IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 16, Issue 4, Ver. V (Jul – Aug. 2014), PP 57-65
www.iosrjournals.org

Implementation of Novel Algorithm (SPruning Algorithm)

Srishti Taneja
Lovely Professional University, School of technology and sciences, NH-1, Punjab, India

Abstract: *Decision trees are very significant for taking any type of verdict related to any field. Today there is ample amount of data but that data is uncooked data therefore to make it cooked data, data mining is done. Data mining concept has been used to enhance the quality of data. Various techniques are there in data mining for growth in information technology. Those techniques are classification; clustering, association rules etc. For classifying the data best method is to generate a decision tree. By using decision tree best decisions can be made as it is hierarchal structure which can be easy to interpret. A novel algorithm is implemented which gives results faster and has enhanced features than conventional algorithms of data mining. In new approach some features of Univariate algorithm (i.e. CART algorithm) and some features of Multivariate algorithm (i.e. M5P algorithm) plus an enhanced feature of new algorithm is included. The new feature of algorithm does enhancement in performance as using it performance is improved. In novel approach pruning of files is done due to which specific data can be accessed. Time is saved by using this algorithm and user can perform the task quickly and efficiently.*

Keywords: *CART, Data Mining, Decision Trees, M5P.*

I. Introduction

Classification tree is a diagram used for categorization of information for decision making process in decision making system. They includes a root node containing only outgoing edges and no incoming edges, leaf node containing only inward and no outgoing edges (symbolize classes), internal nodes containing both incoming as well as outgoing edges (which symbolize conditions) etc.

Decision tree execute categorization in two phases: one is tree growing phase and other is tree pruning phase. Tree clipping is very vital step and is used for perfect tree which is free from any outliers. In this classification of dataset are performed. They utilize greedy algorithm that follow divide and conquer approach. Decision tree is utilized to symbolize classifiers, regression models and at times hierarchical model also. There are 2 approaches for decision trees i.e. univariate decision trees and multivariate decision trees. Multivariate decision trees overcome the disadvantages of univariate approach. A novel algorithm is designed in which pruning concept is enhanced.

1.1 Classification Trees

Depending upon values of features classification of demand is done in classification trees. They are also used to take decision in decision making systems. Other than classification tree there are various methods utilized for classification some of them are neural networks, support vector machines etc.

Decision tree is widely used this is because decision tree is very simple and transparent, it can be understood very easily by anyone there is no need of expert to understand it. Hierarchical representation is done in this which is simple to understand and it's very simple.

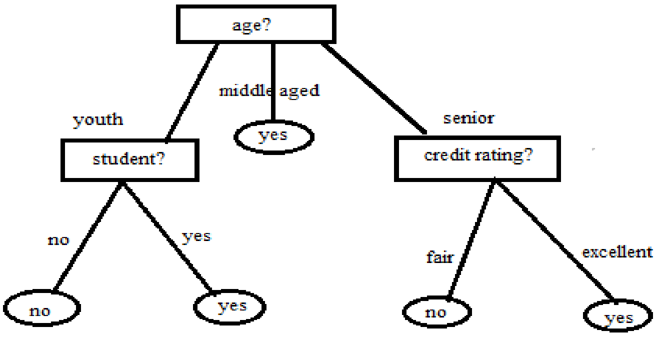


Fig 1.1: Decision Tree

1.2 Characteristics of Classification Trees

Some characteristics of Decision Trees are:

- Elasticity: tree can be shape according to our need
- Speed: processing is very quick as it can be understood quickly
- Interpretability: it can be interpreted easily as it is self explanatory
- Size: complex trees are not preferred as it affects the accuracy
- Hierarchical nature: easily understood as a result cost is also reduced [1]

Trees make use of greedy algorithm to classify the data. Its creation involves two types of stages and they are:

- Growth Phase (build up the tree)
- Pruning Phase(clipping unnecessary tree)

The first phase i.e. tree growing phase follow top-down approach. Tree partitioning is done till we get data items of same class as a residue. The second phase i.e. tree pruning phase track bottom up approach. This phase is crucial phase in which tree is engrave reverse so that prevention can be done from over fitting and also improves correctness of classification tree.

Types of pruning are:-

i) Post pruning (finished after tree creation)-Firstly full tree is drawn and following that its pruning is done i.e. outliers are removed from the tree in foundation up manner. If error improvement is noticed by pruning then sub tree is replaced by leaf.

