International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-09

IMPROVED APPROACH AND IMPLEMENTATION FOR DECISION TREE CLASSIFICATION USING REDUCED ERROR PRUNING

Shefali Chopra
Research Scholar
Department of Computer Science & Engineering
Doon Valley Institute of Engineering & Technology
Karnal, Haryana, India

Heena
Assistant Professor
Department of Computer Science & Engineering
Doon Valley Institute of Engineering & Technology
Karnal, Haryana, India

ABSTRACT

Decision trees are one of the most profound researched domains in Knowledge Discovery. Regardless of such advantages as the ability to explain the choice procedure and low computational costs, decision trees also generally produce relatively great outcomes in estimation with other machine finding out formulas. Although the best decision tree induction algorithms, such as J48, had been developed a time ago, they continue to be regularly used for solving everyday classification tasks. In this work our aim is to improve the predictive performance of these algorithms by diminishing their three major disadvantages by Pruning Trees. Pruning is a technique in device reading that reduces the dimensions of decision trees by separating parts of the tree that provide little control to categorize instances. Pruning decreases the complexity in the final classifier, and thus improves predictive accuracy by decreasing the over fitting. The key point in construction of decision tree is the choice of the best attribute to split the considered node. The main work done by this research work is it focuses on reducing the complexity of the tree. The complexity of the tree

is reduced by pruning the Decision Tree. During the pruning process the size of the Decision Tree will get reduced. With the help of pruning the entropy or randomness of error will also reduces which inturn increases the accuracy of the Decision Tree.

With the help of this method, complexity of decision tree model is diminished by “reduced error pruning method” and the error arising from the variance is also reduced. Pruning reduces the complexity and thus improves the predictive accuracy by the reduction of that complexity.

Keywords - Data Mining, Decision Tree Classification, Error Pruning

INTRODUCTION

Today the production, engineering, corporate and private associations concerning the region and the world are produced in large numbers of computing processes data. The explosive growth of data and the data to digest skills has outpaced clear. Therefore, methods and tools for automatic data scrutiny invention excavating and vision needs the data. Today's enterprise data warehouse (EDW), increasing proliferation, and tools and methods for data scrutiny vision to include knowledge discovery, data modeling, algorithms, and focuses on visual aspects. Vision data to highlight the connection and construction customers, suppliers, and greater understanding of the inner as well as external processes will be able to get by. This, to identify failures in defect reduction, and increased costs, helping to continuously improve the quality process helps proprietors.

THE DATA MINING PROCESS

Data excavation is depicted as "the non trifling extraction of innate, previously new and conceivably utilitarian data from data. Be that as it may, expelling useful frameworks from data is only the initiating of an iterative technique to loop data into data, the data into vision, and the vision vigorously.

These steps are as follows:

[1] The target data is selected based on end-user goals and understanding of subject area and the number of variables.

[2] This target data now have to be preprocessed to clean up the bad data, decide on missing data, and handle time sequencing and slowly changing data.

[3] The data transformation step allows an effective number of variables to be created for data mining purposes.

[4] Then, it must be decided whether the goal of the knowledge discovery process is classification, regression, clustering, and so on. An appropriate data mining algorithm is chosen for searching for patterns in the data.

[5] The end user has to interpret the returned pattern and evaluate the knowledge obtained by the data mining process.

[6] It is possible to conduct further iteration based on the interpretation of mined patterns.

[7] Finally, the new knowledge has to be integrated into previous knowledge to resolve any conflict and inconsistency.

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-09

DECISION TREES

Contraption choice tree learning ranges are among the most widely examined. Basic leadership abilities and low computational expense, to clear advantages, for example, choice trees, in addition to extra supplementary general inquiry calculations contraption reasonably great relationship with the after generation. Despite the fact that the acknowledged choice tree affectation calculations, truck, for a brief period before the business was, they still habitually used to settle ordinary errands are Association.

Reduced Performance when the Training Set is little. Little example sizes represent a noteworthy test to choice trees, specifically in light of the fact that the quantity of accessible preparing occurrences drops exponentially as the tree limbs out (and the quantity of leaves is limited by the preparation set size). Thus, if the preparation set is too little, the actuation calculation may grow an excessively shortsighted arrangement tree.

Rigid Decision Criteria: Each level of the tree (hubs) on the choice implies that stand out branch (hub) can be chosen (unless plenty trademark illustration, which is a different issue that is not under dialog here is of no worth) is inflexible. This methodology typically functions admirably, however consider the accompanying situation: the tests performed on the trademark X 6 to 10, and x estimation of the properties of a given illustration is 9.99. In such cases, not minimum the likelihood that we might not be right order, or consider an option way ought to consider? This issue is not new and, by utilizing the idea of edge (delicate edges) was examined around 20 years back. Other characterization calculations, unbending nature issue publicizing has been readied. An extraordinary case is exhibited by the calculation so as to better bolster vector machines when a specific hyper plane areas utilizing as opposed to isolating, characterizing dataset utilizes shifting edges.

Outlier Attribute Values:Deep choice tree comprising of a few levels may likewise incorporate on-the tribute. Such wild trees are frequently seen when the preparation set is substantial as far as the quantity of occasions and characteristics. To achieve an order can prompt the accompanying issues need to depend on such a large number of qualities: the arrangement procedure to wreck an anomaly esteem (one that is strange of its class) with just takes a characteristic.

These issues essentially lopsided datasets, where a specific class than another way to deal with see numerous more in wage are developing. In such cases, the greater part class standard dominant part can overpower chine learning systems, along these lines have a circumstance where the minority is overlooked building. For instance, consolidated with an absence of adequate number, exception values it significantly more hard to sort precisely speak to the areas. Keeping in mind the end goal to address the unequal dataset choice trees empower the two practices is frequently utilized:

1. Balanced characterization assignments as opposed to utilizing as leaf arrangement, grouping connected with the better dissemination of uneven datasets assigned leaf is to utilize. This methodology truly the best exchange off amongst positive and

false-positive execution to help the chief to pick, as indicated by their potential outcomes can be utilized to rank test illustrations.

