Predictive Assessment of Learner's Performance Using Decision Trees with Genetic Algorithms

[1]S Neelima, [2]Bodapati Prajna

[1,2]Dept. of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, AP, India

Abstract

We recommend that the configuration and execution of compelling Social Learning Analytics present critical difficulties and open doors for both exploration and endeavor, in three imperative regards. The first is that the learning scene is exceptionally turbulent at present, in no little part because of mechanical drivers. Online social learning is rising as a huge wonder for an assortment of reasons, which we survey, so as to spur the idea of social learning. We finish up by returning to the drivers and inclines, and consider future situations that we may see unfurl as SLA devices and administrations full grown. Conceptual Data Mining is a famous information revelation procedure. In information mining decision trees are of the straightforward and capable basic leadership models. In this anticipate, an algorithm is proposed for predicting a learner's performance utilizing decision trees and genetic algorithm, GDADT algorithm. Id3 algorithm is utilized to make numerous decision trees, each of which predicts the performance of an understudy in view of an alternate list of capabilities. Since every decision tree furnishes us with an understanding to the plausible performance of every understudy; and diverse trees give distinctive results, we are ready to foresee the performance as well as distinguish regions or elements that are in charge of the anticipated result. For higher precision of the acquired results, genetic algorithm is additionally joined. The genetic algorithm is executed on the n-ary trees, by computing the wellness of every tree and applying hybrid operations to acquire numerous eras, each adding to making trees with a superior wellness as the eras increment, lastly bringing about the decision tree with the best exactness. The outcomes so acquired are very reassuring.

Keywords

Genetic Algorithm, Decision Tree Algorithm, GDADT, Inductive Learning

I. Introduction

In Machine Learning people group, and in information mining works, order has its own significance. Order is an imperative part and the exploration application field in the information mining [1]. A decision tree gets its name since it is molded like a tree and can be utilized to decide. ―Technically, a tree is an arrangement of nodes and branches and every branch plummets from a hub to another hub. The nodes speak to the characteristics considered in the decision procedure and the branches speak to the diverse property estimations. To achieve a decision utilizing the tree for a given case, we take the trait estimations of the case and cross the tree from the root hub down to the leaf hub that contains the decision." [2]. A basic issue in counterfeit consciousness (AI) exploration is to defeat the alleged ―knowledge acquisition bottleneck" in the development of information based frameworks. Decision tree can be utilized to take care of this issue. Decision trees can gain information from solid illustrations as opposed to from specialists [3]. Also, for information based frameworks, decision trees have the upside of being understandable by human specialists and of being straightforwardly convertible into generation rules [4]. A decision tree not just gives the answer for a given case, additionally gives the explanations for its decision. So the genuine advantage of decision tree innovation is that it maintains a strategic distance from the requirement for human master. As a result of the above favorable circumstances, there are numerous triumphs in applying decision tree figuring out how to tackle genuine issues. Late years have shown a creating interest and worry in various countries about issue of school frustration and the determination of its guideline contributing components [2]. The extensive game plan of examination [5] has been done on recognizing the segments that impact the low execution of understudies (school disillusionment and dropout) at different informational levels (key, assistant moreover, higher) using the far reaching measure of information that present PCs can store in databases. All these data are a "gold mine" of noteworthy information about understudies. Recognize and find supportive information concealed in significant databases is a troublesome task. An amazingly promising response for fulfill this goal is the usage of data exposure in databases strategies or data mining in direction, called educational data mining, EDM

II. Background

This will first present the primary decision tree algorithms, and some methodologies went for conquering the issues brought about by the ravenous inquiry. After that, classifier frameworks in view of genetic programming will be examined, trailed by an area depicting half and half techniques.

A decision tree algorithm ordinarily enhances some data theoretic measure, similar to data pick up, on a preparation set. The era of the tree is done recursively by part the information set on the free variables. Every conceivable split is assessed by figuring the immaculateness pick up it would bring about in the event that it was utilized to isolate the information set D into the new subsets S={D1, D2,… ,Dn}. The virtue addition is the distinction in immaculateness between the first information set and the subsets as characterized in condition 1 beneath, where P(Di) is the extent of D that is set in Di. The split bringing about the most noteworthy immaculateness addition is chosen, and the technique is then rehashed recursively for every subset in this split.

