Engineering Universe for Scientific Research and Management
(International Journal)
ISSN(ONLINE) : 2319-3069
Vol. 7 Issue 2
February 2015

Decision Tree C4.5 algorithm and its enhanced approach for Educational Data Mining

Preeti Patidar[1], Jitendra Dangra[2], M.K. Rawar[3]
Computer Science dept. LNCT Indore, University RGPV Bhopal, India[1]
Computer Science dept. LNCT Indore, University RGPV Bhopal, India[2]
Computer Science dept. LNCT Indore, University RGPV Bhopal, India[3]
preeti.ppatidar@gmail.com[1], jitendra.dangra@gmail.com[2], ermkrawat@gmail.com[3]

***Abstract*- Data mining tools for educational research issues are prominently developed and used in many countries. Decision Tree is the most widely applied supervised classification data mining technique. The learning and classification steps of decision tree induction are simple and fast and it can be applied to any domain. For this research work student qualitative data has been taken from educational data mining and the performance analysis of the decision tree algorithm C4.5 and proposed algorithm are compared. The classification accuracy of proposed algorithm is higher when compared to C4.5. However the difference in classification accuracy between the decision tree algorithms is not considerably higher.**

This paper describes the use of data mining techniques to improve the efficiency of academic performance in the educational institutions. In this work a real-world experiment is conducted on real-time data. This method helps to identify the students who need to obtain academic records at which grade and learn particular skill to improving their placement possibilities. Prediction for specific student according to skills known by them can be performed. In this study C4.5 classifier and proposed algorithm with ensemble techniques such as boosting and bagging have been considered for the comparison of performance of both the algorithms according to parameters accuracy, build time, error rate, memory used and search time for the classification of datasets.

Keywords: Data Mining (DM), Educational Data Mining (EDM), Classification Model, Decision Tree Algorithm (DT), C4.5 classifier, CART, Ensemble learning, Prediction.

I. Introduction

Nowadays, information and data are stored everywhere, mainly on the Internet. To serve us, information had to be transformed into the form, which people can understand. This transformation represents a large space for various machine learning algorithms, mainly classification. The quality of the transformation heavily depends on the precision of classification algorithms in use. The precision of classification depends on many aspects. Two of most important aspects are the selection of a classification algorithm for a given task and the selection of a training set. Here focus is on experiments with training set samples, to improve the precision of classification results. At present, two approaches are there, the first approach is based on an idea of making various samples of the training set. By a selected machine learning algorithm a classifier is generated for each of these training set samples. In this manner, for *k* variations of the training set, *k* particular classifiers generated. The result will be given as a combination of individual particular classifier; this method is called bagging [1]. Another similar method called Boosting [7] does experiments over training sets as well. In this method weights of training examples are used. Higher weights are imputed to incorrectly classified examples i.e. the importance of these examples is emphasised. After the weights are updated, a new (base) classifier is generated. A final classifier is calculated as a combination of base classifiers. The presented paper focuses on the bagging method in combination with Decision trees in the role of base classifiers.

Data Mining can be used in educational field to enhance our understanding of learning process to focus on the identification, extraction and evaluation of variables related to the learning process of students. Data mining [6] is the process of analyzing data from various perspectives and summarizing it into useful, meaningful with all relative information. There are many DM algorithms and tools that have been developed for feature selection, clustering, rule framing and classification. DM tasks can be divided into 2 types: Descriptive – to discover general interesting patters in the data and Predictive – to predict the behavior of the model on available data.

The inspiration for this work came from the study of many research work done in the area of educational data mining. Many institutions abroad have developed student analysis system and are using it. India has more number of educational institutions but very few use student analysis systems. Mining in educational environment is called Educational Data Mining. Educational data mining is an interesting research area which extracts required, previously unknown patterns from educational database for better understanding, to improve educational performance and assessment of the student learning process. Various algorithms and techniques such as Clustering, Regression, Neural Networks, Classification, Association Rules, Nearest Neighbor and Genetic Algorithm etc., are used for knowledge discovery from databases [18]

1.1 DECISION TREE

A decision tree is a tree like structure, where rectangles are used to denote internal node and ovals are used to denote leaf nodes. All internal nodes can have two or more child nodes. All internal nodes contain splits, which test the value of an expression of the attributes. Connections from an internal node to its children are labeled with distinct outcomes of the test and each leaf node has a class label associated with it. Decision tree are commonly used for acquiring information for the purpose of decision -making. Decision tree starts with a root node on which it is for users to take actions, from this node each node is spitted accordingly for decision tree learning algorithm recursively. The concluding result is a decision tree in which each branch represents a possible scenario of decision and its outcome. Two operations are there in decision tree as follows:

Training : The records of students with known result is trained as attributes and values which is used for generating the decision tree based on the information gain of the attributes.

Testing: The unknown records of students are tested with the decision tree developed from the trained data for determining the result.

1.1.1 C4.5

This algorithm is a successor to ID3 developed by Quinlan Ross [1]. It is based on Hunt's algorithm like ID3. C4.5 handles both categorical and continuous attributes to build a decision tree. C4.5 splits the attribute values into two partitions to handle continuous attributes based on the selected threshold such that all the values above the threshold as one child and the remaining as another child. Missing attribute values can be handled using C4.5. To build a decision tree, C4.5 uses Gain Ratio as an attribute selection measure which removes the biasness of information gain when there are many outcome values of an attribute. Initially, calculate the gain ratio of each attribute; the root node will be the attribute whose gain ratio is maximum. Pessimistic pruning is used in C4.5 to remove unnecessary branches in the decision tree to improve the accuracy of classification.