It includes:

a) Minimal Cost complexity Pruning
b) Reduced Error pruning
c) Pessimistic Pruning

ii) Pre Pruning (finished during tree creation)-In this creation of tree is stopped in between. Stop when values of all attributes are analogous or when all instances belong to same class. [1]

1.3 Approaches of Trees

2 approaches for decision trees are:-

- Univariate Decision Tree: This is used for small information. Throughout this approach, a characteristic is used at internal nodes and after that splitting is done. E.g. X<7, y>=9 etc. We can create these type of trees by number of algorithms and these algorithms are ID3, c4.5 etc. J48 algorithm is an extension of ID3 algorithm and it possibly creates a small tree. They uses divide and conquer approach to growing decision tree.
- Multivariate Decision Tree: It is utilized for enormous dataset. It removes issue of Univariate approach i.e. inaccuracy of data is removed by this. This approach of decision tree uses additional than one attribute. Test condition in these trees may be as x + y>10. It is a non linear mixture of attributes at every test nodes. Univariate decision tree may include noisy data but multivariate decision tree can remove noise and makes data accurate. This approach is considered to be far improved than Univariate decision tree approach [1].

1.4 Algorithms Used For Creation of Decision Tree

- **C4.5**: It is successor of two algorithms CLS and ID3.It is predecessor of C5.0 algorithm. Similar to these algorithms it creates classifiers as classification trees and it also generates classifiers in further constant rule set form. It uses divide and conquer strategy to create tree. In this depth first construction of tree is done and it uses information gain. At each node sorting of continuous attributes is done and this algorithm can't be used for large dataset. It includes two criteria for testing order and they are information gain and gain ratio. Both numeric and nominal value can be used in this algorithm. Pruning is done in this using single-pass algorithm for avoiding over fitting and it is done in bottom up manner. Residual in this is two or more outputs.

Features of C4.5 are:

i) Handling missing values
ii) Candidate testing is done
iii) Avoid over fitting
iv) Pruning of decision trees [9].

- **Naive Bayes**: It is very significant algorithm and also known idiot's bayes. This algorithm is used for various purposes. It can be developed in easy method and complex recursive parameters schemes are not required. It means it may be iteratively applied to large amount of data sets [9].

- **CART**: Full form of this algorithm is ‘Classification and Regression Trees’ which are used to show main objective in the evolution in the field of Artificial Intelligence, data mining and Machine Learning. Its records can be found in almost any field, appear in domains such as electrical engineering, biology etc. Its classification tree is a binary recursive partitioning method which is used for processing continuous and nominal parameters. It follows divide and conquer strategy. No binning is required in data. Number of trees is generated as a resultant. It utilize concept of pruning known as minimal cost complexity pruning. Missing values can be handled by CART algorithm. It is also utilized to generate regression trees. It involves different steps:

i) Construction of tree
ii) Discontinue the construction of tree
iii) Pruning of tree
iv) Optimization of tree

Features of CART are:

i) Class balancing
ii) Automatic missing value handling
iii) Dynamic feature construction
iv) Probability tree estimation

Different characteristics are:

i) Splitting done in this is binary and features are taken one by one i.e. at a time only one feature is considered.
ii) Criterion to split the values is based on information gain and Gini index [9].

- **J48**: This algorithm is an execution of C4.5 algorithm .C4.5 was a adaptation of J48 algorithm which was used former and developed by J.Ross Quinlan. There are two techniques for pruning which is supported by J48.In this algorithm nodes are replaced by leaf. It reduces number of test for specific path. In this bottom up approach is used starting from leaves and shift to origin. Two methods of pruning are: sub tree replacement sub tree raising [9].
- **Ridor**: Firstly default rule is generated and then tree is generated for exception. Exceptions are also generated by this algorithm and this algorithm can’t be used for large dataset.
- **M5P**: This algorithm is a multivariate tree algorithm which is utilized for noise removal and also used for huge database .By this regression trees can be generated whose leaves are amalgamation of multivariate linear models. Error reduction can be done by this algorithm. M5P algorithm is advantageous as compared to other algorithms. It is used for both categorical and continuous variables and for missing values. It is preferred over CART as in CART regression trees are much smaller and are more accurate than that of M5P.

Features of M5P algorithm are:

i) Error estimates
ii) Simplification of linear models
iii) Pruning
iv) Smoothing

Steps involved in M5P are:
Step 1: Construction of tree
Step 2: Pruning of tree
Step 3: Smoothing of tree [10].