2. The actuality that the grouping much of the time uneven dissemination of capacities is utilized, it is prescribed to abstain from pruning. Along these lines, the order tree unequal activities commonly experience the ill effects of more elevated amounts et cetera. Like uneven datasets, multi-level characterization issues likewise represent a danger to trees, particularly when the quantity of classes goes past an unassuming size. Truly multi-class works, its actual leaf from crashing a misclassification will bring about a test occasion can be clarified by. This adjusted paired order capacities, which leaves a false and arbitrary populace of wrecking still a plausibility, equivalent to the rate of things with the class may bring about a right arrangement is not a good fit for.

DECISION TREESAS A CLASSIFIER

Unearthing data, a choice tree is a prescient model that both classifiers and relapse model can be utilized to exemplify. Scutiny forms, the supplementary hand, choice trees choices and their outcomes mean a various leveled model. Chiefs to get a handle on the procedure expects to distinguish the doubtlessly choice trees are utilized.

A choice tree is utilized for the operations of the Union, it is more proper to the tree is set apart as a union. After it is utilized for relapse capacities, it is yelled relapse tree.

The choice tree, as tree hubs that are embedded, which implies that it is a fruitful one with a hub tree root is yelled that no approaching edges. Every single supplementary hub precisely one approaching edge". A hub with the active fringes an "inward" or "Test" hub is checked. Every single supplementary hub "leaves" are yelled (likewise "Terminal" or "choice" perceived as hubs). In the choice tree, each interior hub quality estimation of the speculation objective, precise discrete spaces into two or more sub-partition the space in the case. The least difficult and most intermittent case, every one highlighting a lone Test, for instance, is separated by worth space qualities gets it. Regarding numerical properties, the condition is alluded to an arrangement.

Every leaf of a class to speak to the most fitting target quality is apportioned. On the other hand, a potential vector leaf (vector similitude) to the objective property estimation showing the probability of an exact sense. A choice tree is regardless of whether the explanations behind a potential client to deal with a mailing will answer depicts another illustration. Internal hubs, epitomized as circles, while the leaves are set apart as triangles. Two or more divisions, each interior hub (ie, not a leaf) can create. Every hub compares with a precise portrayal and divisions are steady with the extent of qualities. The element has the upside of the advantages of these extensions is to set up a division.

According to the results of examinations with examples of the path, from the root of a tree down to the leaf by passing are classified. In particular, we

start with the basics of a tree; We attribute that corresponds to a route to consider; And we have to split attribute corresponds to portray eminent worth. Next we have to consider that the division node appears.

THE HIERARCHICAL NATURE OF DECISION TREES

Another characteristics is their hierarchical nature of decision trees. Imagine that you, according to countless patients after treatment to identify the desire to develop a health system testing. Founded on the results of an examination, the doctor may order tests present or supplementary workshop.

Classifiers manufactured data mining frameworks that are ordinarily utilized as a part of gadgets. Such frameworks takes contribution as the accumulation of cases, each identifying with one of a little number of classes and their qualities to a specific arrangement of attributes portrayed, and a classifier to deliver another class that is correct the case can foresee which has a place with snatch.

These notes portray C4.5 , a relative of the CLS and ID3 . Like CLS and ID3, C4.5 classifiers produces communicated as decision trees, additionally more understandable as classifiers can assemble ruleset. We will diagram C4.5 calculations utilized, his successor C5.0 highlighted some progressions, and open examination closes with a few issues.C4.5 calculation to turn the data sharing that data produces a decision tree. Utilizing profundity first technique decision tree develops.

C4.5 calculation to test all the conceivable data that is partitioned and can have a test that gives data that best comprehends chooses. This test of ID3 inclination for vast decision trees evacuated. Each discrete characteristic, as the quantity of various advantages to a test highlight to endless after creation are utilized. Each nonstop characteristic, the data is sorted, and a sweep of the data determined entropy pick up at an alternate cost in every parallel introduced on the cut is figured. This procedure proceeds for all elements is recapped. C4.5 decision tree calculation for sorting developing licenses. This expands the mistake rate on the preparation data, yet indispensably, evaluate data on blunder rates cut disregarded. Further numerical properties C4.5 calculation, missing benefit, and can manage hard data. The quest for benefit and misfortune:

Advantages:

1) C4.5 can handle both continuous and discrete attributes. In order to handle continuous attributes, it creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
2) C4.5 allows attribute values to be marked as? For missing. Missing attribute values are simply not used in gain and entropy calculations.
3) C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

Disadvantages:

1) C4.5 constructs empty branches; it is the most crucial step for rule generation in C4.5.We have found many nodes with zero values or close to zero values. These values neither contribute to generate rules nor help to construct any class for

classification task. Rather it makes the tree bigger and more complex.

2) Over fitting happens when algorithm model picks up data with uncommon characteristics. Generally C4.5 algorithm constructs trees and grows it branches 'just deep enough to perfectly classify the training examples'.
3) Susceptible to noise.

J48

J48 C4.5 calculation is the execution of an open premise. There are two approaches to bolster sorting by J48 soon perceived as a sub-tree choice you may be, the leaf hubs in the tree with the decision to supplant the work with the precise way of the first as diminishment in the quantity of test. This leaves the course toward mechanical tree finished and a backward assessment from the begin with the procedure works. Like the J48 after the solicitation at the highest point of the tree toward the source and out and about, for example, the expansion of a supplementary sub-tree hubs is expanding by substitution propelled hubs. This calculation is partitioned on the appraisal data and a test dataset that will give the best results in the advantages of the data determination can be stressed by. Discrete properties also, these calculations uneven advantages and every one component to test double element number as some of an outcome with a test thought uneven benefit development to every one element considered won't go on without serious consequences.

Ruleset classifiers

To comprehend the mind boggling decision trees, for instance, since data concerning a class is ordinarily disseminated all through the tree can be troublesome. C4.5 as an option formalism ... the following class An and B and C and X contains a rundown of laws, laws for each one class, while all the while gave are gathered. Ahead of schedule for this situation by case law whose conditions are satisfied by the disclosure of grouped; If no law is delighted, the case has been alloted to a default class.