There are a few distinctive decision tree algorithms, two of the all the more understood are GDADT [6] and CART [7]. Marginally distinctive immaculateness capacities are utilized, GDADT streamlines entropy E, (condition 2) while CART improves the gini record (GDI in condition 3.) In the conditions beneath, C is the conceivable classes, p is the evaluated class likelihood and t is the present tree hub.

Contrasted with enhancing GDI, entropy tends to prompt littler and purer nodes, which is good for issues with an unmistakable basic relationship, however mediocre when the information contain a considerable measure of clamor or are feeling the loss of a genuine relationship [8].

At the point when no parts enhancing immaculateness can be observed, the tree should be pruned to evacuate excessively particular nodes to enhance the speculation capacity of the tree. Pruning is ordinarily performed by picking the best sub-tree taking into account the blunder rate on a concealed acceptance information set.

A. Improving Imperfect Trees

Numerous specialists have attempted to enhance decision tree execution by considering a few successive parts rather than just the following. In any case, most studies have demonstrated that this methodology for the most part neglects to enhance the execution, and that it even might be hurtful.

There have likewise been a few endeavors to enhance imperfect decision trees utilizing a second stage where the tree is changed utilizing another pursuit system. A case of this methodology is [12], where a tree is initially made utilizing fluffy rationale seek. In the second stage, the terminal nodes in the tree are conformed to be streamlined in general preparing set. Different cases are [1] where a problematic decision tree is enhanced utilizing dynamic programming and [2] where multi-direct writing computer programs is connected also. It ought to be noted, notwithstanding, that regardless of the possibility that these algorithms enhance the imperfect decision trees, they are not genuinely utilizing worldwide hunt, since they are needy of the underlying structure of the tree.

B. Genetic Programming for Classification

Typically a full nuclear representation is utilized for GP grouping. A nuclear representation utilizes iotas as a part of inward and leaf nodes [11]. Each inner molecule speaks to a test comprising of a characteristic, an administrator and a quality, where the administrator is a Boolean capacity. Leaf nodes contain iotas speaking to a class of the anticipated quality.

GP order representations can be isolated into the Michigan and Pittsburgh approaches [6]. In the Michigan approach, every individual encodes a solitary expectation principle, while in the Pittsburgh approach every individual encodes an arrangement of forecast standards.

Another imperative issue to consider is that traditional GP depends on the fundamental supposition of conclusion [10]. To accomplish conclusion, the yield of any hub in a GP tree must have the capacity to handle all conceivable guardian nodes. This normally turns into an issue when a dataset contains both clear cut and ceaseless credits since they should be taken care of by various capacities.

One approach to handle this is to utilize compelled syntactic structures, see [3], where an arrangement of standards characterizes permitted sub nodes for each non terminal capacity. These guidelines are then upheld while making new trees and amid hybrid and transformation. Another marginally more adaptable arrangement is specifically GP [4], which rather characterizes the permitted information sorts for every contention of each non-terminal capacity and the returned sorts of all nodes.

III. Recent Advances in Decision Trees

In Data mining, the problem of decision trees has also become an active area of research. In the literature survey of decision trees we may have many proposals on algorithmic, data-level and hybrid approaches. The recent advances in decision tree learning have been summarized as follows: A parallel decision tree learning algorithm expressed in MapReduce programming model that runs on Apache Hadoop platform is proposed by [5]. A new adaptive network intrusion detection learning algorithm using naive Bayesian classifier is proposed by [6]. A new hybrid classification model which is established based on a combination of clustering, feature selection, decision trees, and genetic algorithm techniques is proposed by [7]. A novel roughest based multivariate decision trees (RSMDT) method in which, the positive region degree of condition attributes with respect to decision attributes in rough set theory is used for selecting attributes in multivariate tests is proposed by [8]. A novel splitting criteria which chooses the split with maximum similarity and the decision tree is called mstree is proposed by [9]. An improved ID3 algorithm and a novel class attribute selection method based on Maclaurin-Priority Value First method is proposed by [10]. A modified decision tree algorithm for mobile user classification, which introduced genetic algorithm to optimize the results of the decision tree algorithm, is proposed by [1]. A new parallelized decision tree algorithm on a CUDA (compute unified device architecture), which is a GPGPU solution provided by NVIDIA is proposed by [2]. A Stochastic Gradient Boosted Decision Trees based method is proposed by [3]. A modified Fuzzy Decision Tree for the fuzzy rules extraction is proposed by [4]. Obviously, there are many other algorithms which are not included in this literature. A profound comparison of the above algorithms and many others can be gathered from the references list.