Classification Tree based on C4.5 uses the training samples to generate the model. The data classification process can be described as follows.

- learning using training data
- Classification using test data

C4.5 uses information gain ratio which is an impurity-based criterion that employs the entropy measure as an impurity measure.

Definition 1 (Information Entropy): Given a training set *T*, the target attribute takes on *n* different values, and then the entropy of *T* is defined as:

$$Entropy(T) = -\sum_{i=1}^{n} Pi \log_2 Pi$$

where Pi is the probability of *T* belonging to class *i*.

Definition 2 (Information Gain): The information gain of an attribute *A*, relative to the collection of examples *T* is:

$$InfoGain = Entropy(T) - Entropy(A,T)$$

$$InfoGain = Entropy(T) - \sum_{i=1}^{n} \frac{|Ti|}{|T|} Entropy(Ti)$$

where Si is the partition of *S* induced by the value of attribute *A*.

Definition 3 (Gain Ratio): The gain ratio "normalizes" the information gain as follows:

$$GainRatio(A,T) = \frac{InfoGain(A,T)}{SplitEntropy(A,T)}$$

$$GainRatio(A,T) = \frac{InfoGain(A,T)}{-\sum_{i=1}^{n} {}^{|Ti|}/_{|T|} \log_2 {}^{|Ti|}/_{|T|}}$$

1.1.2 CART:

CART stands for Classification And Regression Trees introduced by Breiman, it is also based on Hunt's algorithm. It handles both continuous and categorical attributes to build a decision tree. It handles missing values. CART uses Gini Index as an attribute selection measure to build a decision tree. Dissimilar to ID3 and C4.5 algorithms, CART produces binary trees using binary splits. Gini Index measure does not use probabilistic assumptions like C4.5. CART uses cost complexity pruning to remove the unreliable branches from the decision tree to improve the accuracy. Similar to CART, C4.5 can also deal with both nominal and continuous variables. CART uses Gini index which is an impurity-based criterion that measures the divergences among the probability distributions of target attribute's values.

Definition4 (Gini Index): Given a training set *T* and the target attribute takes on *n* different values, then the Gini index of T is defined as

$$Gini(T) = 1 - \sum_{i=1}^{n} Pi^2$$

Where *Pi* is the probability of *T* belonging to class *i.*

Definition 5 (Gini Gain): Gini Gain is the evaluation criterion for selecting the attribute A which is defined as

$$GiniGain(A,T) = Gain(T) - Gini(A,T)$$

$$GiniGain(A,T) = Gain(T) - \sum_{i=1}^{n} \frac{|Ti|}{|T|} Gini(Ti)$$

Where Ti is the partition of T induced by the value of attribute A. CART algorithm can deal with the case of features with nominal variables as well as continuous ranges. Pruning a tree is the action to replace a whole sub-tree by a leaf node. CART uses a pruning technique called "minimal cost-complexity pruning" which assuming that the bias in the re-substitution error of a tree increases linearly with the number of leaves. Formally, given a tree *E* and a real number $\alpha>0$ which is called the "complexity parameter", then the cost-complexity risk of *E* with respect to α is:

$$R_\alpha(E) = R(E) + \alpha.|E|$$

Where |E| is the number of terminal nodes (i.e. leaves) and R *(E)* is the re-substitution risk estimate of *E*.

1.1.3 Ensemble of Classifiers:

In this work, we focus on ensembles of decision trees classifiers and compare them with the C4.5 classifiers. Decision tree ensembles tend to produce very accurate results on a variety of datasets due to the reduction in both bias and variance component of the generalization error of the base classifier [5]. May be that the researchers regard only a single decision tree such as C4.5 or CART (due to interpretability), which is not strong enough to compare to the classification results.

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a vote of their predictions. Bayesian averaging is the original ensemble method. Merely more recent algorithms include error correcting output coding Boosting and Bagging. The approach of Ensemble systems is to improve the confidence with which we are making right decision through a process in which various opinions are weighed and combined to reach a final decision. We propose a meta-algorithm to get more accurate models.

Some of the reasons for using Ensemble Based Systems [15]:

- Too much or too little data: The amount of data can be too large to be analyzed effectively by a single classifier. Resampling techniques can be used to overlap random subsets of inadequate training data and each subset can be used to train a different classifier.
- Statistical Reasons: To reduce the risk of selecting a poorly performing classifier combination of the outputs of several classifiers is formed by applying average.
- Confidence Estimation: A properly trained ensemble decision is usually correct if its confidence is high and usually incorrect if its confidence is low. By applying this approach the ensemble decisions can be used to estimate the posterior probabilities of the classification decisions.
- Divide and Conquer: A particular classifier is unable to solve certain problems. The decision boundary for different classes may be too complex. In such cases, complex decision boundary can be estimated by combing different classifiers appropriately.
- Data Fusion: A single classifier is not adequate to learn information contained in data sets with heterogeneous features (i.e. data obtained from various sources and the nature of features is different). Applications in which data from different sources are combined to make a more informed decisions are

referred as Data Fusion applications and ensemble based approaches are most suitable for such applications.

1.1.4 Bagging

Bagging [23] is a name derived from bootstrap aggregation. It is the first effective method of ensemble learning and is one of the simplest methods of arching. The meta-algorithm which is a peculiar case of model averaging was originally designed for classification and is usually applied to DT models, but it can be used with any type of model for classification or regression. The method usages multiple versions of training set by using the bootstrap i.e. replacement using sampling. Each of these data sets is used to train a different model. The outputs of the models are aggregated by averaging (in the case of regression) or voting (in the case of classification) to create a single output.