C4.5 rulesets early (unpruned) decision trees are modern. Follows the cause of each of the leaf trees along the trail after a model law and the states of the class name leaf falls. This law, thus, disposing of the aftereffects of the discovery of the state of each of the accompanying is clear. Dropping a condition secured by law in various cases could rise N, and for cases that don't fit the classification of law selected by the expansion number e, and negative blunder rate is less driven as the above. The most cynical of a slope climbing calculation mistake rate is found to drop conditions are utilized.

To complete the procedure, clear laws, a subset of the class is chosen for every one thusly. The square subset preparing matters are masterminded to lessen mistake and a default class is chosen. A definite choice ruleset ordinarily pruned the quantity of leaves on the trees are not exactly the law.

Loss of C4.5 rulesets important measure of CPU time and memory that they require. In one test, extending from 10,000 to 100,000 cases, tests were drawn from an expansive dataset. For decision trees, 10 100K moving instances of CPU time on a PC 44. The expansion of 1.4 61 s, a factor essential for rulesets time, notwithstanding, expanded from 300 a factor of 32 9715-S.

C5.0

C4.5 was superseded in 1997 by a commercial system C5.0 (or C5.0 for short). The changes encompass new capabilities as well as much-improved efficiency, and include:

1) A variant of boosting , which constructs an ensemble of classifiers that are then voted to give a final classification. Boosting often leads to a dramatic improvement in predictive accuracy.
2) New data types (e.g., dates), not applicable values, variable misclassification costs, and mechanisms to pre-filter attributes.
3) Unordered rulesets—when a case is classified, all applicable rules are found and voted. This improves both the interpretability of rulesets and their predictive accuracy.
4) Greatly improved scalability of both decision trees and (particularly) rulesets. Scalability is enhanced by multi-threading; C5.0 can take advantage of computers with multiple CPUs and/or cores.

RESEARCH ISSUES WITH C4.5

We have over and again communicated Associates that decision trees are listened, the issue is unraveled. We don't concur with this proposition, and will close with two or three open scutiny issues.

We have more than once communicated Associates that decision trees are listened, the issue is explained. We don't concur with this proposition, and will close with a few open scutiny issues. Stable tree. It is very much perceived that the blunder rate of a tree on the cases that it was built (Resubstitution blunder rate) on the inconspicuous matters mistake rate (prescient blunder rate) is not exactly. For instance, with 20,000 cases a surely understood at Credit dataset, C4.5 resubstitution blunder rate to 4%, however the forget one (20000 times) cross-acceptance mistake rate of 11.7% is. As this illustrates, out of 20,000 cases a lone tree that is built which regularly influences!. Today, we gauge that a non-inconsequential tree-building calculation that once in a while was supplanted by a lone aside from may create. For the steady trees, resubstitution forget one cross-accepted mistake rate blunder rate around the tree consultancy that immaculate size ought to have.

Complex tree disintegrating. Gathering classifiers, improve quality, sacking, measuring randomization, or supplementary techniques, produced by the prescient exactness regularly expanded. Presently, given that a little number of decision trees a single (exceptionally intricate) tree is precisely equivalent to the yield of the tree early races are conceivable, however we can be correlative way? That is, a mind boggling tree is a little accumulation of straightforward tree that in the wake of being chosen mutually, as similarly complex tree can be separated to convey results? Such deterioration conceivable decision trees ought to be helped with the production of remarkable.

LITERATURE SURVEY

Gilad Katz, 2014, ConfDTree The test study in this paper demonstrates that the proposed post-handling technique reliably and essentially enhances the prescient execution of decision trees, especially for

little, imbalanced or multi-class datasets in which a normal change of 5%»9% in the AUC execution is accounted for.

Brijain R. Patel et al., 2014 In this paper,data unearthing is the system of finding or expelling new diagrams from giant data sets including strategies from measurements and artificial insight. Affiliation and figure are the techniques used to make out crucial data classes and gauge likely pattern.

Michal Wozniak et al., 2014 In this paper, a present center of extreme scutiny in framework affiliation is the mix of incalculable classifier game plans, that can be made seeking after whichever the alike or dissimilar models and/or datasets building approaches. These plans present data blend of affiliation decisions at unique levels vanquishing constraints of built up routes set up on single classifiers. This paper introduces a state-of-the-art study on a few classifier courses of action (MCS) from the purpose of consider Hybrid Intelligent Systems.

Delveen Luqman Abd et al., 2013 In this paper, a similarity in the midst of three grouping's calculations will be found out, these are (K-Nearest Neighbor classifier, Decision tree and Bayesian system) calculations. The paper will clear up the quality and precision of each and every calculation for relationship in expression of presentation proficiency and period many-sided nature required. For perfect approval energetic, twenty-four-month data examination is driven on a false up premise.

Dursun Delen et al., 2013 In this paper,ascertaining the steady presentation utilizing an arrangement of business measures/proportions has been a fascinating and testing misfortune for endless analysts and professionals. Distinguishing proof of factors (i.e., business measures/proportions) that can decisively conjecture the steady presentation is of extraordinary consideration regarding every decision creator.

Nirmal Kumar et al., 2013 In this paper, LCC of an earth diagram constituent is sought after for manageable use, affiliation and protection hones. Hoisted speed, raised exactness and simple delivering of laws by contraption finding calculations can be used to create pre-characterized laws for LCC of soil diagram constituents in developing decision prop plans for earth use masterminding of a territory.

Leszek Rutkowski et al., 2013 In this paper, in uncovering data streams the most acknowledged instrument is the Hoeffding tree calculation. It utilizes the Hoeffding's joined to find out the littlest number of cases requested at a hub to choose an isolating property. In works the alike Hoeffding connected was used for every assessment reason (heuristic measure), e.g. data addition or Gini list. In this paper it is demonstrated that the Hoeffding disparity is not suitable to determine the hidden issue.