IV. Proposed Algorithms

A. ID3 ALGORITHM

Iterative Dichotomiser 3 is a simple decision tree learning algorithm Inductive learning is the learning that is based on induction. In inductive learning Decision tree algorithms are very famous. For the appropriate classification of the objects with the given attributesinductive methods use these algorithms basically. These algorithms are very important in the classification of the objects. The algorithm is implemented in the java language.

B. ID3 Algorithm Follows These Steps

1. Choosing Attributes

The order in which attributes are chosen determines how complicated the tree is.ID3 uses information theory to determine the most informative attribute. A measure of the information content of a message is the inverse of the probability of receiving the message:

information1 (M) = 1/probability (M)

Taking logs (base 2) makes information correspond to the number of bits required to encode a message:

Information (M) = -log2 (probability (M))

2. Entropy

It is a measure in the information theory, which characterizes the impurity of an arbitrary collection of examples. If the target attribute takes on c different values, then the entropy S relative to this c-wise classification is defined as:

Entropy(S) = ∑ -P (I) log2 p (I)

where p (I) is the proportion of S belonging to class I,

S is the entire sample set.Logarithm is base 2 because entropy is a measure of the expected encoding length measured in bits.

For e.g. if training data has 14 instances with 6 positive and 8 negative instances, the entropy is calculated as

Entropy ([6+, 8-]) = 22 (6 /14) log (6 /14) (8 /14) log (8 /14) −−= 0.985

Main point to note here is that the more uniform is the probability distribution, the greater is its entropy.

3. Information Gain

It measures the expected reduction in entropy by partitioning the examples according to this attribute. The information gain, Gain(S, A) of an attribute A, relative to the collection of examples S, is defined as
Gain(S, A) = Entropy(S)-∑ ((|Sv|/|S|)*Entropy (Sv))
Where,∑ is each value v of all possible values of attribute
Sv is subset for which attribute A has value v
|Sv| =number of elements in Sv
|S|=number of elements in S
We can use this measure to rank attributes and build the decision tree where at each node is located the attribute with the highest information gain among the attributes not yet considered in the path from the root.

4. Rules of Classifying

If the entropy of the attribute is 0, it is a homogeneous node and there is no need to classify further. If the entropy of the attribute is 1, it is a heterogeneous node and there is a need to classify further.

5. Advantages of Using ID3

- Understandable prediction rules are created from the training data.
- Builds the fastest tree.
- Builds a short tree.
- Only need to test enough attributes until all data is classified.
- Finding leaf nodes enables test data to be pruned, reducing number of tests.
- Whole dataset is searched to create tree.

6. Disadvantages of using ID3

- Data may be over-fitted or over-classified, if a small sample is tested.
- Only one attribute at a time is tested for making a decision.

Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.

C. Genetic Algorithm

Genetic Algorithms (GAs) have been successfully applied to solve search and optimization problems. The basic idea of a GA is to search a hypothesis space to find the best hypothesis. A pool of initial hypotheses called a population is randomly generated and each hypothesis is evaluated with a fitness function. Hypotheses with greater fitness have higher probability of being chosen to create the next generation. Some fraction of the best hypotheses may be retrained into the next generation, the rest undergo genetic operations such as crossover and mutation to generate new hypotheses. The size of a population is the same for all generations in our implementation. This process is iterated until either a predefined fitness criterion is met or the present maximum number of generations is reached.