1.1.5 Boosting (Including AdaBoost*)*

AdaBoost stands for "**ada**ptive **boosti**ng", it decreases the weights of correctly classified examples and increases the ones of those classified incorrectly. Boosting is a meta-algorithm which can be viewed as a model averaging method. It is the most widely used ensemble method and one of the most powerful learning ideas introduced in the last two decades. Originally designed for classification, but can also be profitably extended to regression. One first creates a 'weak' classifier, it suffices that its accuracy on the training set is slightly better than random guessing. Iteratively a succession of models is build, each one being trained on a data set in which points misclassified (or, with regression, those poorly predicted) by the previous model are given more weight. Finally, according to their successor all of the successive models are weighted and then the outputs are combined using averaging (for regression) or voting (for classification), thus creating a final model. The original boosting algorithm combined weak learners to generate a strong learner.

II. Background

Data miming consists of a set of techniques that can be used to extract relevant and interesting knowledge from data. Data mining has several tasks such as prediction, association rule mining, clustering and classification. Classification techniques are supervised learning techniques that classify data item into predefined class label. To build classification models from an input data set it is most useful techniques in data mining. The used classification techniques commonly build models that are used to predict future data trends. The ability to perform student's performance prediction is very important in educational environments [12].

Decision tree can be used to visually and explicitly represent decisions and decision making. In DM, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. Decision trees used in data mining are of two main types: Classification tree analysis is when the predicted outcome is the class to which the data belongs and Regression tree analysis is when the predicted outcome can be considered a real number.

Data format:

The data is raw in nature and found in unformatted way. But to work with the data model required to format data first, this process also called the data pre-processing. Data pre-processing includes the different phases to achieve a well formatted and arranged data. Moreover, after processing the data can be categorized into three main parts.

- Data set with only numerical values
- Data set with nominal values
- Data set with both nominal and numerical values.

For the experimental purpose we use the data. Manually generated ARFF data format is used in the proposed work and also dataset that is available online is also used for experiments of machine learning.

ARFF also abbreviated as attribute relationship file format. The Header of the ARFF file contains the name of the relation, a list of the attributes and their types.

2.1 Over fitting:

As we know in constructing decision trees we use training data set. We do this because we want to capture some general underlying functions or trends in that data, usually to be used in prediction. As we are not interested in capturing all the exact nuances and extremities of the training data. It is normally the result of errors or peculiarities that we are not likely to come across again. It is important that we can use our DT model to predict or generalize over future

instances that we might obtain. Over fitting occurs when our decision tree characterizes too much detail, or noise in our training data, this can be stated as: Two hypotheses, H1 and H2 over some data exist with the following relationship:

Training set errors (H1) < Training set errors (H2)
AND
Testing set errors (H1) > Testing set errors (H2)

As well as noise in the training data, it can happen when we don't have much trained data and are trying to extrapolate an underlying hypothesis from it. We want our decision tree to generalize well, but unfortunately if we build a decision tree until all the training data has been classified perfectly and all leaf nodes are reached, then chances are that it we'll have a lot of misclassifications when we try and use it. Methods that we can use to avoid over fitting such as "pruning"

2.2 Pruning:

Over fitting is a significant practical difficulty for decision tree models and many other predictive models. Over fitting happens when the learning algorithm continues to develop hypotheses that reduce training set error at the cost of an increased test set error. There are several approaches for avoiding over fitting in building decision trees.

- Pre-pruning that stop growing the tree earlier, before it absolutely classifies the training set.
- Post-pruning that allows the tree to perfectly classify the training set, and then perform post prune operation on the tree.

The step of tree pruning is to define a criterion that can be used to determine the correct final tree size using one of the following methods; use a distinct dataset from the training set i.e. validation set, to appraise the effect of post-pruning nodes from the tree. By using the training set build a tree and then apply a statistical test to estimate whether pruning or expanding a particular node is likely to produce an improvement beyond the training set.

2.3 Optimal decision tree construction:

The problem of designing a truly optimal DTC seems to be a very difficult problem. In fact it has been shown by Hyafil and Rivest [12] that the problem of constructing optimal binary trees, optimal in the sense of minimizing the expected number of tests required to classify an unknown sample is an NP-complete problem and thus very unlikely of non-polynomial time complexity. It is supposed that the problem with a general cost function or minimizing the maximum number of tests (instead of average) to classify an unknown sample would also be NP-complete. It is also supposed that no sufficient algorithm exists (on the supposition that P ≠ NP) and thereby supply motivation for finding efficient heuristics for constructing near-optimal decision trees.

For construction of DTC various heuristic methods can roughly be divided into four categories:

- Bottom-Up approaches
- Top-Down approaches
- The Hybrid approach and
- Tree Growing-Pruning approaches.

Heuristic based decision trees, also called rule induction techniques, include classification and regression trees (CART) as well as C4.5. CART handles binary splits best, whereas multiple splits are best taken by C4.5. If a tree has only two-way splits, it is considered a binary tree, otherwise a ternary tree. For most of their applications, decision trees start the split from the root (root node) into leave nodes, but on occasion they reverse the course to move from the leaves back to the root. Figure 2.1 is a graphical rendition of a decision tree (binary). The algorithms differ in the criterion used to drive the splitting. C4.5 relies on measures in the realm of the Information Theorem and CART uses the Gini coefficient (SPSS, 2000). Rule induction is fundamentally a task of reducing the uncertainty (entropy) by assigning data into partitions within the feature space based on information-theoretic approaches.