II. Review of Literature

S. Anupama Kumar et al. (2013) designed an approach for tree generation. She described that by feedback job security option of teacher’s, student’s feedback helps in developing good quality teaching learning atmosphere. For evaluation of teachers tool used can be regression trees. For classification of data she used an approach by which accuracy of data is improved. She firstly input the data, then model is applied after that a tree is obtained on which pruning is done. In this paper algorithms used for assessment are: - REP Tree & M5P algorithm. In this pruning can be done so as to improve accuracy of algorithm. REP Tree is bottom up strategy but M5P follow top down approach.

In M5P algorithm the predictable error reduction equation will be:

$$\Delta \text{error} = \frac{m}{|S|} * \beta(i) * \left[\text{stdev}(S) - \sum \left(\frac{|Si|}{|S|} \text{stdev}(Si) \right) \right] \quad (1)$$

Both algorithms develop model tree. In REP Tree regression function is used and in M5P we use linear regression function. It was find out that performance and accuracy of REP Tree is far improved than M5P algorithm. Pruning plays a vital role in developing a tree with lesser time. Accuracy of Pruned REP tree algorithm is improved to develop tree as compared to M5P tree. [1]

Dr. Neeraj Bhargava et al. (2013) had given a review of basic concepts of data mining i.e. he told about decision tree, characteristics of tree, approaches used, and also described about pruning concept in tree by which accuracy of data will improve. He designed a novel approach of decision tree for huge amount of data. He discussed that data mining is removal of noisy data and knowledge can be obtained from it. Decision tree are classification trees that uses two approaches for better outcome.
2 approaches for classification tree are:-

- Univariate decision tree: This approach is used for small data. During this approach, one attribute is taken at internal nodes and then splitting is performed.
- Multivariate decision tree: This approach is used for large dataset. Univariate tree may sometimes results in inaccurate tree and multivariate decision tree removes this issue

  J48 algorithm's rules slow for huge and raucous data. Space complexity is extremely more because values are repetitively in arrays .As a result use of M5P to create decision and regression tree, in M5P algorithm P stands for prime. Multivariate approach is better than Univariate approach as it allow dealing with huge quantity of data. In his research he performed experiment on data using univariate as well as multivariate technique as a result he concluded that multivariate technique is much better. This is because multivariate approach is applicable for large data and also removes outliers from data i.e. it improves accuracy of data [3].

W.Nor Haizan W. Mohamed et al. (2012) discussed about Reduced Error Pruning technique in decision tree algorithms. He discussed that Classification trees are most accepted and well-organized technique used in data mining. Various tree algorithms are used to create decision trees, some of them are complex and some are simple it depends upon size of data. Complex trees are difficult to recognize. To better understand pruning methods, an experiment was conducted using Weka application to compare the performance in term of complexity of tree structure and accuracy of classification for J48, REPTree, PART, JRip, and Ridor algorithms using seven standard datasets from UCI machine learning repository. In data modeling, J48 and REPTree generate tree structure as an output while PART, Ridor and JRip generate rules. In additional J48, REPTree and PART using REP method for pruning while Ridor and JRip using improvement of REP method, namely IREP and RIPPER methods. The experiment result shown performance of J48 and REPTree are competitive in producing better result. Pruning is most vital for purification of data. There are two standard classes of techniques for pruning are: pre-pruning and post pruning. J48 and REPTree create more accuracy of classification and simple tree arrangement. Ridor algorithm is used for best performance. He compared various algorithms and finally concluded that J48 and RepTree algorithm improves accuracy, Ridor improves performance and JRip improves performance in terms of complexity [2].

Goyal Anshul et al. (2012) compared 2 decision tree algorithms, he presents that classification is used in every field of life. Classification utilized to categorize each entry in a dataset into already defined classes or groups. He carried out a survey to make a performance evaluation of Naïve Bayes and j48 classification algorithm. Naive Bayes algorithm is based on chance and j48 algorithm is based on choice. A tree is generated using J48 algorithm and the tree generated is improved as compared to other algorithms. J48 is a simple classifier method to construct a decision tree, well-organized consequence has been taken from dataset of a bank using weka tool. Naive Bayesian classifier also present superior outcome. The experiments consequences shown are for accuracy of classification for doing analysis of cost. The consequence in the research on datasets also shows that the efficiency and accuracy of j48 and Naive bayes algorithm is superior. [4]