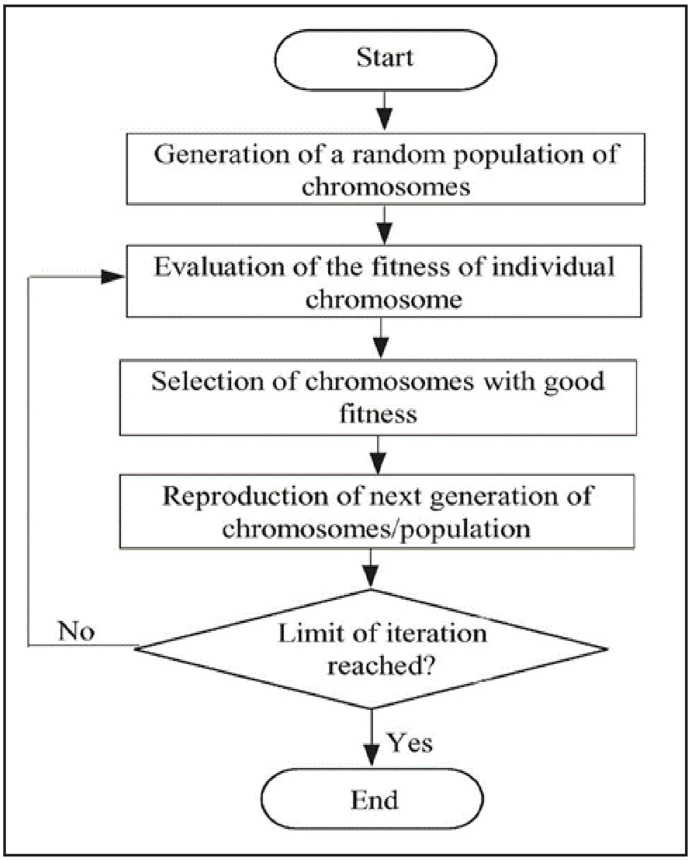
Start
Generation of a random population of chromosomes
Evaluation of the fitness of individual chromosome
Selection of chromosomes with good fitness
Reproduction of next generation of chromosomes/population
Limit of iteration reached?
No
Yes
End

Fig. 1:

A Genetic Algorithm generally has four components. A population of individuals where each individual in the population represents a possible solution. A fitness function which is an evaluation function by which we can tell if an individual is a good solution or not. A selection function which decides how to pick good individuals from the current population for creating the next generation. Genetic operators such as crossover and mutation which explore new regions of search space while keeping some of the current information at the same time. Genetic algorithm consists of three basic processes: encoding, crossover and mutation, selection.

A. Encoding

1. Solution Representation

We use a genetic algorithm to obtain a solution for the rough-set decision problem. First, we must find an encoding of potential solutions to the problem. This encoding must be meaningful to the problem and corresponding solution. Our problem is to find the minimum reduct of a decision table. The natural way to describe the reduct is bitmap. According to the reduct definition, the reduct representation is only a subset of attributes. Therefore, we create a binary string, the length of which is the same as the number of attributes in the decision table. Each bit of the string corresponds to one of the attribute of the decision table. If a certain bit is set to one that means the corresponding attribute belongs to the reduct. It is possible that when the genetic algorithm ends, we may have multiple solutions because these strings have the same fitness score.

B. The First Population

We randomly select a number of individuals from the distinction table. In addition, the distinction table is going to be involved in the fitness score computation for each individual. We first need to transform our decision table into a distinction table. Once the distinction table is formed, we can select a certain number of strings (each row is a string) from it to build the first population of the GA.

C. Genetic Operators

1. Crossover

Crossover is actually a recombination of parents. It is the most important operation of the GA. New individuals generated by crossover are part of the next generation. Performing crossover is also problem-specific. We just select two strings, and at some position, we recombine them, and two new offspring strings are born. Whether the offspring are part of the next generation (population) depends on their fitness score.