For improving results of machine learning classification algorithms Bagging is used. In case of classification into two possible classes, a classification algorithm creates a classifier H: D = {-1,1} on the base of a training set of example descriptions (in our case played by a document collection) D. The bagging method creates a sequence for classifiers H_m, m = 1...M in respect to modifications of the training set. A compound classifier is formed by combining these classifiers. The prediction of the compound classifier is given as a weighted combination of individual classifier predictions:

$$H(di) = sign\left(\sum_{M=1}^{M} am\ Hm(di)\right)$$

Experiment is performed using the following bagging algorithm [1] for multiple classification into several classes.

1 Initialization of the training set *D*
2 for *m = 1, ..., M*

- Creation of a new set *Dm* of the same size *D* by random selection of training examples from the set *D* (some of examples

can be selected repeatedly and some may not be selected at all).

- Learning of a particular classifier *Hm*: *Dm* → *R* by a given machine learning algorithm based on the actual training set *Dm*.

3 Compound classifier *H* is created as the aggregation of particular classifiers *Hm: m = 1, ...,M* and an example *di* is classified to the class *cj* in accordance with the number of votes obtained from particular classifiers *Hm.*

$$H(di, cj) = sign\left(\sum_{M=1}^{M} am\ Hm(di, cj)\right)$$

If it is possible to influence the learning procedure performed by the classifier *Hm* directly, classification error can be minimized also by *Hm* while keeping parameters $\alpha\, m$ constant.

III. Proposed Work

Educational data mining concerned with developing methods for exploring the unique types of data that come from the educational domain. The discipline focuses on analyzing educational data to develop models for improving learning experiences and improving institutional effectiveness. The scope of educational data mining includes areas that directly impact students; for example mining course content and the development of recommender systems. Other areas within EDM include analysis of educational processes including course selections, admissions and alumni relations. Moreover, applications of specific DM techniques such as association, web mining, rule mining, classification and multivariate statistics are also key techniques applied to educationally related data. These data mining methods are largely exploratory techniques that can be used for prediction and forecasting of learning and institutional improvement needs, also the techniques can be used for modeling individual differences in students and provide a way to respond to those differences thus improve student learning.

Our empirical studies on student's database have identified two data mining techniques that generate rules with considerable different parameters. Two algorithms C4.5 decision tree classifier and proposed algorithm to predict the result of student is applied on educational data mining.

In the previously published papers we have performed analysis on the student data using many data mining techniques and finally selected C4.5 decision tree algorithm and proposed new algorithm for predicting the performance of students. Unlike the recent research trends that focused on predicting overall grading of students during their studies, this paper orients itself in identifying student's placement levels according to their skills known. It was found that from study that obtained accuracy and error rate figure was better in proposed algorithm than C4.5 Decision tree classifier. There are two levels at which the system functions. At one level they can use various techniques to perform analysis on student data and generate the necessary output for those methods that prove useful. This output is fed into the second level where it is implemented and used for performing prediction on the real data.

3.1 Data Preparation

On existing and real-time data base both C4.5 and proposed algorithm is applied. Existing data-set consist of ARFF file format which is available for experimental purpose. Real-time student's dataset consist of records with different attributes. The academic data was extracted from the student management system of the college. Other details were collected from through questionnaires and than all the attributes are transformed into categorical values as student's final year Grade i.e. VII and VIII semester results (Grade A, B or C) and skills known (Yes or No).

3.2 Prediction

In prediction, the goal is to develop a model which can infer a single aspect of the data (predicted variable) from some combination of other aspects of the data (predictor variables). For a limited data set prediction requires having labels for the output variables, here a label represents some trusted "ground truth" information about the output variable's value in specific cases. In some cases, however, it is important to consider the degree to which these labels may in fact be approximate, or incompletely reliable [22]. Prediction has two key uses within educational data mining. In some cases, prediction methods can be used to study what features of a model are important for prediction, giving information about the underlying construct. It's a common approach in programs of research that attempt to predict student educational outcomes, without predicting intermediate or mediating factors first. In a second type of usage,

prediction methods are used in order to predict what the output value would be in contexts where it is not desirable to directly obtain a label for that construct (e.g. in previously collected repository data, where desired labeled data may not be available, or in contexts where obtaining labels could change the behavior being labeled, such as moulding affective states, where self-report, video, and observational methods all present risks of altering the construct being studied).

In classification, the predicted variable is a categorical or binary variable. Some popular classification methods include DT, logistic regression (for binary predictions), and support vector machines. In regression, the anticipated variable is a continuous variable. Some popular regression methods within EDM include neural Networks, support vector machine and linear regression. For each type of prediction, the input variables can be either categorical or continuous; different prediction methods are more effective, depending on the type of input variables used. In discovery with a model, model of a phenomenon is developed via clustering, prediction, or in some cases knowledge engineering. So this model is used as a component in another analysis, like relationship mining or prediction. In the prediction case, the created model's predictions are used as predictor variables in predicting a new variable.

IV. Related Work

Estimation and prediction may be viewed as types of classification. The following table 1 shows the comparison within the working of existing algorithms. These algorithms are the most influential data mining algorithms in the research community [17]. Different classification algorithms are categorized in following table 1:

Classification Algorithm	Type
Statistical	Regression Bayesian
Distance	Simple distance K nearest neighbors
Decision tree	ID3 C4.5 CART SPRINT
Neural network	Propagation NN supervised learning Radial base function network
Rule based	Genetic rules from DT Genetic rules from NN Genetic rules without DT and NN

Decision tree models can be compared and evaluated according to the following criteria:

(1) Measure: The ability of the model to correctly classify the unseen data on the basis of Entropy information gain or Gini indexing.