Chengjun Zhan et al. (2011) used a multivariate decision tree algorithm for a request of problem solving related to lane clearance. Lot of research is being attempted for prediction of time of incident clearance. It is being analyzed that due to lane blockage congestion occurs therefore to forecast lane clearance time. Rather than incident clearance time will be profitable. Earlier no model for prediction of lane clearance was developed. After a long research an algorithm is being developed for solving the problem of incidents the algorithm is being used for clearance of lane, the algorithm was M5P. M5P algorithm is advantageous as compared to other algorithms. It is used for both categorical and continuous variables and for missing values. There are so many variables (like quantity of vehicles, number of sterile lanes etc.) due to which clearance time is being affected. Model developed is when compared to conventional models then it is concluded that this model gives superior prediction outcome.

Today problem of congestion is major issue for everyone as a result lot of time is wasted. To predict the delay and length of queue blockage time is most important aspect. This time period is a function of various factors like type of incident, the quantity of blocked lanes. Due to bad weather conditions also incidents probability may also get increased. Management of incident is being done at Traffic Management Centers (TMCs). So many algorithms and data mining techniques are applied for prediction of lane clearance but it is being concluded that no algorithm is 100% accurate but M5P gives better results as compared to all other algorithms utilized earlier. In the research paper read a novel approach is designed and utilized for development of model for prediction of lane clearance. The approach uses the M5P algorithm for prediction of lane clearance time, this algorithm is beneficial as compared to other algorithms as it can be used for both categorical and continuous variables and it can also missing values problem. M5P algorithm is best compared to conventional algorithms [5].

Venkatadri. M et al. (2010) presents usually used classification algorithms and those are: neural networks, decision trees etc. Classification trees are widely used technique. In the research paper relative analysis is done of a variety of decision tree algorithms. Decision tree divides a data set of records using depth first greedy or breadth first approach.

There are 2 stages to perform classification and those are:-

- Tree Growing (or building): In this top down approach is followed, tree is iteratively partitioned till all data objects fit in to the same class label.
- Tree Pruning: In this bottom approach is followed, full grown tree is cut reverse to avoid over fitting and develop accuracy of tree.

Comparison is done among a variety of algorithms and it is concluded that SPRINT and Random Forest algorithms have good quality accuracy as compared to ID3, BFTree , C4.5, BEST FIRST TREE, SPRINT ,CART, SLIQ. [6]

C. Deepa et al. (2010) designed Tree Based Modeling. She discussed that compressive power of the elevated performance and she used classification algorithms like Multilayer Perceptron, M5P Tree models and Linear Regression for tree creation. The conclusion can be finished that tree based models present fine i n strength forecast of concrete mix. For building the model time taken is far above the ground when compared to other algorithms. Linear regression takes less time to construct the models. In M5P model Error rate is less than other two algorithms. When compared to MLP and Linear Regression algorithm M5P tree algorithm is much improved. The supervised learning issue of the MLP can be solved with an algorithm called back-propagation algorithm. The model tree algorithm used in this task is based on M5P. To create relationship between two variables linear regression is used, it attempts to well a linear equation to examine data. The predicted far above the ground strength values are not a lot superior to the investigational compressive strength values in the real time data set. As compared to other algorithms M5P algorithm has revealed the lowest RMSE. It also has elevated association among the other algorithms used for model tree generation. M5P algorithm is beneficial as compared to other algorithms. It is used for both categorical and continuous variables and for missing values [7].

III. Current Work

3.1 Problem Formulation

Many algorithms are already designed for generation of classification tree. Different algorithms have different efficiency or accuracy issue. Accuracy of algorithm differs from application to application; it may be possible that an algorithm is accurate in one but not much accurate in another area. Analysis of various algorithm was done by which a proposal came into mind that a new algorithm design will improve performance. Classification tree had 2 approaches: Univariate and Multivariate decision tree. Both decision trees have different algorithms for implementation purpose. They may have some issues like in Univariate noise of data remains in it. For that issue Multivariate decision tree's algorithm can be used.

