Handling Missing Value in Decision Tree Algorithm

Preeti Patidar
Computer Science Dept.
LNCT, Indore (M.P.), India

Anshu Tiwari
Computer Science Dept.
LNCT, Indore (M.P.), India

ABSTRACT

Nowadays all the decisions making and large data analysis is made using computer applications. In such kind of application we use the data mining techniques to analyses them. Different domains of research like management, engineering, medical, education are frequently using these techniques. Data mining in educational system is an emerging discipline that focuses on applying data mining tools and techniques on educational data. Educational data mining is used to study the data available in the educational field and bring out the hidden knowledge from it.

In this research work, data mining techniques is used to make smart decisions for the student, additionally this technique is used to analysis the performance of the students in educational domain, to make analysis and making decisions here we are using C5.0 decision tree. Comparative study is done on ID3, C4.5 and C5.0. Among these classifiers C5.0 gives more accurate and efficient output with comparatively high speed. Memory usage to store the rule set in case of the C5.0 classifier is less as it generates smaller decision tree. This research work supports high accuracy, good speed and low memory usage as proposed system is using C5.0 as the base classifier. The classification process here has low memory usage compare to other techniques because it generates fewer rules. Accuracy is high as error rate is low on unseen cases. And it is fast due to yielding pruned trees.

Keywords

Data Mining, Decision Tree, Educational Data Mining, C4.5 and C5.0 Algorithm .

1. INTRODUCTION

Data mining conceptions and methods can be applied in several fields like marketing, medicine, education, real estate, customer relationship management, engineering, web mining etc. Educational data mining is a new emerging technique of data mining that can be applied on the data related to the field of education. There are increasing research interests in using data mining in education. This new emergent field, called Educational Data Mining; it is related with developing methods that discover knowledge from data originating from educational environments.

Educational Data Mining uses many techniques such as Decision Trees, K- Nearest neighbour, Naive Bayes, Neural Networks and many others. The efficiency of various decision tree algorithms can be examined based on their accuracy and time taken to derive the tree.

When data is mined in educational environment it is called educational data mining. Educational data mining is the process of transforming raw data compiled by education systems in useful information that could be used to take informed decisions and answer research questions [14]. The educational data mining examines the unique ways of applying data mining methods to solve educational problems. When data mining is applied in educational system some classification algorithms like decision tree, neural network, genetic algorithm etc. is applied on educational data so the desired out comes can be predict.

Today, one of the biggest challenges that educational institutions face is the explosive growth of educational data and to use this data to improve the quality of managerial decisions. Data mining techniques are analytical tool around that can be used to extract meaningful knowledge from these large data sets. To face these challenges different systems are use such as ERP, Data wear housing etc.

Decision making of classroom processes involves observing a student's conduct, analysing historical data and estimating the effectiveness of pedagogical strategies. However when students work in electronic environments, this informal supervising is not possible; educators must look for other ways to attain this information. So the solution we are proposing will help to predict all those out comes.

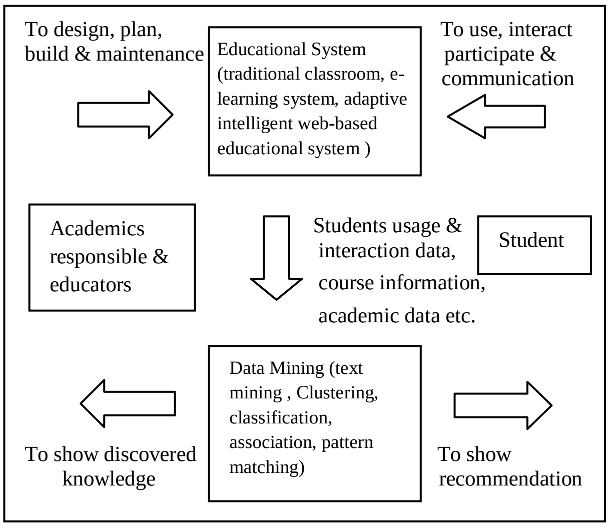
To design, plan, build & maintenance
Educational System (traditional classroom, e-learning system, adaptive intelligent web-based educational system)
To use, interact participate & communication
Academics responsible & educators
Students usage & interaction data, course information, academic data etc.
Student
Data Mining (text mining , Clustering, classification, association, pattern matching)
To show discovered knowledge
To show recommendation

Fig1: The cycle of applying data mining in educational systems

Employing data mining in this manner can assist researchers and practitioners to discover new patterns and trends within large amount of educational data.