(2) Procedure: The procedure used to construct a decision tree either in top down or breadth first manner.

Two main pruning strategies:

- *Post pruning*: takes a fully-grown decision tree and discards unreliable parts. Possible strategies for post pruning are error estimation, significance testing, MDL principle. Bottom-up pruning is applied in C4.5 decision tree algorithm.
- *Preprinting*: stops growing a branch when information becomes unreliable. It simplifies a decision tree to prevent over fitting to noise in the data. Stops growing the tree when there is no statistically significant association between any attribute and the class at a particular node.

Literature Survey

This work examined the use of decision tree ensembles in biomedical time-series classification. Given algorithms are shown to be accurate and fast, as they construct diverse classifiers in little time, and vote strongly for the target class [5]. J.R.Quinlan [4] performed experiments with ensemble methods Bagging and Boosting by opting C4.5 as base learner.

In this work three different supervised machine learning techniques is applied in cancer classification, namely C4.5, bagged and boosted decision trees. Classification task is performed on seven publicly available cancerous microarray data and compared the classification/prediction performance of these methods. They observed that ensemble learning often performs better than single decision trees in this classification task[8]. Jinyam LiHuiqing Liu et.al. [9]

Experimented on ovarian tumor data to diagnose cancer using C4.5 with bagging and without bagging.

Han and Kamber [10] describes data mining software that allow the users to analyze data from different dimensions, categorize it and summarize the relationships which are identified during the mining process. This work attempts to propose a framework called Faculty Support System (FSS) that would enable the faculty to analyze their student's performance in a course. Supervised association rule mining is used to identify the factors influencing the result of students and C4.5 DT algorithm to predict the result of student. This work concentrated on the identification of factors that contribute to the success or failure of students in a subject and predict the result. [14].

We have thus proposed a novel and effective three-stage learning technique - partition, bag each partitioned subset, and learn[16]. The objective of this work is to evaluate the performance of employee using Decision Tree algorithm. The employee data are evaluated for giving promotion, yearly growth and career progress. To provide yearly increment for an employee, evaluation is performed by using past historical data of employees [18].

A cancer prediction system based on data mining is proposed in this work. This system estimates the risk of the lung, breast and skin cancers. This system is validated by comparing its predicted results with patient's prior medical information and it was analyzed by using weka system. Objective of this model is to provide the earlier warning to the users, and it is also cost efficient to the user [21].

In this work, an ensemble learning algorithm is applied within a classification framework that already got good predictive results. Ensemble technique is applied here which takes individual classifier, to combine them to improve the individual classifier result with a voting scheme. An algorithm is proposed here which starts by using all the available experts and removes them one by one focusing on improving the ensemble vote [23].

V. Proposed Model

Majority of students in higher education join a course for securing a good job. Therefore taking a wise career decision regarding the placement after completing a particular course is crucial in a student's life. An educational institution contains a large number of student records. Therefore finding patterns and characteristics in this large amount of data is a difficult task. Higher Education is categorized into professional and non-professional education.

Professional education provides professional knowledge to students so that they can make their stand in corporate sector. Professional education may be technology oriented or it may be totally concentrating on improving managerial skills of candidate. Here algorithms are applied on student's technical data base like their final year marks and skills known, and prediction is performed on some pattern to know the placement level of the student.

5.1 System Architecture:

Proposed system architecture is shown in below figure. There are many sub components in the architecture that cater intermediate results for new sub system.

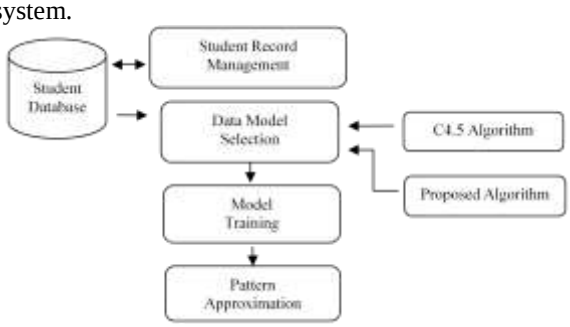

Fig 1: System Architecture

Different sub components of system architecture are:

- **Student Database:** At student's data base has been collected according to the requirements like final year result and skills known by them to.
- **Student Record Management:** In this part student's dataset is form and for managing the dataset for student records in this component options for adding, updating and deleting records is present.
- **Data Model Selection:** In this section data model is selected by user so that data analysis can be performed to develop a model.
- **C4.5 algorithm:** This is a decision tree classifier implemented for growing decision tree.
- **Proposed algorithm:** In this part a proposed algorithm is implemented which is formed by applying modifications in C4.5 algorithm to get improved results.

- **Model training:** In this section selected data model algorithm is used to process the data from the available data base and formulates decision tree for data pattern approximation.
- **Pattern approximation:** In this part according to existing data user feed their information to get prediction for placement status.

5.2 C4.5 Decision tree classifier:

To select the best decision tree algorithm for predicting the results we analyzed the student data with two different decision tree algorithms. Number of folds cross validation is used in the experiment.

INPUT: Tentative data set D which is showed by discrete value attributes.