In novel algorithm('SPruning Algorithm') approach, some features of Univariate algorithm i.e. CART is being used like CART algorithm is used for clustering of data, classification of data is also done using CART algorithm and some features of Multivariate tree's algorithm i.e. M5P is used like M5P is used for searching. New algorithm includes one enhanced feature due to which less time is taken as compared to existing algorithms and that feature is pruning. Pruning was performed in earlier designed algorithms also but new algorithm prunes the data in more accurate manner. Earlier pruning of data was done but in this new approach pruning of files is being due to which user's time is saved. As a result of doing this performance of classification tree can be improved when compared to earlier implemented algorithms. Less time will be taken by new algorithm and complexity will be improved. Data is being pruned and clusters of data are formed into different files and user can access particular data according to his need due to which he can save his precious time and money.

3.2 Objectives

An algorithm is implemented which is better than other algorithms already designed for decision tree generation. Existing algorithms have many issues regarding accuracy of data after mining of the data which creates problems for an organization and for solving those problems an algorithm is designed. This novel algorithm solves issues of existing algorithms. It includes one enhanced feature due to which improvement in performance is done. Pruning feature is improved in new algorithm as a result user time and money is saved.
Algorithm designed will be used:

- To enhance some features of conventional algorithms
- To improve performance
- To combine some features of existing algorithms
- To classify the files

Complexity of novel algorithm is calculated in this research.

3.3 Research Methodology

Data is collected and this raw data is mined for knowledge extraction. Then decision tree are generated using tree generating tools and different mining techniques (algorithms are also applied for generation of decision trees) are to be applied on data and trying to extract some novel, previously unknown knowledge in order to improve decision making process. Raw data is converted to knowledge discovery in the form of well structured graphical representation known as decision trees. We try to explore how to improve the performance of decision tree for that a novel approach of algorithm is implemented which gives better results as compared to previously designed algorithms of decision trees. Preprocessing of that data is done i.e. data is stored in different formats of files (pdf, text, doc, excel).After that on the processed data novel algorithm is applied which is blend of some features of two algorithms plus some additional features. User doesn't have to waste time by searching whole data therefore to save time pruning feature is enhanced. Earlier pruning of data was done but in novel approach pruning of data files is done by which accessing of only particular file which is needed is done. Using novel algorithm classification of files is done i.e. files are classified into different panels according to the type of file. Then user can search any particular word from any particular file i.e. pruning of files is being done. Pruning of files saves user time and money as a result complexity improves. Decision tree for explaining searching concept are also drawn. Searching concept is used in classification of data files and in word searching which is explained through decision tree representation.

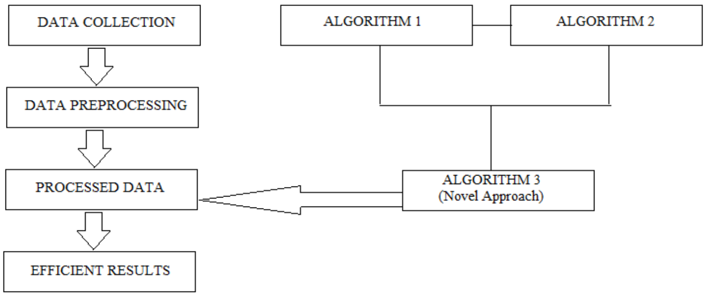


Fig 3.1: Flowchart of proposed work

Above flowchart depicts the flow of data i.e. after collection of data, preprocessing i.e. storage of data in different data files is being done and on that processes data a novel algorithm is applied which improves the complexity than conventional algorithms.

In traditional algorithms pruning of data was done which was not that much time saving approach as user have to access all the data but by using novel algorithm pruning of data files is done by which time complexity improves as user can search only required document and can generate a separate decision tree rather than decision tree of whole data.
Features of novel algorithm:-

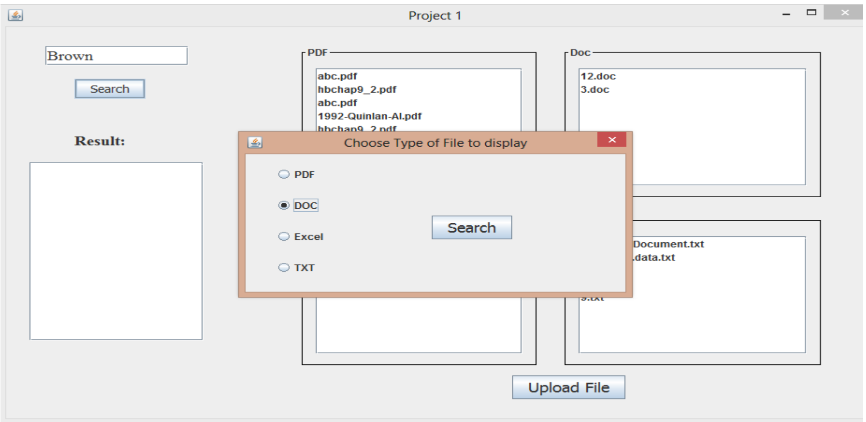
- Classification: Classes are generated of data files i.e. data in pdf files is stored in pdf class, text files are stored in txt class and similarly for all types of files
- Clustering: When user search any word then searching is pruned to clusters i.e. user can search data from only a particular cluster required which saves time and money.
- File pruning: Earlier pruning of data was done which was not that much efficient as in traditional algorithm whole data is taken once and on that data pruning is done. That approach takes time but in novel approach pruning of data files is done i.e. user can search particular word from a particular type of file rather than searching whole data.