Richa Sharma et al., 2013 In this paper, a try has been settled on to build up a decision tree affiliation (DTC) calculation for relationship of remotely distinguished satellite data (Land sat TM) utilizing open premise support. “The decision tree is made by recursively apportioning the unearthly distribution of the preparation dataset utilizing WEKA, open premise data uncovering programming. The sorted picture is differentiated nearby the photo arranged utilizing established ISODATA grouping and Maximum Likelihood Classifier (MLC) calculations.

Anuja Priyama et al., 2013 In this paper, at the present period, the quantity of data put away in instructive database is rising quickly. These databases incorporate concealed data for upgrade of understudy's execution. Relationship of data items is a data exhuming and vision affiliation technique used in social affair practically identical data protests together. There are incalculable affiliation calculations realistic in works however decision tree is the most typically used due to its simplicity of executing and less demanding to fathom differentiated to supplementary affiliation calculations.

Susan Lomax et al., 2013 In this paper, in the most recent decade there has been rising custom of data uncovering strategies on wellbeing data for finding utilitarian patterns or diagrams that are used in analysis and decision making. Data unearthing strategies, for example, grouping, affiliation, relapse, affiliation law exhuming, CART (Classification and Regression Tree) are broadly used in medicinal services area.

A.S. Galathiya et al., 2012 In this Research work, Analogy is made in the midst of ID3, C4.5 and C5.0. In the midst of these classifiers C5.0 gives additional exact and viable yield nearby modestly raised pace.

Raj Kumar et al., 2012 In this paper, affiliation is a perfect finding strategy that is used for parceling the data into unique classes as indicated by a little compels. In supplementary words they can say that affiliation is method of summing up the data as indicated by divergent cases.

Rodrigo Coelho Barros et al., 2012 In this paper, exhibits a study of developmental calculations anticipated for decision tree impelling. In this connection, the majority of the paper concentrates on ways that develop decision trees as a substitute heuristics to the built up top-down tear and-vanquish approach. Furthermore, they exhibit somewhat elective techniques that make utilization of transformative calculations to improve specific constituents of decision tree classifiers.

Smith Tsang et al., 2011 In this paper ,founded decision tree classifiers work close by data whose advantages are perceived and exact. They spread such classifiers to handle data nearby speculative data.

J. R. Otukei et al., 2010 In this paper, earth spread change evaluation is one of the primary solicitations of remote identified data. Various pixel built up affiliation calculations have been industrialized over the previous years for the scutiny of remotely distinguished data. The most striking contain the greatest probability classifier (MLC), prop vector components (SVMs) and the decision trees (DTs).

Kalpesh Adhatrao et al., 2009 In this paper, an instructive affiliation needs an inexact earlier vision of selected understudies to figure their presentation in up and coming scholastics. This helps them to perceive enthusing understudies and also gives them a chance to wage thoughtfulness regarding and upgrade the individuals who ought to conceivably get to be lower grades. As a determination, they have industrialized a course of action that can conjecture the presentation of understudies from their first presentations utilizing musings of data exhuming techniques underneath Classification.

Thair Nu Phyu et al., 2009 In this paper affiliation is a data unearthing (machine learning) technique used to figure bunch enrollment for data occasions. In this paper, they show the straight to the point affiliation systems. Endless primary sorts of affiliation strategy enveloping decision tree incitement, Bayesian networks, k-closest associate classifier, case-based thinking, hereditary calculation and hairy rationale systems. The point of this study is to outfit an exhaustive investigation of divergent affiliation techniques in data mining.

Matthew N. Anyanwu et al., 2009 In this paper, relationship of data articles set up on a predefined vision of the items is a data exhuming and vision affiliation technique used in social event similar data questions together. It can be depicted as directed finding calculations as it doles out class marks to data objects built up on the association in the midst of the data things close by a pre-characterized class name.

OBJECTIVES AND PROPOSED OUTCOME

1. To study various clustering and classification algorithm in data mining and their application in data mining.
2. To create a REP Tree.
3. Reduced the complexity of Decision Tree.
4. To detect the missing values and study the imbalance issue.
5. To reduce the size of tree using REP Tree and Confidence Estimation.
6. To reduce the entropy and absolute error.
7. To increase the accuracy of the tree by using REP Tree.
8. To compare the performance of proposed algorithm with existing Decision Tree algorithm.

POST PRUNING ALGORITHM

Tree pruning is completed in order to attain tinier trees and circumvent over-fitting (the algorithm attempts to categorize the training data so well and it becomes too specific to accurately categorize the examination data).

Input**: C4.5 or J48 Decision Tree **T

```
Procedure PostPruning(Data, TreeSplits)
SplitData(TreeSplits, Data, GrowingSet, PruningSet)
Estimate = DivideAndConquer(GrowingSet)
loop
        NewEstimate = Selection(Estimate,PruningSet)
        if Accuracy(NewEstimate, PruningSet) < Accuracy(Estimate,PruningSet)
                exit loop
        Estimate = NewEstimate
return(Estimate)


Procedure DivideAndConquer(Data)
Estimate = Ø
while Positive(Data) != Ø
        Leaves = Ø
        Instance = Data
  while Negative(Instance) != Ø
        Leaves = Leaves ∪ Find(Leaves, Instance)
        Instance = Instance(Leaves,Instance)
  Estimate = Estimate ∪ Leaves
  Data = Data - Instance
return(Estimate)
```

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-09

We tear adaptation representative maintained a frank and law framework for the win. A generating set and a pruning set: initial training data subset to tear in two. In the early period, no attention to data that is considered sound and a full description of positive and negative examples of covers have shown any production set is paid for. Guesstimate voraciously emerging from each guesstimate more literals to deletion and remove laws result in a reduction of predictive accuracy as measured on pruning should set clear next time. Simplification of the current estimate of the set of selection sort subroutine sets selects guesstimate with highest accuracy. Simplification that normally are attempting to remove an entire section, or a section of the final factual removed. Representative variants Additionally, the removal of a section every factual literals or a literal removal of the last scene I like the best choice to maintain operators can supplement simplification. So the best accuracy of simplification is not un-pruned down to the accuracy of the estimate, the new estimate will tolerate Cutting representative. It is best pruned guesstimate recapped the accuracy is below that of its predecessor.