There are several major Data Mining techniques have been developing and using in data mining projects recently including clustering, association, prediction, sequential patterns, classification [14]. Here classification technique is used to develop an adaptive learning system for educational environment.

1.1 Classification

Classification is a definitive data mining technique based on machine learning. Classification is used to assort each particular in a set of data into one of predefined set of classes or groups. Classification method makes use of numerical techniques such as decision trees, neural network, linear programming and statistics.

1.2 Decision Tree

Decision tree is tree data structures that represent sets of decisions at foliage nodes. This data structure gives a set of rules for the classification of a training dataset. Decision tree is one of the most frequently used data mining approach because of its transparency. In decision tree technique, the root of the decision tree is a simple question or condition that has multiple answers. Each answer then leadings to a set of questions or conditions that help us to determine the data so that we can make the final decision based on it. Final result is a decision tree in which each branch represents a possible scenario of decision and its outcome.
To build decision tree, row to column and column to row relationships are established. Then all possible outcome instances are tested to check whether they are falling under the same class or not. If all the cases are falling under the same class, the node is delineated with single class name, differently choose the splitting attribute to classify the instances.

The three broadly used decision tree learning algorithms are: ID3, and C4.5, CART.

1.1.1 CART- CART stands for Classification and Regression trees, introduced by Breiman. It is based on Hunt's algorithm. CART addresses both continuous and categorical attributes to build a decision tree. CART handles missing values. Gini Indexing is used in CART as an attribute selection measure to build a decision tree. Dissimilar to ID3 and C4.5 algorithms, CART produces binary splits. Therefore, it produces binary trees. Gini Index measurement does not use probabilistic assumptions like ID3, C4.5. CART uses cost complexness pruning to remove the unreliable branches from the decision tree to improve the accuracy.

1.1.2 ID3 (iterative dichotomiser) - This is a decision tree algorithm introduced in 1986 by Quinlan Ross. ID3 is based on Hunts algorithm. The tree can be built in two stages. The two stages are tree building and pruning. Basic idea of ID3 Algorithm is to construct the decision tree by applying a top-down, greedy search through the given sets to test each attribute at every tree node.ID3 uses information gain measure to select the splitting attribute. It merely accepts categorical attributes in building a tree model. It does not give exact outcome when there is noise. To dispatch the noise pre-processing technique has to be used. Continuous attributes can be handled using the ID3 algorithm by discrediting 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 by default it can be applied after building data model.

1.1.3 C4.5 and C5.0 – C4.5 and C5.0 both algorithms are successor of ID3, developed by Quinlan Ross. It is based on Hunt's algorithm. C4.5 handles both continuous and categorical attributes to construct a decision tree. In order to address continuous attributes, C4.5 separates 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. Both C4.5 and C5.0 also handles missing attribute values. C4.5 usages Gain Ratio as an attribute selection measure to develop a decision tree. It withdraws the biasness of information gain when there are many outcome values of an attribute.

At first, calculate the gain ratio of each attribute. Gain ratio of the root node will be maximal. C4.5 uses pessimistic pruning to remove unnecessary branches in the decision tree to improve the accuracy of classification. The main attribute of data mining is that it subsumes Knowledge Discovery (KD) is a nontrivial process of identifying legal, novel, possibly useful and ultimately understandable patterns in data processes, thereby contributing to predicting trends of outcomes by profiling performance attributes that supports effective decisions making.

C4.5 selects the test that maximizes gain ratio value. The difference between ID3 and C4.5 algorithm is that C4.5 algorithm uses multi-way splits, whereas ID3 uses binary splits. In order to reduce the size of the decision tree, C4.5 uses post-pruning technique; whereas an optimizer combines the generated rules to eliminate redundancies. The improved version of C4.5 is C5.0, which includes cross-validation and boosting capabilities.

