</i>	<i>Fmeasure</i>	<i>Mcc</i>	<i>Roc Area</i>	<i>Prc Area</i>	<i>Class</i>
1.000	1.000	0.000	1.000	1.000	1.000	1.000	false
1.000	1.000	0.000	1.000	1.000	1.000	1.000	true

Table 5. C4.5accuracy Rate

<i>TP Rate</i>	<i>FP Rate</i>	<i>Precision Recall</i>	<i>Fmeasure</i>	<i>Mcc</i>	<i>Roc Area</i>	<i>Prc Area</i>	<i>Class</i>
0.938	0.125	0.938	0.938	0.813	0.949	0.964	False
0.875	0.063	0.875	0.875	0.813	0.949	0.843	True

9. Results and Discussion

Both the classification techniques have been functional on the sample dataset to build the classification model. The techniques are: the ID3 decision tree algorithm and the C4.5algorithm^[13]. After applying the appropriate action before processing and preparation methods, we try to study the structure of the data visually and figure out the circulation of values. Together the algorithm gives different accuracy. The standard attribute does not show any understandable outcome, while the Category and Sex have shown some effect in some of the experiments for predicting the dropout students. Machine learning algorithms such as the C4.5 algorithm can become skilled at predictive models from the student dropout data accumulated from the previous years, because the precision value for dropout student is highest.

10. Conclusion

The study can be further extensive to revise the concert of other classification techniques with larger sample dataset. From the above results it is clear that classification techniques can be applied to educational

data for predicting the beginner's outcome and improve their results. The sample data in the table 7 is used as a training data set to find the most likely students who are about to dropout from school. This prediction will help the Teachers to identify the student who is weak and help them to score better marks. The C4.5 and ID3 decision tree predict the number of learners who are likely to fail or pass. The result given to the trainer helps to step forward to improve the performance of the learners who have been predicted to fail. Since the application of data mining offers advantageous for school education system, but still the techniques can be applied in the other areas of education to optimize the resources, to predict the detainment of faculties in the institution and find out the solution of how to reduce student dropout rate in school education system. In future work, we can apply different data mining techniques on an expanded data set with more distinct attributes to get more accurate results.

References

- [1] <http://www.penpencileraser.com/blog/education-system-in-india-improvising-techniques-and-concepts-for-the-upcoming-schools/>.
- [2] Romero C, Ventura S. Educational data mining: a survey from 1995 to 2005. *J Expert Syst Appl* 2007,1:135-146.
- [3] Scheuer O, McLaren BM. Educational data mining. In: *The Encyclopedia of the Sciences of Learning*. New York, NY: Springer; 2011.
- [4] Baker RSJd. Data mining for education. In McGaw B, Peterson P, Baker E, eds. *International Encyclopedia of Education*. 3rd ed. Vol. 7. Oxford, UK: Elsevier; 2010,112-111.
- [5] Baker RSJd, Yacef K. The state of educational data mining in 2009: a review and future visions. *J EduData Min* 2009, 17.
- [6] Vincent GUIJARRO.K, Les Arbres de Décisions L'algorithme ID3, Elissa, "Title of paper if known," unpublished.
- [7] N. Kwad, C. H. Choi, Input feature selection for classification problem [J], *IEEE Trans on Neural Networks*, 2002,13(1): 143-159.
- [8] Ankur Shrivastava and Vijay Choudhary, Comparison between ID3 and C4.5 in Contrast to IDS Surbhi Hardikar, *VSRD-IJCSIT*, Vol. 2 (7), 2012, 659-667.
- [9] Carlos Márquez Vera, Cristóbal Romero Morales and Sebastián Ventura Soto, "Predicting of school failure and dropout by using data mining techniques", *The IEEE Journal of Latin-American Learning Technologies (IEEE-RITA)*, Vol. 8, No. 1, pp 7-14, Feb 2013.
- [10] A. Peña-Ayala, "Review: Educational data mining: A survey and a data mining-based analysis of recent works," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1432-1462, 2014.15.R.S.J.D Baker and K.Yacef, "The State of Education".
- [11] J.R. QUINLAN, Induction of Decision Trees, 1986, *Machine Learning* 1:81-106.
- [12] *Indian Journal of Science and Technology*, Vol 8(15), DOI: 10.17485/ijst/2015/v8i15/74555, July 2015.
- [13] <http://rulequest.com/see5-comparison.html> (cité dans l'article de R. Quinlan).
- [14] Johan Baltié, DataMining: ID3 et C4.5, Promotion 2002, Spécialisation S.C.I.A. Ecole pour l'informatique et techniques avancées.
- [15] S.V.Manikanthan and D.Sugandhi "Interference Alignment Techniques For Mimo Multicell Based On Relay Interference Broadcast Channel " *International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE)* ISSN: 0976-1353 Volume- 7 ,Issue 1 – MARCH 2014.
- [16] T. Padmapriya and V. Saminadan, "Distributed Load Balancing for Multiuser Multi-class Traffic in MIMO LTE-Advanced Networks", *Research Journal of Applied Sciences, Engineering and Technology (RJASET)* - Maxwell Scientific Organization, ISSN: 2040-7459; e-ISSN: 2040-7467, vol.12, no.8, pp:813-822, April 2016.
- [17] S.V.Manikanthan and K.Baskaran "Low Cost VLSI Design Implementation of Sorting Network for ACSFD in Wireless Sensor Network", *CiT International Journal of Programmable Device Circuits and Systems*, Print: ISSN 0974 – 973X & Online: ISSN 0974 – 9624, Issue :November 2011, PDCS112011008.
- [18] Rajesh, M., and J. M. Gnanasekar. "Congestion control in heterogeneous wireless ad hoc network using FRCC." *Australian Journal of Basic and Applied Sciences* 9.7 (2015): 698-702.