Algorithm Confidence for attributes

Input: *Node node: a node in the tree*

1: IF IsLeaf (node) ←true or node. Size < m THEN RETURN

2: split attribute ← node. split attribute

3: node.num of instances ← Get Num of Instances per ClassID (node)

4. FOREACH (Split Attribute Value ai)

5: FOR (i = 0; i <Get Num of Att Values (split attribute); i + +)

6: FOR (j = if + 1; j <Get Num of Att Values (split attribute); j + +)

7. cj1 ← Get Attribute Value (split attribute, i)

8: cj2 ← Get Attribute Value (split attribute, j)

9: Percof instances with Val2 ←Get Num of Instances with Val (cj1; ai)

10: perc of instances with Val 2 ← Get Num of Instances with Val (cj2; ai)

11: IF (perc of instances with Val 1 == 0 && perc of instances with val 2 == 0)

12: CONTINUE

13: ELSE

14: IF (Attribute Proportion Dierect In Class (a_1, c_j1, c_{j2}))

15: IF (Proportion (cj1, ai) > Proportion (c_{j2}; $_{ai}$))

16: Mark as Superior (c_{j1}; c_{j2}; $_{ai}$)

17: ELSE

18: Mark as Superior (cj2; c_{j1}; ai)

19: END FOR

20: END FOR

21: END FOR

ACCURACY MEASURES

Accuracy class alongside an unbalanced allocation count is not sufficient to assess the model. There are estimated to an accuracy rate , while matters concerning the quality of a derivative classifier one can mislead. In such conditions, the dataset significantly more than the wholesale segment of the

minorities, including examples, the bulk of class selection and accuracy can achieve good performance. Therefore, in these cases, measures of sensitivity and specificity to measure accuracy can be used as an alternative. "Sensitivity (also known as memory) to assess how well the classifier to identify positive samples, and is defined as

$$Sensitivity = \frac{true\_positive}{positive}$$

where true positive corresponds to the number of the true positive samples and positive is the number of positive samples.

Specificity measures how well the classifier can recognize negative samples. It is defined as

$$Sensitivity = \frac{true\_negative}{negative}$$

where true negative corresponds to the number of the true negative examples and negative the number of samples that is negative.

Another well-known performance measure is precision. Precision measures how many examples classified as positive class are indeed positive. This measure is useful for evaluating crisp classifiers that are used to classify an entire dataset. Formally:

$$Accuracy = Sensitivity . \frac{positive}{positive + negative} + Sensitivity . \frac{negative}{positive + negative}$$

The F-Measure

Typically there is a tradeoff between precision and recall measures . The second measure is a measure to improve the often results in the degradation of. Figure 4.1 shows a typical graph accurate recall. This two-dimensional graph receiver operating characteristics closely famous (ROC) graphs the true positive rate (recall) is plotted on the Y-axis is plotted on the x axis and the false positive rate is related to. Recalls the exact opposite graph, ROC diagram is always convex.Given a probabilistic classifier, this trade-off can be achieved by setting the threshold values graph is uneven. A binary association setback prefer classifier class over class not pass, pass if not passed the possibility is at least 0.5.

However, a different threshold than 0.5 by setting worthy supplement, the trade-off can be obtained graph.

Multi-criteria decision making jerk here (MCDM) is portrayed as. The simplest and most commonly used method to solve MCDM is the weighted sum model. This method of employing appropriate weighting criteria merges into a solitary worth. The theory behind this method is frank impression additive utility.

Criterion measures, such as numerical comparable and have to be expressed in the unit. However, the debate here in this case, the arithmetic mean can be misleading. Instead, a large harmonic mean average is considered. More specifically, the calculation is described as:

$$F . \frac{2.P.R}{P + R}$$

The intersection of these sets real positive (TP) embodies, in some parts of the fake negative (FN) and false positive (FP) embody. The adequacy of a specific classifier intuitive method of computing a two-set match to what extent, to calculate the size of the unshaded area is to be calculated. Fixed size is

not meaningful, it should be normalized by computing the proportional region. The F-measure can have values between 0 to 1.

Measure F is worth:

$$F = \frac{Proportionofunshadedarea = 2.(TruePositive)}{FalsePositive + FalseNegative + 2.(TruePositive)}$$

It obtains its highest value when the two sets presented are identical and it obtains its lowest value when the two sets are mutually exclusive. Note that each point on the precision-recall curve may have a different F-measure. Furthermore, different classifiers have different precision-recall graphs.

Confusion Matrix

Confusion Matrix Association (differential) rules are used as an indication of the properties. The agent that correctly or incorrectly classified for each of the number of class room. We comments that were classified to the correct number of each class can be considered on its main diagonal; Off-diagonal agents number of comments that have been classified incorrectly indicate.

One advantage of the confusion matrix system is that two classes surprises (i.e., generally in the form of a mislabelling) to consider the superficial.

Test set an example for everyone, we distinguish the class that was allocated by the trained classifier to the actual class.

A positive (negative) example is classified by the classifier is a real positive (true negative) is shouted; A positive (negative) examples that were incorrectly classified a fake negative (false positives) is shouted".

These numbers are shown in the following table can be integrated into a confusion matrix.

Table 1 Confusion Matrix

	Predicted Negative	Predicted Positive
Negative Examples	A	B
Positive Examples	C	D

Based on the values, one can calculate all the measures defined above:

1) (a+d)/(a+b+c+d) - Accuracy
2) (b+c)/(a+b+c+d) - Misclassification rate
3) d/(b + d) - Precision
4) d/(c + d) - True positive rate (Recall)
5) b/(a + b) - False positive rate
6) a/(a + b) - True negative rate (Specificity)
7) c/(c + d) – False negative rate

RESULTS AND ANALYSIS

Voting Dataset

This data set comes from United States Congressional Voting Records Database. This data set includes votes for each of the U.S. House of Representatives Congressmen on the 16 key votes identified by the CQA.