Both C4.5 and C5.0 can produce classifiers expressed either as decision trees or rule sets. In many applications, rule sets are preferred because they are simpler and easier to understand than decision trees, but C4.5's rule set methods are slow and memory- hungrier. C5.0 embodies new algorithms for generating rule sets, and the improvement is substantial. In C4.5, all errors are treated as adequate, but in practical applications some classification errors are more serious than others. C5.0 allows a separate cost to be determined for each predicted/actual class pair; if this option is used, C5.0 then builds classifiers to minimize expected miss classification costs rather than error rates.

2. BACKRGOUND

Decision trees are trees that classify instances by sorting them based on characteristic measures. Each node in a decision tree presents a feature in an instance to be classified, and each branch demonstrates a value that the node can simulate. Instances are classified starting at the root node and sorted based on their feature values [15]. Decision tree rules provide model transparency so that a user can understand the basis of the model's previsions, and therefore, be comfortable working on them and explaining them to others.

Comparative analysis on C4.5, C5.0 and ID3 is done in the below section. C4.5 is the successor algorithm of ID3 and C5.0 is the successor algorithm of C4.5. C5.0 algorithm has many features like:

- C5.0 algorithm can respond on noise and missing data.
- C5.0 provides boosting.
- A prominent decision tree may be difficult to read and comprehend.
- C5.0 provides the option of viewing the large decision tree as a set of rules which is easy to understand.
- Over fitting is figured out by the C5.0 and abbreviate error pruning technique.
- C5.0 can also predict which attributes are relevant in classification and which are not. This technique, known as Winnowing is especially useful while dealing with high dimensional datasets.

2.1 Algorithm C5.0-

```
Input: Example, Target attribute, Attribute
Output: decision tree algorithm:
1. Check for the base class
2. Construct a decision tree using training data
3. Find the attribute with the highest information gain
4. For each tied, apply the decision tree to determine
   its class since the application of a given tuple to a
   decision tree is relatively straightforward.
```

Base cases are the following for the algorithm C4.5 and C5.0:

- All the instances from the training set belong to the same class (a tree leaf labeled with that class is returned).
- The training set is empty (renders a tree leaf called failure).
- Attribute list is empty (yields a leaf labeled with the most frequent class or the disjunction of all the classes).

Output: Decision tree which properly classifies the data.

2.2 Comparison with Current Algorithms

2.2.1 C4.5 Improvements from ID3 algorithm

- Handling both discrete attributes and continuous. In order to treat continuous attributes, C4.5 produces a threshold and splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values - C4.5 tolerates attribute values to be labeled as '?' for missing. Lacking attribute values are simply not used in gain and entropy calculations.
- Treating attributes with differing costs.
- Pruning trees after creation - C4.5 goes backward through the tree once it's been produced and attempts to remove branches that do not help by replacing them with leaf nodes.

2.2.2 C5.0 advances from C4.5 algorithm

- Speed - C5.0 is faster than C4.5 (various orders of magnitude)
- Memory usage - C5.0 is more memory efficient than C4.5. C5.0 commonly uses an order of magnitude less memory than C4.5 during rule set construction.
- Accuracy - the C5.0 rules sets have noticeably lower error rates on unseen cases. Sometimes C4.5 and C5.0 rule sets have the same predictive accuracy, but C5.0 rule set is smaller.
- Smaller decision trees - C5.0 gets similar outcomes to C4.5 with substantially smaller decision trees.
- Support for boosting - Boosting improves the trees and gives them more accuracy.
- Weighting - C5.0 allows weighting different attributes and misclassification types.

2.3 Problem of Current System

Consequences in data mining with decision trees- There are some issues in acquiring decision trees which include:

- Determining how profoundly to produce the decision tree
- Handling continuous attributes
- Taking an appropriate attribute selection measure
- Treating training data with missing attribute values
- Handing attributes with disagreeing costs
- Improve computational efficiency

2.4 Solution via Proposed System

Avoiding over-fitting the data- In case of noisy data or in case of too small training set, is really difficult to classify. In either of these cases, this simple algorithm can produce trees that *over-fit* the training examples.