- Tree generation: Decision trees are generated to explain the concept of searching used in this new algorithm. Searching is done while doing classification and in searching a word process. So to explain how a word is searched and how class of file is searched decision trees are generated. [8]

IV. Implementation Work

Novel approach is blend of different features like:

- **Classification of files**: Classification of files is done by new algorithms. Files of different types are stored in their respective class i.e. a pdf file stores in pdf class, text file in text class and so on.
- **Enhanced Pruning in searching method**: In former algorithms when users have searched anything then they have to search from whole data and to go through each and every file which is time consuming. So to remove that disadvantage enhanced method of pruning is introduced in novel algorithm. In novel approach when user have to search any word and if he know that word is in pdf file then instead of searching all the files user search only specific cluster i.e. pdf file as a result time is saved as compared to earlier algorithms.



Brown

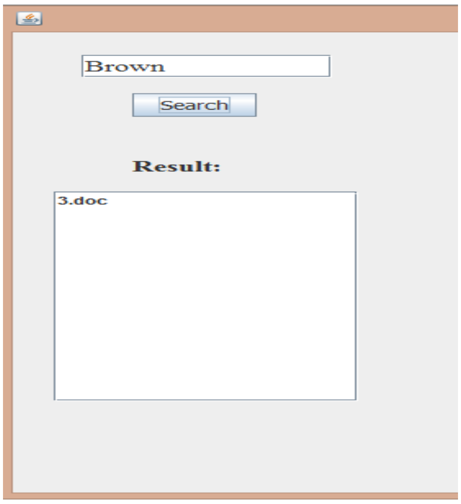
PDF
abc.pdf
hbchap9_2.pdf
abc.pdf
1992-Quinlan-AI.pdf
hbchap9_2.pdf

Doc
12.doc
3.doc

Document.txt
.data.txt
9.txt

Choose Type of File to display
○ PDF
● DOC
○ Excel
○ TXT

Figure 4.1: Pruning of files



Brown

Result:
3.doc

Figure 4.2: Result after pruning of files

As shown in above snapshots of pruning when word to be searched is entered in Textbox and search button is clicked. Various clusters are shown and user makes his search specific by clicking any one of the required type. As a result time does not waste and task is done easily. Only pdf files which contain particular word are shown in result window.

- **Tree generation (explaining searching method)**:

Decision trees are generated which explains the concept of searching. Searching method is used in classification as well as in searching any word from file.

In case of classification: Searching is done from last character i.e. last character of file extension is matched with the last character of class name and when extension name matches the class name then file will get stored in that particular class. For E.g. suppose file name is s.pdf and class names are pdf, doc etc. then last character of extension is matched first with last character of all class names. It matches with last character of pdf class, similarly next character is searched and finally if extension name matches class name i.e. pdf then file get stored in that class.

In case of searching a word from file: Searching concept of word searching is similar to concept of searching in classification as explained above i.e. searching is done from last character to first character.

Snapshots of Decision Trees generated for explaining concept of Searching are:-

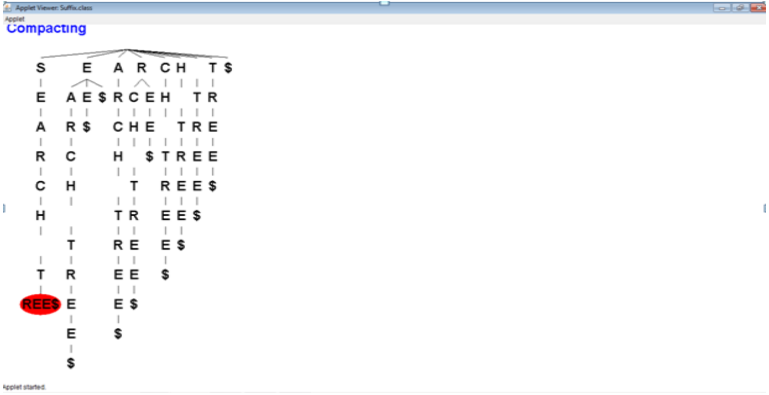


Figure 4.3 Decision Tree (a)