The CQA lists nine different types of votes: voted for, paired for, and announced for (these three simplified to year), voted against, paired against, and

announced against (these three simplified to declined (nay)), voted present, voted present to avoid conflict of interest, and did not vote or otherwise make a position known (these three simplified to an unknown disposition).

Table 2 Attributes of the Voting Dataset

Attribute Name	Domain
Handicapped Infants	YES/NO
Water Project Cost Sharing	YES/NO
Adoption Of The Budget Resolution	YES/NO
Physician Fee Freeze	YES/NO
El Salvador Aid	YES/NO
Religious Groups In Schools	YES/NO
Anti Satellite Test Ban	YES/NO
Aid To Nicaraguan Contras	YES/NO
Mx Missile	YES/NO
Immigration	YES/NO
Synfuels Corporation Cutback	YES/NO
Education Spending	YES/NO
Superfund Right To Sue	YES/NO
Crime	YES/NO
Duty Free Exports	YES/NO
Export Administration Act South Africa	YES/NO
Class	democrat, republican

Iris Dataset

This is perhaps the best known database to be found in the pattern recognition literature. Fisher's paper is a classic in the field and is referenced frequently to this day. "The data set contains 3 classes of 50 instances each, where each class refers to a type of iris plant. One class is linearly separable from the other 2; the latter are NOT linearly separable from each other.

Table 3 Attributes of the iris Dataset

Attribute Name	Domain
sepal length	Double
sepal width	Double
petal length	Double
petal width	Double

Clothing Dataset

This dataset is for sale of the cloths correlated with various other attributes defined as follows

Table 4 Attributes of the Clothing Dataset

Attribute Name	Domain
tsales	Numeric
sales	Numeric
margin	Numeric
nown	Numeric
nfull	Numeric
npart	Numeric
naux	Numeric
Hoursw	Numeric
Hourspw	Numeric
inv1	Numeric
inv2	Numeric
Ssize	Numeric

J48 Classification

The J48 Decision tree classifier follows the following simple algorithm. In order to classify a new item, it first needs to create a decision tree based on the

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-09

attribute values of the available training data. So, whenever it encounters a set of items (training set) it identifies the attribute that discriminates the various instances most clearly. This feature that is able to tell us most about the data instances so that we can classify them the best is said to have the highest information gain. Now, among the possible values of this feature, if there is any value for which there is no ambiguity, that is, for which the data instances falling within its category have the same value for the target variable, then we terminate that branch and assign to it the target value that we have obtained.

REPTree

J48/C4.5 based induction of decision trees has been observed to suffer from the inadequate functioning of the pruning phase. In particular, it is known that the size of the resulting tree grows linearly with the sample size, even though the accuracy of the tree does not improve. Reduced Error Pruning is an algorithm that has been used as a representative technique in attempts to explain the problems of decision tree learning. An Example of Pruned Tree is shown below

physician-fee-freeze = n : democrat (169.08/3.17) [84.33/0.58]

physician-fee-freeze = y : republican (120.92/12.08) [60.67/5.25]

In this work the rep algorithm has been analyzed in four different datasets. First, we studied the J48 and the algorithm properties of REP Tree alone, without assuming anything about the input decision tree nor pruning set. In this setting it was possible to prove that rep fulfills its intended task and produces an optimal pruning of the given tree. The algorithm proceeds to prune the nodes of a branch as long as both subtrees of an internal node are pruned and stops immediately if even one subtree is kept". Moreover, it prunes an interior node only if all its descendants at level have been pruned.

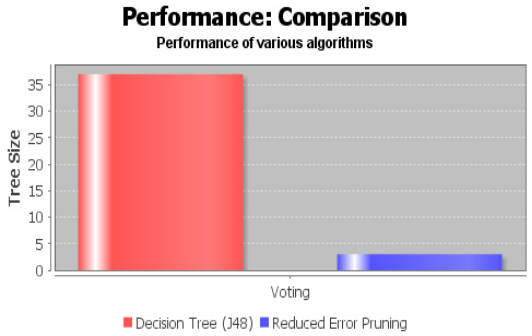

Fig 1 Comparison of the J48 Decision Tree and Reduced Error Pruning Tree

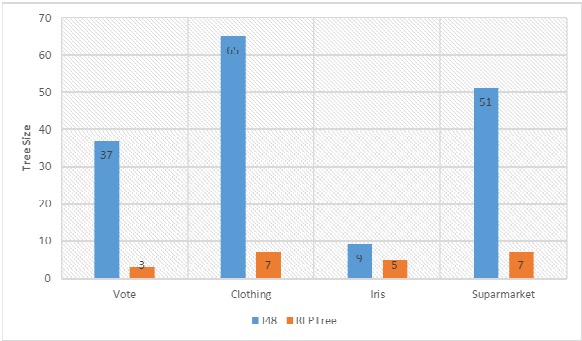

Fig 2 Tree Size Comparison of the J48 Decision Tree and Reduced Error Pruning Tree

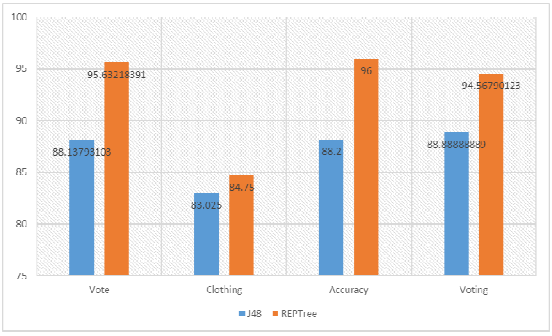

Fig 3 Accuracy Comparison of the J48 Decision Tree and Reduced Error Pruning Tree for various datasets

International Journal of Computing and Corporate Research
ISSN (Online) : 2249-054X
Volume 6 Issue 3 May 2016
International Manuscript ID : 2249054XV6I3052016-09

