C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning

Rutvija Pandya
Diploma Computer Engineering Department, Gujarat Technological University
Atmiya Institute of Tech & Sci
Rajkot

Jayati Pandya
Bachelar in Computer science and Application,Saurashtra University
K.P.Dholakiya Infotech
Amreli

ABSTRACT
Data mining is a knowledge discovery process that analyzes data and generate useful pattern from it. Classification is the technique that uses pre-classified examples to classify the required results. Decision tree is used to model classification process. Using feature values of instances, Decision trees classify those instances. Each node in a decision tree represents a feature in an instance to be classified.
In this research work ID3, C4.5 and C5.0 Compare with each other. Among all these classifiers C5.0 gives more accurate and efficient result. This research work used C5.0 as the base classifier so proposed system will classify the result set with high accuracy and low memory usage. The classification process generates fewer rules compare to other techniques so the proposed system has low memory usage. Error rate is low so accuracy in result set is high and pruned tree is generated so the system generates fast results as compare with other technique. In this research work proposed system use C5.0 classifier that Performs feature selection and reduced error pruning techniques which are described in this paper.
Feature selection technique assumes that the data contains many redundant features. so remove that features which provides no useful information in any context. Select relevant features which are useful in model construction. Cross-validation method gives more reliable estimate of predictive. Over fitting problem of the decision tree is solved by using reduced error pruning technique. With the proposed system achieve 1 to 3% of accuracy, reduced error rate and decision tree is construed within less time.

Keywords
REP, Decision Tree induction, C5 classifier, KNN, SVM

1. INTRODUCTION
This paper describes first about different classification methods. Compare them based on their features.

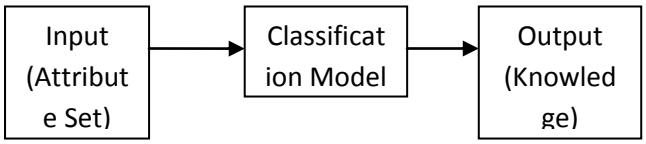

Fig 1: data mining using classification model

The research work is made up from ID3, C4.5 and C5 classifier. In many applications, rule sets are preferred because they are simpler and easier to understand than decision trees.[1] Both C4.5 and C5.0 can produce classifiers demonstrated either as decision trees or rule sets, but C4.5'sruleset methods are slow and it required high memory also. The new algorithm C5.0 represents that how the rule sets are generated with improved features. This research work focuses on high accuracy and low memory usage. C5.0 generates fewer rules so memory usage is low compare to other classifier.

2. SURVEY ON CLASSIFICATION ALGORITHMS

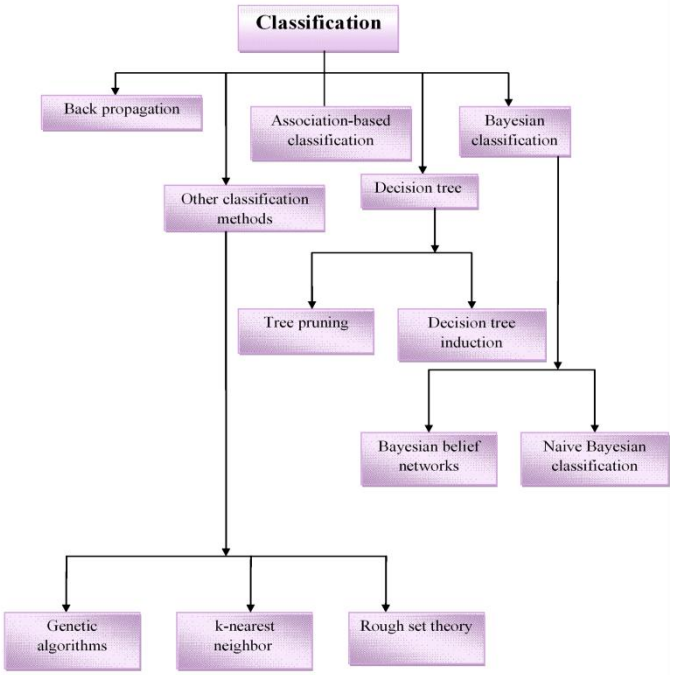

Fig 2: Classification algorithm

2.1 Decision Trees
This method used tree structure to build the classification models. It divides a dataset into smaller subsets. Leaf node represents a decision. Based on feature values of instances, the decision trees classify the instances. Each node represents a feature in an instance in a decision tree which is to be classified, and each branch represents a value. Classification of Instances starts from the root node and sorted based on their feature values. Categorical and numerical data can be handled by decision tress. [2]

2.2 Bayesian classification
Bayesian classification can predict class membership probabilities. the effect of an attribute value on a given class is independent of the value of the other attributes is assumed by the Naïve Bayes algorithm. The Naïve Bayes algorithm scales continuously in the number of predictors and rows and builds rapidly models. Naive Bayes algorithm derives the probability

of a prediction. The probability of event X occurring given that event Y has occurred (P(X|Y)) is proportional to the probability of event Y occurring given that event X has occurred multiplied by the probability of event X occurring ((P(Y|X)P(X)).