OUTPUT: decision tree algorithm T which is created by giving experimental dataset.

i) Create the node N;
ii) If instance is related to the same class
iii) Then return node N as leaf node and marked with CLASS C;
iv) IF attribute List is null, THEN
v) Return node N as the leaf node and signed with the most common CLASS;
vi) Selecting the attribute with highest information gain in the attribute List, and signing the test_attribute;
vii) Validation the node N as the test_attribute;
viii) FOR the well-known value of each test_attribute to divide the samples;
ix) Producing a new branch which is fit for the test_attribute = ai from node N;
x) Let Ci is the set of test_attribute= a_i in the samples;
xi) IF Ci = null THEN
xii) Adding a leaf node and labelled with the most common CLASS;
xiii) ELSE we will add a leaf node return by the Generate_decision_tree.

5.3 Proposed Algorithm:

Much of the research in learning has tended to focus on improved predictive accuracy so that the performance of new systems is often reported from this perspective. It is easy to understand why this is so, accuracy is a primary concern in all applications of learning and is easily measured as opposed to intelligibility which is more subjective while the rapid increase in computers performance cost ratio has deemphasized computational issues in most applications in the active sub area of learning decision tree classifiers.

The data for classifier learning systems consists of attribute value vectors or instances. Both bootstrap aggregating or bagging and boosting a manipulate the training data in order to generate different classifiers Bagging produces replicate training sets by sampling with replacement from the training instances. Boosting uses all instances at each repetition but maintains a weight for each instance in the training set that respects its importance adjusting the weights causes the learner to focus on different instances and so leads to different classifiers. In either case the multiple classifiers are then combined by voting to form composite classifiers. In bagging each component classifier has the same vote while boosting assigns different voting strengths to component classifiers on the basis of their accuracy.

Ensemble methods are learning algorithms that construct a set of classifiers and then classify new data points by taking a vote of their predictions. The main objective of ensemble methodology is to try to improve the performance of single classifiers by inducing several classifiers and combining them to obtain a new classifier that outperforms every one of them. The most widely used ensemble learning algorithms are AdaBoost and Bagging whose applications in several classification problems have led to significant improvements. These methods provide a way in which the classifiers are strategically generated to reach the diversity needed by manipulating the training set before learning each classifier.

Enhanced C4.5 algorithm

1. There are n base learners, known as "data modal" for classifying a set of data.
2. Data may inconsistent by value therefore sometimes a data model learner performs faster and second will perform slow process.
3. Therefore, if a classification set have (D_1, D_2,… D_n) data models to learn them, a cross validation process can works as

$$Accuracy_{Di} = \sum_{i=0}^{n} \frac{A_{Di}}{i}$$

Therefore normalize to normalize the weights, can be calculated by calculating average weight

$$W_{Di} = W_{D1} + W_{D2} \dots W_{Dn}$$

Engineering Universe for Scientific Research and Management
(International Journal)
Vol. 7 Issue 2
February 2015

$$W = \frac{1}{n}\sum_{i=0}^{n} W_{Di}$$

To scale the weights vectors in a week learner the difference from base line can be calculated as

$$\partial^2 = \sum (W - \overline{W_{Dn}})^2$$

Therefore

$$\partial = \sqrt{\sum (W - \overline{W_{Dn}})}$$

If $W_{Dn} \leq \partial$ than required to distribute weights for second learner.

5.4 GUI Implementation:

By using the provided support of visual studio 2008 the whole system is designed for efficient user navigation. In the given system first screen is given using figure:

Given figures contains screenshots of the proposed work at first login form appears, after getting successful login. Menu items of our application can be accessed. Menu bar contains following items: File, Data model & Real time data and decision tree. Through file menu manual data can be generated. Data modal menu item contains existing data sets and decision tree algorithm is applied. Last menu item contains real time data and both algorithms are applied here on manually generated data.

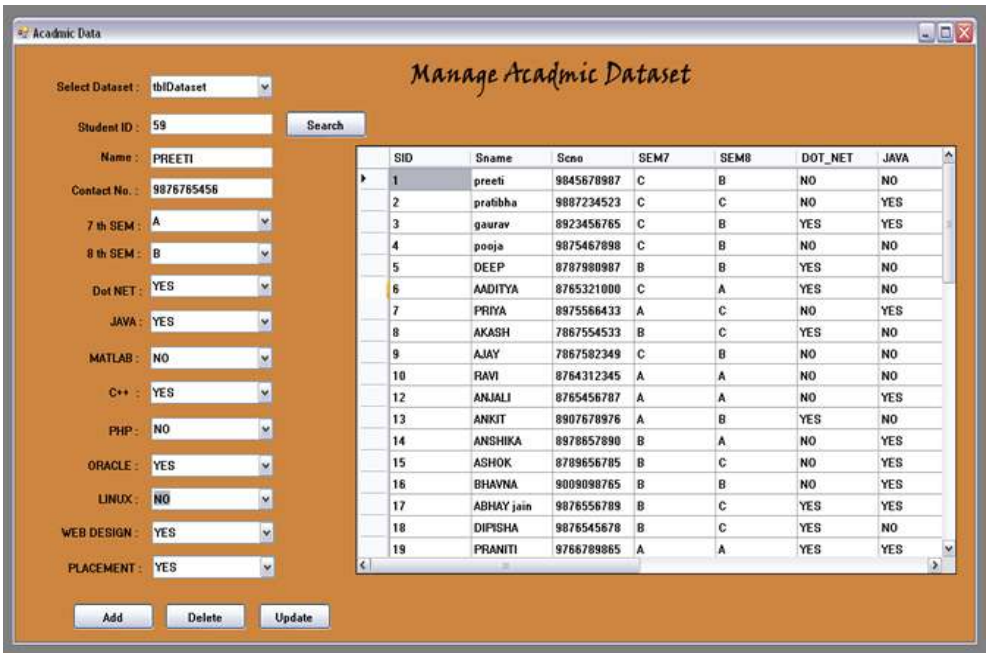

Fig 2: Student Data Management Screen

Engineering Universe for Scientific Research and Management
(International Journal)
ISSN(ONLINE) : 2319-3069
Vol. 7 Issue 2
February 2015