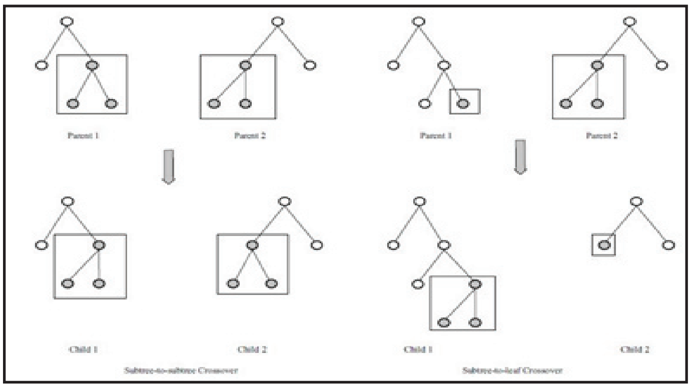

Fig. 2: Cross over

D. Mutation

We perform mutation after performing crossover. We implement the mutation by one bit flip. In other words, randomly select the position of the string and change it from 1 to 0 or vice versa.

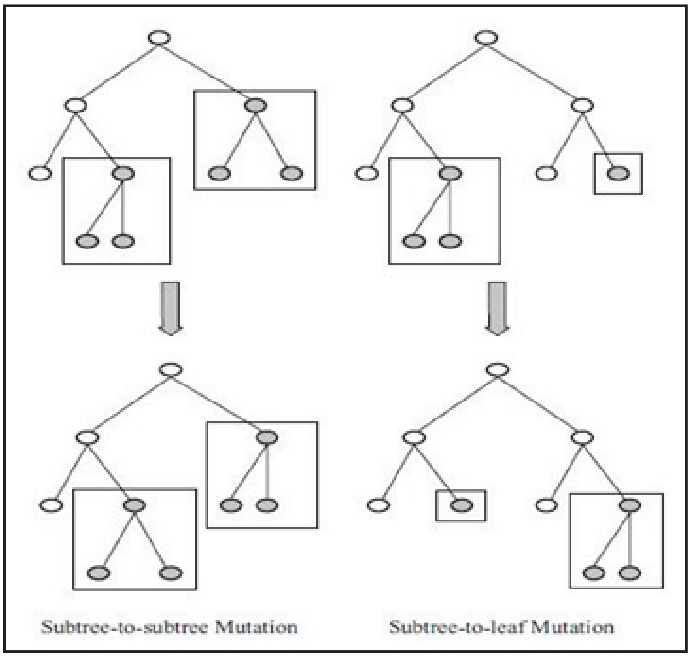

Fig. 3:

E. Selection

1. Fitness Function

It is very important to define a proper fitness function for the GA. The fitness function is used to assess the fitness of a string compared to the rest of the population. Each string (individual) gets its own fitness score computed from the fitness function. Individuals with the highest scores are selected or maintained after Xover and mutation. Finally, the GA stops after a certain number of generations, and the individual with the highest score in the last generation is the solution.

Fitness = Nc / N+1

Nc the number of correct classified cases

N is the number of cases covered by the rule

V. 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. GDADT approach to missing attribute values is another example of a method from thisgroup. GDADT 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.

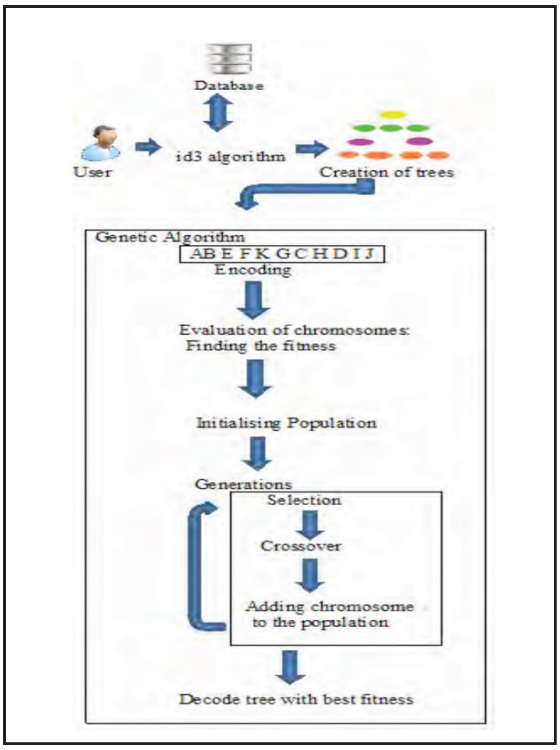

Fig. 4: Proposed System Architecture

Genetic based Data Adopted Decision Tree (GDADT) Algorithm

Input: Data set, D = {a1, a2…an}

Output: Resulted Tree (Selected Instances)