There are several approaches to avoiding over-fitting in decision tree acquisition. These can be grouped into two categories:

- Approaches that stop growing the tree earliest, before it arrive at the point where it absolutely classifies the training data.
- Approaches that allow the tree to over-fit the data and then post prune the tree. First approach is more direct compare to post pruning. Post pruning is the acceptable approach as in case of the first approach it is difficult to know that when to stop. It is still not known that how to determine the correct tree size.

Related Work:

A survey of top-down decision trees induction algorithms. It has been shown that most algorithms fit into a simple algorithmic framework whereas the differences concentrate on the splitting criteria, stopping criteria and the way trees are pruned [1].

An approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. They design, implement, and evaluate a series of pattern classifiers and compare their performance on an online course dataset. Intensifying multiple classifiers leads to a significant improvement in classification operation. Moreover, by learning an appropriate weighting of the features used via a genetic algorithm (GA), then further improve prediction accuracy is performed [2].

The equally important issue of how to handle the test costs associated with querying the missing values in a test case. In this paper proposed work shows a test-cost-sensitive earning framework for designing classifiers that minimize the sum of the misclassification cost and the trial costs. In the model of TCSL, imputes are selected for testing intelligently to get both the sequential test strategy and the batch test scheme. Experiments demonstrate that the method surmounts other competing algorithms and observe that the decision tree algorithm which is used here is a kind of simple one in that it simply follows the tree sequentially to obtain a next missing value [3].

Research is following different methods of data mining to construct imputative models in accordance with different types of missing data. When the missing data is uninterrupted, Neural Networks and regression models are used to build imputative models. For the categorical missing data, C5.0, CART, the logistic regression model and neural network are employed to construct imputative frameworks. The results demonstrated that the regression model was found to provide the best estimate of continuous missing data; but for categorical missing data, C5.0 model established the best method [4].

An implementation of a J48 algorithm analysis tool on data collected from surveys on different specialization students of faculty, with the intention of differentiating and predicting their choice in continuing their education with post university studies (master degree, Ph.D. considers) with decision trees [5].

A novel framework that aims to improve the accuracy of the existing imputation methods is proposed. The new framework consists of three modules, namely mean pre-imputation, confidence intervals, and boosting, and can be applied to many of the existing imputation methods, including data driven, model based, and ML based. To demonstrate the advantages of the proposed framework, it was used with two imputations methods: NB ML-based imputation method and HD data-driven imputation method. The results show that a significant improvement of imputation accuracy can be achieved by applying the proposed framework and that the accuracy of the framework-based methods was, on average, the highest among the considered methods [6].

A simple and effective method for dealing with missing data in decision trees applied for categorization. This approach is called missingness incorporated in attributes or assigns. Hence research worker put forwards the procedure of missingness comprised within attributes as a conceptually and computationally mere method for dealing with missing data in decision trees when classification is the goal. It is intimately colligated to addressing “missing as class” intrinsically, deriving that approached path for usage with continuous as well as categorical variables. [7].

An improved ID3 algorithm to overcome deficiency of general ID3 algorithm which tends to take attributes with many values. The presented algorithm makes the constructed decision tree more clear and understandable [8].

A methodological analysis of discovering social action patterns in collaborative learning activities for use in improving activity design, and in particular for restructuring existing designs involving face-to-face social actions to enhance their social dynamics and thus better ensure the achievement of a specified aim [9].

A system to identify and classify Telugu (a south Indian language) characters extracted from the palm leaves, habituating Decision Tree approach. The decision tree is formulated by applying SEE5 algorithm, which is advancement from the predecessor C4.5 and ID3 algorithm [10].

Effectively apply Euclidean Distance for selecting a subset of robust features using smaller storage space and getting higher Intrusion detection functioning. On the evaluation stage, three different test data sets are used to evaluate the performance of proposed approach on C5.0 classifier. Observational results demonstrate that the proposed approach based on the Euclidean Distance can improve the performance of a true positive intrusion detection rate especially for detecting known attack patterns [11].