2.3 Rule-Based Classification

A rule-based classification technique used a collection of if ... then ... rules for classifying records. For rule-based classifier, extract a set of rules that show the relationships between the attributes of a dataset and the class label. Here the if part of a rule is known as the precondition of rule. The then part of the rule is consequent of rule. The Coverage of a rule is the number of instances that satisfy the previous rule. The Accuracy of a rule is the fragment of instances that satisfy both the previous and consistent rule, normalized by those satisfying the antecedent. Ideal rules should have both high coverage and high accuracy rates. [3]

2.4 K-Nearest Neighbour

*K*NN instance-based classifier operates on unknown instances. According to some function, Classification can be done by relating the unknown to the known. In k-nearest neighbor classification unknown sample is given and the classifier searches the pattern space which is used for the k training samples. These samples are closest to the unknown sample. "Closeness" is defined in terms of Euclidean distance. The unknown sample is assigned to the most common class among its k nearest neighbours.

2.5 SVM

Support Vector Machine (SVM) classification technique analyze data and recognize patterns from them. From statistical learning theory, SVM a machine learning algorithm is derived. SVM classification use a very small sample set and generate pattern from that.

3. COMPARISION OF CLASSIFICATION ALGORITHMS

Table 1 Feature comparison

Feature	Decision Tree	Naive Bayes	K-Nearest Neighbour	Rule-Based	Support Vector Machine
Learning Type	Eager Learner	Eager Learner	Lazy learner	Lazy learner	Eager Learner

4. C5 CLASSIFIER

The classifier is tested first to classify unseen data and for this purpose resulting decision tree is used. C4.5 algorithm follows the rules of ID3 algorithm. Similarly C5 algorithm follows the rules of algorithm of C4.5. C5 algorithm has many features like:

- The large decision tree can be viewing as a set of rules which is easy to understand.
- C5 algorithm gives the acknowledge on noise and missing data.
- Problem of over fitting and error pruning is solved by the C5 algorithm.
- In classification technique the C5 classifier can anticipate which attributes are relevant and which are not relevant in classification.

4.1 Comparison- Current Algorithms:

4.1.1 Improvement in C4.5 from ID3 Algorithm:

C4.5 algorithm handles both continuous and discrete attributes. For handling continuous attributes, C4.5 creates a threshold and then makes the list of attributes having value above the threshold and less than or equal to the threshold. C4.5 algorithm also handles the training data with missing attribute values. It allows attribute values to be marked as '?'. In gain and entropy calculations the missing attribute values are not used. C4.5 pruning trees after its creation. Once the tree is created this algorithm goes back to it and replace the branches that do not help in classification.

4.1.2 Improvement in C5 from C4.5 Algorithm:

C5 is faster than C4.5**.** Memory usage is more efficient in C5 than C4.5. C5 gets smaller decision trees in comparison with C4.5. The C5 rule sets have lower error rates on unseen cases. So comparing with C4.5 the accuracy of result is good with C5 algorithm.C5 automatically allows removing unhelpful attributes.

4.2 Problem of Current System:

There are some difficulties in learning decision trees. It is difficult to take decision that how deeply to grow the decision tree. It is also difficult to choose an appropriate attribute selection measure and manage training data with missing attribute values.

4.3 Solution via Proposed System:

For the solution of above discussed problems are generally arise while using too small training set. It is very difficult to classify this kind of training set. The result produced by simple algorithm may *over-fit* the training examples. To avoid over-fitting in decision tree learning two different approaches is used. The first approach is that, do not produce the tree before it reaches the point where the training data is perfectly classifies. The second approach is, post prune the tree when the tree over-fit the data.

Accuracy	Good in many domains	Good in many domains	High – Robust	Significantly high	Significantly high
Scalability	Efficient for small data set	Efficient for large data set	-	Efficient for Uncertain data set	-
Interpretability	Good	-	-	-	-
Speed	Fast	Very fast	Slow	-	Fast with active learning
Transparency	Rules	No rules (black box)	Rules	Rules	No rules (black box)

4.4 Proposed Algorithm:

4.3.1 Feature Selection:

Data with extremely high dimensionality has presented serious challenges to existing learning methods [6]. Because of large number of features, a learning model generate the result which is over-fited, so the performance is degraded. For reducing dimensionality, the feature selection is a widely used technique. It chooses a small subset of the relevant features from the actual set which usually provides better learning performance and model interpretability. Using Feature selection technique the computational cost of learning can also reduce.