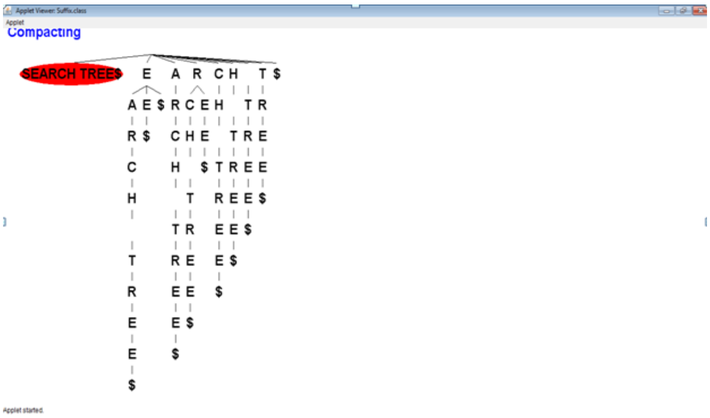


Figure 4.4 Decision Tree (b)

V. Conclusion

In the research work a novel algorithm (SPruning algorithm) is designed which is better than earlier designed algorithms as it improves the performance. In this new approach enhancement is done in some features. There is combination of some features of existing algorithms plus some enhanced features of new algorithm. Due to enhancement in features new algorithm works quickly. In SPruning algorithm classification of files is done i.e. pdf files are placed in pdf class, doc files are in doc class and so on. Searching concept is used for storage of files in their respective class i.e. when class name is similar to extension of file then that file get stored in that class for that searching is done. Pruning concept is enhanced in this novel approach proposed as earlier pruning of data is done but in this novel approach pruning of data files are done as a result user can access particular file rather than accessing whole data. Due to this time is saved. When user wants to search any word from data and he knows that word is in one of the pdf file then he can prune his search to pdf files and search the required word. Decision trees are generated to explain the concept of searching. By this approach time performance and efficiency is enhanced as compared to conventional algorithms.

Enhancement by adding some new features in novel algorithm can be done in this research work in future. Pruning concept can be modified such that less time will be taken for searching.

References

[1]. S.Anupama Kumar," A Naïve Based approach of Model Pruned trees on Learner's Response*", I.J.Modern Education and Computer Science, Vol. 9, pp. 52-57, 2012*

[2]. W. Nor Haizan W. Mohamed , " A Comparative Study of Reduced Error Pruning Method in Decision Tree Algorithms", *IEEE International Conference on Control System,Computing and Engineering, pp.23 - 25 Nov. 2012*

[3]. Dr. Neeraj Bhargava, "Decision Tree Analysis on J48 algorithm for Data Mining", *International Journal of Advanced Research in Computer Science and Software Engineering, June 2013*

[4]. Goyal Anshul, "Performance Comparison of Naïve Bayes and J48 Classification Algorithms*", International Journal of Applied Engineering Research Vol. 7, no. 11, 2012*

[5]. Chengjun Zhan, "Prediction of Lane Clearance Time of Freeway Incidents Using the M5P Tree Algorithm", *IEEE transactions on intelligent transformation systems, Vol. 12, no. 4, December 2011*

[6]. Venkatadri .M," A Comparative Study on Decision Tree Classification Algorithms in Data Mining", *International Journal of Computer Applications in Engineering, Technology and Sciences, April-September 2010*

[7]. C.Deepa," Prediction of the Compressive Strength of High Performance Concrete Mix using Tree Based Modeling", *International Journal of Computer Applications Vol. VI, no. 5, September 2010*

[8]. Srishti Taneja," Hybrid Approach for Classification Tree Generation", *International Journal of Emerging Trends & Technology in Computer Science Volume 3,Issue 1,February 2014*

[9]. http://www.slideshare.net/asad.taj/top10-algorithms-data-mining

[10]. Du Zhang & Jeffrey J.P. Tsai, Advances in Machine Learning Applications in Computer Engineering(2000)