Fig 3: Classification Screen on Real-Time Dataset

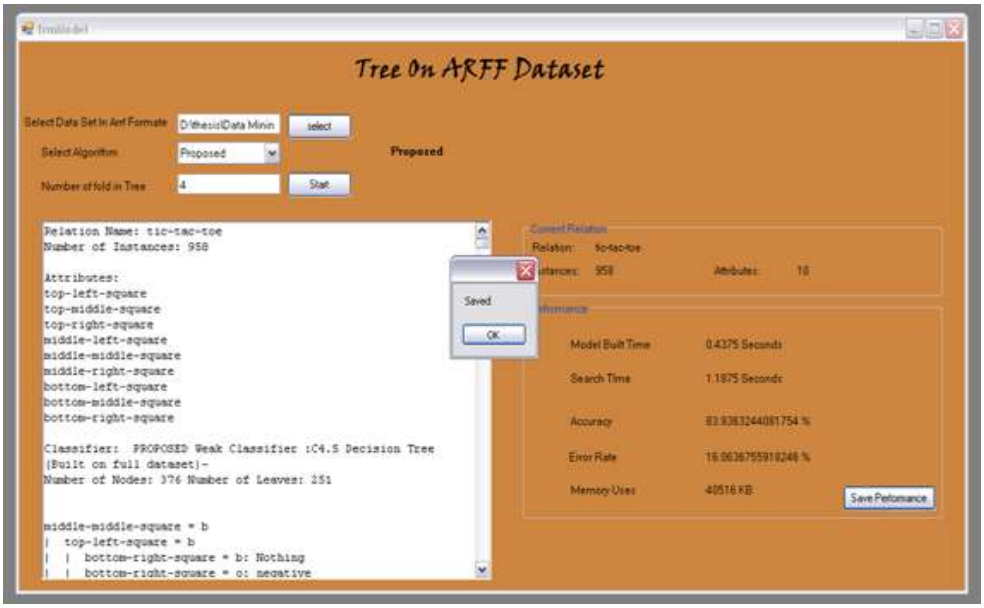

Fig 4: Classification Screen on Arff Dataset

VI. Result Analysis

Data Mining is gaining its popularity in almost all applications of real world. One of the DM technique i.e., classification is an interesting topic to the researchers as it is accurately and efficiently classifies the data for knowledge discovery. In decision tree, rules are extracted from the training dataset to form a

tree structure, and this rule will be applied to the classification of testing data. Decision trees are so popular because they produce human readable classification rules and easier to interpret than other classification methods. Here Classification task is used in student's educational database to predict students performances on the basis of their skills learn. Information like academic records, technical skills known was collected from the students' previous record, to predict the placement status. This study helps to predict whether to knowing particular skill and having better academic record will be help for their placement. In this paper we have chosen classical C4.5 and enhanced C4.5 for performance analysis. The C4.5 algorithm recursively classifies data until it has been classified as perfectly. This technique gives maximum accuracy on training data.

The accuracy percentage of each of the algorithm according to 5 different parameters is shown in Table 2.

Parameters/ Classified Instances	Existing Dataset(ARFF)		Real Time Dataset	
	C 4.5 Classifier	Proposed Algorithm	C 4.5 Classifier	Proposed Algorithm
Accuracy	76.55%	82.9%	73.76%	79.10%
Error Rate	23.45%	17.10%	26.24%	20.89%
Memory used	34128 KB	40724 KB	30836 KB	30919 KB
Search Time	0.24 Sec	0.63 Sec	0.31 Sec	0.42 Sec
Build Time	0.26 Sec	0.44 Sec	0.34 Sec	0.46 Sec

6.1 Prediction Form

In this work we have constructed an expert system. That predicts the placement status according to skills known. It helps the students to enhance their technical skills and academic records also .This prediction system consists of various functional units listed below:

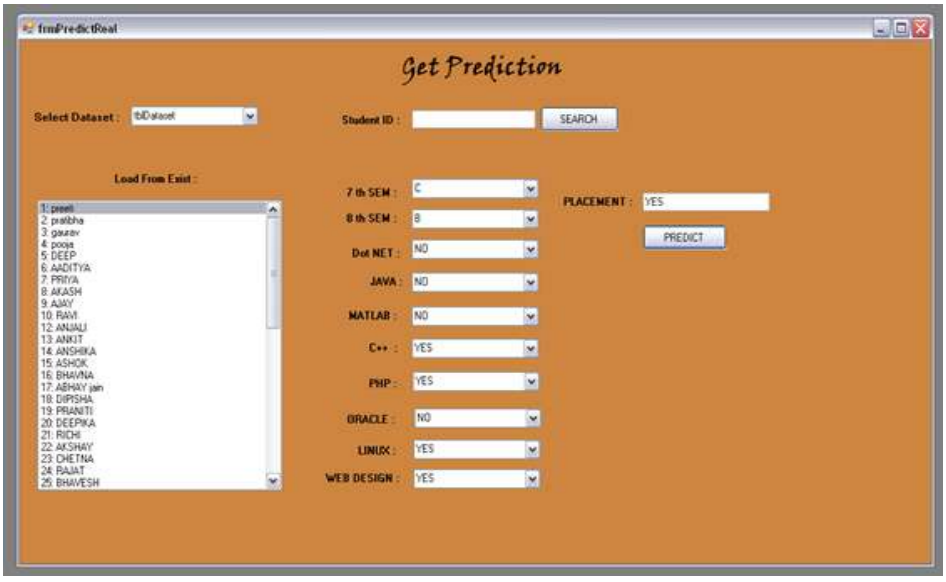

Fig 5: Classification Screen on Arff Dataset

Data mining based students placement prediction system is used to predict the placement status. Once the user opens prediction form, they need to answer the queries, either they have that particular skill or not. Then the prediction system finally predicts the result and answer is yes or no either it can be placed.