Work extinction of the model of decision-tree classification to accommodate data tuples having numerical attributes with uncertainty described by absolute pdf’s. Here altered classical decision tree progressing algorithms (based on the framework of C4.5 [3]) is proposed to build decision trees for classifying such data [12].

Comparison of different data imputation an approach used in filling missing data and proposes a combined approach to estimate accurately missing attribute values in a patient database. Introduced study suggests a more robust technique that is likely to supply a value closer to the one that is missing for effective classification and diagnosis [13].

3. PROBLEM IDENTIFICATION AND PROPOSED WORK

Large amount of data is generated daily in the routine of educational institutions and other relevant research labs. These data is random and unstructured by nature, this data is related to the students’ performance, activity and there nature or behaviour. Using the data mining techniques these data can be improved and used for different kind of analysis in the domain of educational growth and advancement.

Classification performance can degrade if data contain missing attribute values. Many methods consider missing information in a simple manner, such as substituting missing values with the global or class-conditional mean/mode. Throughout the different procedures and due to gap between communicating facts some data is missing or incomplete. The incomplete information generates the ambiguity during the data analysis of any kind of system. Missing values are a common experience in real-world data sets. This situation can complicate both induction (a training set where some of its values are missing) as well as classification (a new instance that miss certain values).

There is requirement to handle the different missing attributes of data which is not found during the building of data formats, additionally after filling the problem of missing attribute values recover the decision making facts by which system generates the future prediction of the students behaviour, growth and gap of the performance.

Reason for missing attribute values can be that the attribute value was not placed into the table because it was forgotten or it was placed into the table but later on was mistakenly erased. Sometimes a respondent refuse to answer a question. Such a value, that matters but that is missing, will be called lost.

The problem of missing attribute values is as important for data mining as it is for statistical reasoning. In both disciplines there are methods to deal with missing attribute values. In general, methods to handle missing attribute values belong either to sequential methods (called also pre-processing methods) or to parallel methods (methods in which missing attribute values are taken into account during the main process of acquiring knowledge).

Sequential methods include techniques based on deleting cases with missing attribute values, substituting a missing attribute value by the most common value of that attribute, allotting all potential values to the missing attribute value, substituting a missing attribute value by the mean for numeral attributes, attributing to a missing attribute value the corresponding value taken from the closest case, or replacing a missing attribute value by a new value, computed from a new data set, considering the original attribute as a decision.

The second group of methods to handle missing attribute values, in which missing attribute values are taken into account during the main process of acquiring knowledge is represented, for example, by a modification of the LEM2 (Learning from instance Module, version 2) rule inductance algorithm in which rules are induced form the original data set, with missing attribute values considered to be "do not care" conditions or lost values. C4.5 approach to missing attribute values is another example of a method from this group. C4.5 induces a decision tree during tree propagation, dissevering cases with missing attribute values into fractions and adding these fractions to new case subsets.

Missing values make it difficult for analysts to realize data analysis. Three types of problems are commonly related with missing values:

- Loss of efficiency;
- Complications in handling and analysing the data;
- Bias resulting from differences between missing and complete data.

Although some methods of data analysis can cope with missing values on their own, many others require complete databases. Standard statistical software works only with complete data or uses very generic methods for filling in missing values.

Thus to overcome the discussed problem in the section of problem domain required to handle the missing attributes in the suggested data base and predict them to complete the uncompleted information.

Additionally in this project we provide the following solutions for data analysis based model for the educational data mining.

- Handle missing values of the supplied data base.
- Prepare data model for student future analysis.
- Prepare data model for student behaviour analysis.
- Prepare data model for student performance analysis.

The existing methods for dealing with missing values can be divided into two main categories:

- Missing data removal.
- Missing data imputation.