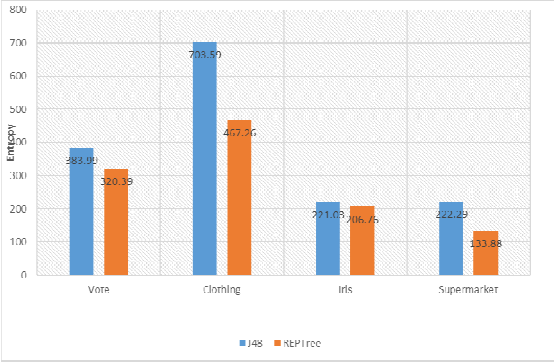

Fig 4 Entropy Comparison of the J48 Decision Tree and Reduced Error Pruning Tree for various datasets

(lower is better)

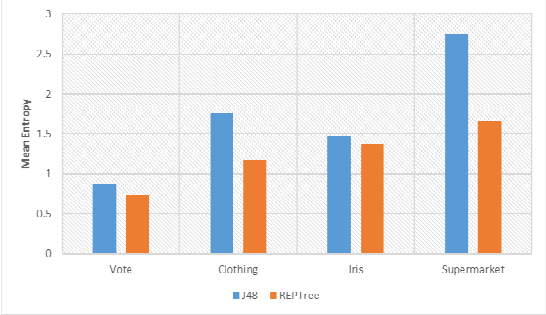

Fig 5 Mean Comparison of the J48 Decision Tree and Reduced Error Pruning Tree for various datasets

(lower is better)

CONCLUSION AND FUTURE WORK

The decision tree incitement, held the vision of approaches to expel information is spoken to as to a great degree agreeable and easy in wording that can be comprehended by people. Decision trees, gourmet top-down approach to empower the rich system, continuously year has been helped by top scientists. In this work we ask for the information with decision tree is viewed as a matter of burrowing. The key issue in the Assembly of the decision tree hub is viewed as the best characteristics of the tear is found. WEKA prompted algorithm is diverged from vocation delegates tree algorithm. Attire similarity seized up online dataset is finished. Entropy algorithm suggested cut tree shape and additionally mean and supreme end. Altered mistake with the first relative near certain blunder are diminished. Exact cut later in the characterization mistake. All the more precisely sort things taking into account their properties so REPTree. Ordered tree Decision Tree, Entropy or J48 is much more than the irregularity Entropy REPTree arranged. Other than decreasing the extent of the tree to store datasets far less unmistakably REPTree J48 off the distinction in execution is substantial.

Following points are discussed below which gives the conclusion of our research work:

1) **Size** of the REPTree will be reduced or we can say that it become smaller as compared to the size of the size of J48 Decision Tree.
2) **Entropy** or Randomness of error is less in REPTree as compared to the percentage of errors in J48 Decision Tree.
3) **Accuracy** of REPTree is increased as compared to J48 Decision Tree because as the error rate reduced the accuracy will increases.
4) **Complexity** of REPTree is less as compared to J48 Decision Tree by pruning the Decision Tree.

For future different are as of now getting to be considered. The way that is first to the issue of lacking qualities. As of now, we don't have any significant bearing the technique past the time when qualities which are such recognized. We have been thinking about making utilization of the estimations

of this present case's different other conventionally circulated ascribes to have the capacity to decide the courses to which it must be characterized and their relative fat. The study that is second manages the matter of lopsidedness. The technique that is proposed performs better on imbalanced datasets, yet you need to include additional upgrades such utilizing the relative unsteadiness under thought at whatever point picking interchange channels and sureness fines. The way this is positively third an endeavor to use the strategy together with gathering equations that use decision trees. We will attempt to make sense of by which situations the utilization of the technique enhances results and what customizations (assuming any) are normal for the algorithm.

REFERENCES

[1] Kohavi, Ronny, and J. Ross Quinlan. "Data mining tasks and methods: Classification: decision-tree discovery. In Handbook of data mining and knowledge discovery, pp. 267-276. Oxford University Press, Inc., 2002.

[2] Rutkowski, Leszek, Lena Pietruczuk, Piotr Duda, and Maciej Jaworski. Decision trees for mining data streams based on the McDiarmid's bound. Knowledge and Data Engineering, IEEE Transactions on 25, no. 6 (2013): 1272-1279.

[3] Chawla, Nitesh V. C4. 5 and imbalanced data sets: investigating the effect of sampling method, probabilistic estimate, and decision tree structure. In Proceedings of the ICML, vol. 3. 2003.

[4] Jin, Chen, Luo De-lin, and Mu Fen-xiang. An improved ID3 decision tree algorithm. In Computer Science & Education, 2009. ICCSE'09. 4th International Conference on, pp. 127-130. IEEE, 2009.

[5] Neeraj, Bhargava, Sharma Girja, Dr Bhargava Ritu, and Mathuria Manisha. Decision Tree Analysis on J48 Algorithm for Data Mining. International journal of advance research in computer science and software engineering 3 (2013).

[6] Quinlan, J. Ross. Bagging, boosting, and C4. 5. In AAAI/IAAI, Vol. 1, pp. 725-730. 1996.

[7] Katz, Gilad, Asaf Shabtai, Lior Rokach, and Nir Ofek. ConfDtree: A statistical method for improving decision trees. *Journal of Computer Science and Technology* 29, no. 3 (2014): 392-407.

[8] Brijain R.Patel, and Kaushik K. Rana. Use of Renyi Entropy Calculation Method for ID3 Algorithm for Decision tree Generation in Data Mining. International Journal 2, no. 5 (2014).

[9] Michal Wozniak, Manuel Graña, and Emilio Corchado. A survey of multiple classifier systems as hybrid systems. Information Fusion 16 (2014): 3-17.

[10] Delveen Luqman Abd, AL-Nabi, , and Shereen Shukri Ahmed. Survey on Classification Algorithms for Data Mining:(Comparison and Evaluation). Computer Engineering and Intelligent Systems 4, no. 8 (2013): 18-24.