VII. Conclusion

This system can be very easily implemented by any educational institution. It can be used by faculties who do not have any knowledge on data mining techniques. Although there are so many benchmarks comparing the performance and accuracy of different classification algorithms, there are still very few experiments carried out on Educational datasets. In this work, we compare the performance and the interpretation level of the output of different classification techniques applied on educational datasets. Our experimentation shows that there is not one algorithm that obtains significantly better classification accuracy, so ensemble of classifier is created. Future work can concentrate on other student data analysis techniques that would mine other useful knowledge.

References:

[1] J. R. Quinlan, "Introduction of DT", 1986 Journal of Machine learning", pp. 81-106.
[2] J. R. Quinlan, "C4.5: Programs for Machine Learning", Publishers: Morgan Kaufmann, 1992.
[3] Breiman, L., "Bagging Predictors". Machine Learning, 1996.
[4] J.R. Quinlan, "Bagging, Boosting and C4.5", 14th National Conference on Artificial Intelligence" 1994.
[5] Schapire, R.E., Freund, Y., Bartlett, P., Lee, and W.S., "Boosting the margin: A new explanation for the effectiveness of voting methods", 14th International Conference on Machine learning, Morgan Kaufmann, San Francisco, pp. 322–330, 1997.
[6] J. Han and M. Kamber, "Data Mining: Concepts and Techniques", 2000, Morgan Kaufmann Publishers,.
[7] Nitesh Chawla, Lawrence O. Hall, Steven Eschrich, "Creating Ensembles of Classifiers" , International Conference on DM IEEE, 2001
[8] Tan,Gilbert, " Ensembling machine learning on gene expression data for cancer classification", Proceedings of New Zealand Bioinformatics Conference, Wellington, New Zealand, ,13-14, February 2003.
[9] Jinyan LiHuiqing Liu, Limsoon Wong and See-Kiong Ng, "Discovery of significant rules for classifying cancer diagnosis data", Bioinformatics 19, Oxford University Press 2003.
[10] Han,J. and Kamber, M., "Data Mining: Concepts and Techniques", The Morgan Kaufmann Series in Data Management Systems, Series Editor, 2006.
[11] Tom Diethe, John Shawe-Taylor, Jose L. Balcazar, ," Comparing classification methods for predicting distance students' performance", Workshop and Conference Proceedings JMLR, pp. 26-32, 2011
[12] Anshu Katare, Anant Athavale, Dr. Vijay, "Behavior analysis of different decision tree algorisms", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 1, Issue 1, pp. 43-47, August 2011.
[13] S.K Yadav. And S. Pal, "A prediction for Performance Improvement of Engineering Students using Classification", WCSIT (World of Computer Science and Information Technology Journal), Vol.2, pp- 51-56, 2012.
[14] T. Venkatachalam ,J. Shana, "A Framework for Dynamic Faculty Support System to Analyze Student Course Data", Volume second, Issue 7th, pp.478-482, July 2012.
[15] Jinyu Wen ,Haibo He, Yuan Cao,Yi Cao, "Ensemble learning for wind profile prediction with missing values", 2011.
[16] Mattia Bosio, Pau Bellot, Philippe Salembier, Albert Oliveras Verg´es," Ensemble learning and hierarchical data representation for microarray classification", 2013.
[17] T.Miranda Lakshm, Dr.V.Prasanna Venkatesan A.Martin, R.Mumtaj Begum, " An Analysis on Performance of Decision Tree Algorithms using Student's Qualitative Data", I J Modern Education and Computer Science, pp.18-27, 2013.
[18] P. Thangaraj, N. Magesh M.E., "Evaluating The Performance Of An Employee Using Decision Tree Algorithms", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 4, , pp. 2814-2830 April – 2013.
[19] S. K. Yadav, B.K Bharadwaj. and S Pal., "Data Mining Application: A comparative study for Predicting Student's Performance", International Journal of Innovative Technology and Creative Engineering (IJITCE), Vol. 1, No. 12, pp. 13-19.
[20] K. Usha Rani ,G. Sujatha, Dr., "An Experimental Study on Ensemble of Decision Tree Classifier", International Journal of Application or Innovation in Engineering & Management (IJAIEM), Volume 2, Issue 8, August 2013, pp 300-306.
[21] A.Priyanga, S.Prakasam,, " Effectiveness of Data Mining - based Cancer Prediction System (DMBCPS)", International Journal of Computer Applications, Volume 83 – No 10, pp 11-17, December 2013.
[22] Ryan S.J.d. Baker, "Data Mining for Education", Carnegie Mellon University, Pennsylvania, USA.

[23] Yahya Abu Hasan ,Yin Zhao, "Fine Particulate Matter Concentration Level Prediction by using Tree-based Ensemble Classification Algorithms", International Journal of Advanced Computer Science and Applications (IJACSA), Vol. 4, No.5, pp. 21-27, 2013.