The removal of missing values is concerned with discarding the records with missing values or removing attributes that have missing entries. The latter can be applied only when the removed attributes are not needed to perform data analysis. Both removals of records and attributes result in decreasing the information content of the data. They are practical only when a database contains a small amount of missing data and when the ensuing analysis of the remaining complete records will not be biased by the removal. Another method belonging to the same category proposes substituting missing values for each attribute with an additional category. Although this method provides a simple and easy-to-implement solution, its usage results in substantial problems occurring during the subsequent analysis of the resulting data.

4. SYSTEM ARCHITECTURE

A decision tree is a tree structure which classifies an input sample into one of its possible classes. Decision trees are used to extract knowledge by Inferring decision making rules from the huge amount of available information. Decision tree is a useful tool in classification. A decision tree classifier has a simple form which can be compactly stored and that efficiently classifies new information. Decision tree classifiers can perform reflexive feature selection and complexity reduction, while the tree structure gives easily understandable and interpretable information regarding the predictive or generalization ability of the data. A decision tree recursively divides a data set into smaller subdivisions on the basis of tests applied to one or more features at each node of the tree.

Missing attribute values are a common occurrence in data, either through errors made when the values were recorded or because they were judged irrelevant to the particular case. Such lacunae affect both the way that a decision tree is constructed and its use to classify a new case. There are two general approaches to deal with the problem of missing values: They could be ignored (removed) or imputed (filled in) with new values. The first solution is applicable only when a small amount of data is missing. Since in many cases databases contain relatively large amount of missing data, it is more constructive and practically viable to consider imputation.

Missing values impact decision tree construction in three different ways:

- Affects how impurity measures are calculated.
- Affects how to administer instance with missing value to child nodes.
- How to construct a decision tree when some records have missing values?

The suggested data model works with the below given architecture in fig.2.

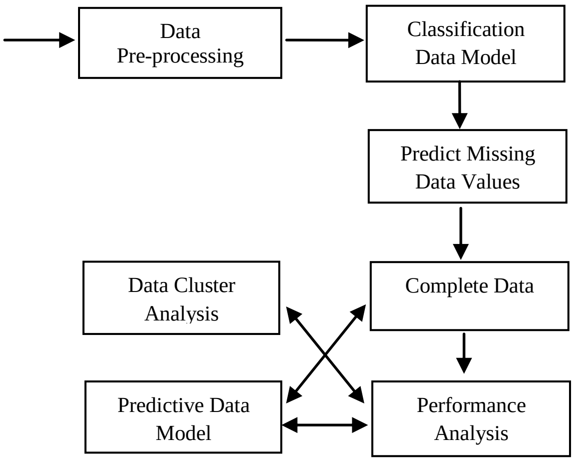

Fig2: Architecture of Recommender System

The above given system accepts the students' data base as input apply data pre-processing on the database and cleans the unwanted symbols and objects from the data so the pre-processed data can be ready for the data mining.

One of the data mining techniques is classification is applied on the pre-processed data and then required to make the data model based on the available data from database. As we know that some of the data in student data base is missing so to handle these values we predict the missing values in the data.

The incomplete database with missing value attributes is compared with the complete database, so that missing value can be predict. The complete data base is again used for cluster analysis and as predictive model. These two model used to provide the all our desired data model as outcome.

5. CONCLUSION AND FUTURE WORK

Data mining is a substantial tool for facilitating organizations to enhance decision making and examining new patterns and relationships among large amount of data.

Main core of this work is to review role of data mining techniques in education system. Educational Data Mining has been introduced as an upcoming research area, thus a tool is developed that use a default algorithm for each task and parameter-free DM algorithms to simplify the configuration and execution for non-expert users.

An approach dealing with missing attribute values in both training data and test data has been proposed. Experiments have been done by comparing it with the basic method that uses mode to fill missing values, an attribute trees method that builds a decision tree for each attribute, and a previously proposed ordered attribute trees method. We come up with the observation that our method performs consistently better in different domains of datasets.

The decision tree is the only learning scheme that is applied in the presentation of the method and in the tests, because of its nature to capture the relations between attributes. Further experimentation on data which contain both nominal and numeric attributes using other learners is expected.