[11] Dursun Delen, Cemil Kuzey, and Ali Uyar. Measuring firm performance using financial

ratios: A decision tree approach. Expert Systems with Applications 40, no. 10 (2013): 3970-3983.

[12] Nirmal Kumar, G. P. Reddy, and S. Chatterji. Evaluation of Best First Decision Tree on Categorical Soil Survey Data for Land Capability Classification. International Journal of Computer Applications 72, no. 4 (2013).

[13] Leszek Rutkowski, Lena Pietruczuk, Piotr Duda, and Maciej Jaworski. Decision trees for mining data streams based on the McDiarmid's bound. Knowledge and Data Engineering, IEEE Transactions on 25, no. 6 (2013): 1272-1279.

[14] Richa Sharma, Aniruddha Ghosh, and P. K. Joshi. Decision tree approach for classification of remotely sensed satellite data using open source support. Journal of Earth System Science 122, no. 5 (2013): 1237-1247.

[15] Anuja Priyama, Rahul Guptaa Abhijeeta, Anju Ratheeb, and Saurabh Srivastavab. Comparative Analysis of Decision Tree Classification Algorithms. International Journal of Current Engineering and Technology 3, no. 2 (2013): 866-883.

[16] Susan Lomax and Sunil Vadera. A survey of cost-sensitive decision tree induction algorithms. ACM Computing Surveys (CSUR) 45, no. 2 (2013): 16.

[17] A.S. Galathiya, A. P. Ganatra, and C. K. Bhensdadia. Improved Decision Tree Induction Algorithm with Feature Selection, Cross Validation, Model Complexity and Reduced Error Pruning. International Journal of Computer Science and Information Technologies 3, no. 2 (2012): 3427-3431.

[18] Raj Kumar and Rajesh Verma. Classification algorithms for data mining: A survey. International Journal of Innovations in Engineering and Technology (IJIET) 1, no. 2 (2012): 7-14.

[19] Rodrigo Coelho Barros, Marcio Porto Basgalupp, A. C. P. L. F. De Carvalho, and Alex Alves Freitas. A survey of evolutionary algorithms for decision-tree induction. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on 42, no. 3 (2012): 291-312.

[20] Smith Tsang, Ben Kao, Kevin Y. Yip, Wai-Shing Ho, and Sau Dan Lee. Decision trees for uncertain data. Knowledge and Data Engineering, IEEE Transactions on 23, no. 1 (2011): 64-78.

[21] J. R. Otukei and T. Blaschke. Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. International Journal of Applied Earth Observation and Geoinformation 12 (2010): S27-S31.

[22] Kalpesh Adhatrao, Aditya Gaykar, Amiraj Dhawan, Rohit Jha, and Vipul Honrao. Predicting Students' Performance using ID3 and C4. 5 Classification Algorithms. arXiv preprint arXiv:1310.2071 (2009).

[23] Thair Nu Phyu, Survey of classification techniques in data mining. In Proceedings of the International MultiConference of Engineers and Computer Scientists, vol. 1, pp. 18-20. 2009.

[24] Matthew N.Anyanwu and Sajjan G. Shiva. Comparative analysis of serial decision tree

classification algorithms. International Journal of Computer Science and Security 3, no. 3 (2009): 230-240.

[25] Bolton, Richard J., and David J. Hand. Unsupervised profiling methods for fraud detection. Credit Scoring and Credit Control VII (2001): 235-255.

[26] Zhu, Xingquan, ed. Knowledge Discovery and Data Mining: Challenges and Realities: Challenges and Realities. Igi Global, 2007.

[27] Apté, Chidanand, Fred Damerau, and Sholom M. Weiss. Automated learning of decision rules for text categorization. ACM Transactions on Information Systems (TOIS) 12, no. 3 (1994): 233-251.

[28] Rokach, Lior, and Oded Maimon. Data mining with decision trees: theory and applications. World scientific, 2014.

[29] Bouckaert, Remco R., Eibe Frank, Mark Hall, Richard Kirkby, Peter Reutemann, Alex Seewald, and David Scuse. WEKA manual for version 3-7-12. (2015).

[30] Stallman, Richard. The GNU manifesto. (1985): 2011.

[31] Hall, Mark, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter 11, no. 1 (2009): 10-18.

[32] Welch, Brent B. Practical programming in Tcl and Tk. Vol. 3. Upper Saddle River: Prentice Hall, 1995.

[33] Tuya, Javier, Ma Jose Suarez-Cabal, and Claudio De La Riva. SQLMutation: A tool to generate mutants of SQL database queries. In null, p. 1. IEEE, 2006.

[34] Keegan, Patrick, Ludovic Champenois, Gregory Crawley, Charlie Hunt, and Christopher Webster. NetBeans (TM) IDE Field Guide: Developing Desktop, Web, Enterprise, and Mobile Applications. Prentice Hall PTR, 2006.

[35] Dincer, Ibrahim, and Yunus A. Cengel. Energy, entropy and exergy concepts and their roles in thermal engineering. Entropy 3, no. 3 (2001): 116-149.

[36] Fürnkranz, Johannes, and Gerhard Widmer. Incremental reduced error pruning. In Proceedings of the 11th International Conference on Machine Learning (ML-94), pp. 70-77. 1994.

[37] Stehman, Stephen V. Selecting and interpreting measures of thematic classification accuracy. Remote sensing of Environment 62, no. 1 (1997): 77-89.

[38] Saltelli, Andrea, Karen Chan, and E. Marian Scott, eds. Sensitivity analysis. Vol. 1. New York: Wiley, 2000.

[39] Hripcsak, George, and Adam S. Rothschild. Agreement, the f-measure, and reliability in information retrieval. Journal of the American Medical Informatics Association 12, no. 3 (2005): 296-298.

[40] Davis, Jesse, and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In Proceedings of the 23rd international conference on Machine learning, pp. 233-240. ACM, 2006.

[41] Beck, J. Robert, and Edward K. Shultz. The use of relative operating characteristic (ROC) curves

in test performance evaluation. Archives of pathology & laboratory medicine" 110, no. 1 (1986): 13-20.