Step-1: Selecting the best attributes of the given data set using genetic algorithm with fitness function, Fitness (W sensitivity) (W specificity) =×+× 1 2

Step-2: Initialize the set of unique values (UV) and the regular intervals (RI)

Step-3: Splitting data into many classes C1, C2….Cn, with different labels

Step-4: Set the Decision node (D) and Best Attributes (Bbest)

Step-5: If the unique value is belongs to the particular class then Split the regular intervals. Else the unique value is belongs to another particular class then Assign the highest probability to the particular class members (instances).

Step-6: Splitting the instances into many groups and updates the intervals.

Step-7: Repeat step 5 until read all unique values

Step-8: Update all the records into the corresponding classes of dataset with best attributes

Step-9: Building the decision tree for updated records using EMSVM algorithm for making final decision.

The proposed algorithm reads the necessary data from the data set by the help of user interface as input and the selected instances are output of the system in the tree format. First, selects the best attributes from the given input dataset by using the genetic algorithm which is developed by the introduction of new fitness function in this algorithm. Second, split the datasets based on the regular intervals and unique values. The best attributes were selects from the data set as decision nodes and split into several classes based on the unique values and time intervals.

VI. Conclusion

GDADTAlgorithm technique for diminutive sample in the automatic classification has enhanced classification consequences. When the necessitate for a sub-sample to be confidential, when intended to be simply sub-samples and the correspondence of every type of vector that is the internal product, and then choose the class of the record comparison to be sub-sample of the equivalent class address the diminutive example, nonlinear and high dimensional pattern recognition presentation of a lot of unique compensation, and can be functional to purpose estimate and extra machine learning problems.

References

[1] Document clustering based on correlation preserving indexing, International Journal of Scientific Research in Computer Science Applications and Management Studies.

[2] Document clustering technique based on noun hyper nyms, International Journal of Electronics and Communication and Technology.

[3] C. C. Aggarwal, J. Han, J. Wang, P.S. Yu,"On demand classification of data streams", In Proceedings of KDD, pp. 503-508, 2004.

[4] Z. Bandar, K. Crockett, D. McLean, J. O'Shea,"On constructing a fuzzy inference framework using crisp decision trees", Fuzzy Sets and Systems, 157, 21, pp. 2809-2832, 2006.

[5] P. A. Berson, Smith, S.J.,"Data Warehousing, Data Mining, & OLAP", McGraw Hill, New York, 1997.

[6] H. R. Bittencourt, R.T. Clarke,“Use of Classification and Regression Trees (CART) to Classify Remotely Sensed Digital Images”, Proceedings of International Geosciences and Remote Sensing Symposium, IGARSS ’03, Vol. 6, pp. 3751-3753, 2003.

[7] L. Breiman, J. Friedman, C.J. Stone, R.A. Olshen, "Classification and Regression Trees", CRC Press, Boca Raton, FL, 1984.

[8] J. Catlett,"On changing continuous attributes into ordered discrete attributes", Springer, Berlin, Heidelberg, Porto, Portugal, 1991.

[9] T. M. Cover,"Elements of Information Theory", Second ed. Wiley- Interscience, New York, NY, 2006.

[10] D. E. Culler, W. Hong, Wireless sensor networks - introduction. Communications of the ACM 47 (6), pp. 30–33, 2004.

[11] U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth: From Data Mining to Knowledge Discovery: An Overview. Advances in knowledge discovery and data mining book contents, 1-34, 1996.

[12] U. M. Fayyad, K.B. Irani, Multi-Interval Discretization of Continuous-Valued Attributes for Classification Learning, Machine Learning, 13th IJCAI, Vol. 2, Chambery, France, Morgan Kaufmann, pp. 1022-1027, 1993.

S Neelima is Pursuing M.Tech (IT) in the Dept. of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.

Prof. Bodapati Prajna is working as a Professor in Dept. of Computer Science and Systems Engineering, Andhra University, Visakhapatnam, Andhra Pradesh, India.