4.3.2 Reduced Error Pruning:

In classification models, it is common practice to discard parts of the model that describe spurious effects in the training sample rather than true features. REP is a technique that removes sections of the tree to reduce the size of decision trees. It removes the part of the tree which provides less power for the classification of instances. The Pruning technique reduced complexity and provides better accuracy by reducing over fitting problem.

4.3.3 Algorithm:

- To make the tree Create a root node
- Check the base case
- **With the use of Genetic Search Apply Feature Selection technique**

  bestTree = Construct a decision tree using training data
- **Apply Cross validation technique**
  1. Divide all training data into N disjoint subsets, $R = R_1, R_2, ..., R_N$
  2. For each j = 1, ..., N do
     - ✓ Test set = R_j
     - ✓ Training set = $R – R_j$
     - ✓ Using Training set, Compute the decision tree
     - ✓ Decide the performance accuracy X_j with the use of Test set
  3. Reckon the N-fold cross-validation technique to estimate the performance

     $= (X_1 + X_2 + ... + X_N)/N$

- **Apply Reduced Error Pruning technique**

  Find the attribute with the highest info gain (A_Best)

Classification:

For each tj ϵ D, apply the DT to determine its class

5. RESEARCH WORK

5.1 Input Parameter:

Example

Input data which is requisite to classify correctly.

Attributes

Input to the algorithm consists of a collection of training cases, with a class attribute and a tuple of values for a fixed set of attributes R = {R1, R2,...., R_n} and a class attribute.

5.2 Output:

Generate the Decision tree which classifies the training data correctly.

6. EXPERIMENTAL RESULTS

With the use of Global Pruning technique and Cross validation technique with algorithm C5, the results are generated as below:

Table 2 Experimental results

Dataset	w/o Global Pruning and Cross Validation		with Global Pruning and Cross Validation	
	Training Data Error	Testing Data Error	Mean Error	Std. Error
Soybean.data	5%	15.5%	11%	1.7%
Banding.data	0.8%	6%	3%	0.9%

7. CONCLUSION

C5 is a classifier which classifies the data in less time compare to other classifier. For generating decision tree the memory usage is minimum and it also improve the accuracy. This proposed system is developed on the bases of C5 algorithm. In the proposed system C5.0 algorithm provides Feature selection, Cross validation and reduced error pruning facilities.

So the further scope of this algorithm is achieved by implementation of new features like Cross Validation and Model Complexity. By implementing all the characteristic of algorithm using weka packages, the accuracy in classification is improved.

8. ACKNOWLEDGMENTS

This research paper is made possible through the help and support from everyone. Especially, please allow me to dedicate my gratitude toward the following significant advisors and contributors: First and foremost, I would like to thank GOD for his unconditional guidance and wisdom as I make my research. Second, I would like to thank my friend for her most support and encouragement for giving us this research. This gives us the experience on how to cooperate and engage ourselves in a serious project. Finally, I sincerely thank to my parents, family, and friends, who provide the advice and support.

9. REFERENCES

[1] XindongWu · Vipin Kumar · J. Ross Quinlan · Joydeep Ghosh · Qiang Yang ·Hiroshi Motoda · Geoffrey J.

[2] McLachlan · Angus Ng · Bing Liu · Philip S. Yu ·Zhi-Hua Zhou · Michael Steinbach · David J. Hand · Dan Steinberg Top 10 algorithms in data mining © Springer-Verlag London Limited 2007

[3] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", International MultiConference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong

[4] Biao Qin, Yuni Xia, Sunil Prabhakar,Yicheng Tu" A Rule-Based Classification Algorithm for Uncertain Data", Department of Computer ScienceIndiana University - Purdue University Indianapolis, USA

[5] A comparative study of decision tree ID3 and C4.5 Badr HSSINA, Abdelkarim MERBOUHA,Hanane EZZIKOURI,Mohammed ERRITALI TIAD laboratory, Computer Sciences Department, Faculty of sciences and techniques Sultan Moulay Slimane University Beni-Mellal, BP: 523, MoroccoM. Govindarajan, Text Mining

Technique for Data Mining Application, World Academy of Science, Engineering and Technology 35 2007

[6] Sohag Sundar Nanda, Soumya Mishra, Sanghamitra Mohanty, Oriya Language Text Mining Using C5.0 Algorithm, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 2 (1) , 2011

[7] cTomM.Mitchel, McGrawHil, Decision Tree Learning, Lecture slides for textbook Machine Learning, , 197

[8] Zuleyka Díaz Martínez, José Fernández Menéndez, Mª Jesús Segovia Vargas See5 Algorithm versus Discriminant Analysis, Spain.

[9] J.R, QUINLAN , Induction of Decision Trees, New South Wales Institute of Technology, Sydney 2007, Australia

[10] A. S. Galathiya, A. P. Ganatraand 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.