Our approach is mainly focusing on the estimation of missing values in training data, and the missing values in test data are simply predicted by the induced training data. We are interested in combining our approach in induction Phase and other approaches in prediction phase, such as dynamic path generation, so as to effectively handle missing values in training data and test data respectively. After handling the missing value prepare a data model for student future analysis as well as prepare data model for student behaviour analysis so students' performance can be increased.

6. REFERENCES

[1] Lior Rokach and Oded Maimon, "Top-Down Induction of Decision Trees Classifiers – A Survey", IEEE transactions on systems, man and cybernetics: part c, VOL. 1, NO. 11, pp. 1-12, 2002.

[2] Behrouz Minaei-Bidgoli , Deborah A. Kashy , Gerd Kortemeyer , William F. Punch, "predicting student performance: an application of data mining methods with the educational web-based system lon-capa", ASEE/IEEE Frontiers in Education Conference, pp. 1-6, 2003.

[3] Qiang Yang, Senior Member, Charles Ling, Xiaoyong Chai, and Rong Pan, "Test-Cost Sensitive Classification on Data with Missing Values", IEEE, VOL. 18, NO. 5, pp.626-638, 2006.

[4] Ruey-Ling Yeh , Ching Liu, Ben-Chang Shia, Yu-Ting Cheng, Ya-Fang Huwang, " Imputing manufacturing material in data mining", Springer Science+Business Media, pp. 109–118, 2007.

[5] Vasile Paul Bre_felean, "Analysis and Predictions on Students' Behavior Using Decision Trees in Weka Environment", ITI Int. Conf. on Information Technology Interfaces, pp. 51-56, 2007, Cavtat, Croatia.

[6] Alireza Farhangfar, Lukasz A. Kurgan, Member and Witold Pedrycz, Fellow, "A Novel Framework for Imputation of Missing Values in Databases", IEEE, VOL. 37, NO. 5, pp. 692-709, 2007.

[7] Twala, B. E. T. H.; Jones, M. C. and Hand, D. J. "Good methods for coping with missing data in decision trees" Pattern Recognition Letters, 29(7), pp. 950–956, 2008.

[8] Chen Jin, Luo De-lin and Mu Fen-xiang, "An Improved ID3 Decision Tree Algorithm", 4th International Conference on Computer Science & Education, pp. 127-130, 2009, China.

[9] Maria Francisca Capponi, Miguel Nussbaum, Guillermo Marshall and Maria Ester Lagos, "Pattern Discovery for the Design of Face-to-Face Computer-Supported Collaborative Learning Activities", Educational Technology & Society, pp. 40–52, 2010.

[10] Panyam Narahari Sastry, Rama krishnan Krishnan and Bhagavatula Venkata Sanker Ram, "Classification and identification of telugu handwritten characters extracted from palm leaves using decision tree approach", ARPN Journal of Engineering and Applied Sciences, VOL. 5, NO. 3, pp. 23-32, 2010.

[11] Anirut Suebsing, Nualsawat Hiransakolwong, "Euclidean-based Feature Selection for Network Intrusion Detection", International Conference on Machine Learning and Computing IPCSIT vol.3, pp. 222-229, 2011.

[12] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho and Sau Dan Lee, "Decision Trees for Uncertain Data", IEEE, VOL.23, NO. 1, pp. 1-15, 2011.

[13] T.R.Sivapriya, A.R.Nadira Banu Kamal and V.Thavavel, " Imputation And Classification Of Missing Data Using Least Square Support Vector Machines – A New Approach In Dementia Diagnosis", IJARAI, Vol. 1, No. 4, pp.29-34,2012

[14] Barahate Sachin R. and Shelake Vijay M, "A Survey and Future Vision of Data mining in Educational Field", Second International Conference on Advanced Computing & Communication Technologies, pp. 96-100, 2012.

[15] A. S. Galathiya, A. P. Ganatra and C. K. Bhensdadia, "Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning", International Journal of Computer Science and Information Technologies, Vol. 3 ,pp. 3427-3431 , 2012, India.

[16] Mingyu Feng, Marie Bienkowski and Barbara Means, "Enhancing Teaching and Learning through Educational Data Mining and Learning Analytics: An Issue Brief", U.S. Department of Education, 2012.