An Efficient Algorithm for Generating Classification Rules

[1]S.Vijayarani, [2]M.Divya
[1,2] Dept. of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore, Tamil Nadu, India

Abstract
Data mining is the process of extracting hidden knowledge from the large data repositories. In data mining, there are several techniques and algorithms are used for extracting the hidden information and finding the relationships between them. Classification is one of the popular techniques of data mining. Classification is a data mining technique which is used to predict group membership for the instances of data. Classification is the task of generalizes the known structure to apply to new data. Classification involves finding rules that partition the data into disjoint groups. Many classification rule algorithms are used to generate the classification rules such as ID3, CART, and uRule. In this research work, we have analyzed the performance of the three classification rule algorithms, namely C4.5, RIPPER and PART algorithms.

Keywords
Data Mining, Classification Rules, C4.5, Ripper, Part.

I. Introduction

Data mining is the process of extracting hidden patterns from large data sets. This process is achieved by combining methods from statistics and artificial intelligence with database management. Data mining is used in the analysis of collections of observations of behavior. Data mining is a searching process done automatically for hidden patterns present in a large database [2].

Data mining is an iterative process. Its progress is defined by discovery, through either automatic or manual methods. Data mining is useful in an exploratory analysis scenario in which there are no predetermined notions which constitutes an "interesting" outcome. Data mining is one of the fastest growing fields in the computer industry. Data mining is reflected in its wide range of methodologies and techniques [8].

These techniques can be applied to a connection of problem sets. Classification deals in generating rules that partition the data into disjoint groups. Classification is a data mining (machine learning) technique used to predict group membership for data instances [4].

The goal of the classification is to assign a class to find previously unseen records as accurately as possible. Classification process consists of training set that are analyzed by a classification algorithms and the classifier or learner model is represented in the form of classification rules [9].

Test data are used in the classification rules to estimate the accuracy. Classification is the identification of new patterns, such as coincidence between duct tape purchases and plastic sheeting purchases [1].

There are various kinds of classification method including decision tree induction, Bayesian networks, k-nearest neighbor classifier, case-based reasoning, genetic algorithm and fuzzy logic techniques. Systems that construct classifiers are one of the commonly used tools in data mining. Such systems take as input a collection of cases, each belonging to one of a small number of classes and described by its values for a fixed set of attributes, and output a classifier that can accurately predict the class to which a new case belongs [7]. The rest of this paper is organized as following. Section II provides a review of literature. The problem definition is given in Section III. Subsequently, our proposed approach is discussed in Section IV. The experimental results are given in Section V. Finally, Section VI, gives the conclusion.

II. Related works

A classification rule is a procedure in which the elements of the population set are each assigned to one of the classes. A classification rule or classifier is a function that can be evaluated for any possible value specifically given the data it will yield a similar classification. In a binary classification, the elements that are not correctly classified are named false positives and false negatives [12].

Some classification rules are static functions. Others can be computer programs. A computer classifier can be able to learn or can implement static classification rules. Given a classification rule, a classification test is the result of applying the rule to a finite sample of the initial data set [7].

There are various classification rule algorithms namely OneR, Ridor, Conjuctive Rule etc. There are two types in extracting classification rules namely direct method and indirect method. In direct method the rules are extracted from data [5].

In indirect method the rules are extracted from other classification models. The classification rules are also known as if then rules. This paper describes about the various rule based classification algorithms namely C4.5, Ripper and Part algorithms. In this paper we review about the role of those three algorithms in various concepts. In the paper "C4.5 algorithm and Multivariate Decision Trees [10], a brief description about the c4.5 algorithm is given which is used to create Univariate Decision Trees. It mainly concentrates on how the algorithm is implemented. The implementation is done through WEKA. C4.5 builds Univariate and Multivariate Decision Tree. In the paper "Towards the use of C4.5 algorithm for classifying Banking Dataset" [11] has given a detailed view of C4.5 algorithm. In this C4.5 algorithm, the process and the result which utilizes C4.5 for classifying banking dataset is given. To utilize C4.5 algorithm, rules have been generated from dataset. Banking dataset is taken for experiment. C4.5 algorithm performs well in the construction of decision trees and extracts rules from the banking dataset. In the paper "Rule-based Text Categorization Using Hierarchical Categories" [6] provides an efficient method for hierarchical document categorization based on the rule learning algorithm RIPPER (for Repeated Incremental Pruning to Produce Error Reduction). The problem of automatic document categorization by using the rule learning algorithm RIPPER is pointed out. To decrease the failure of retrieval, the RIPPER algorithm is extended to hierarchical RIPPER algorithm. In the paper, "FURIA: An algorithm for unordered fuzzy Rule Induction" [3] gives an algorithm namely FURIA which is an extension of RIPPER algorithm is used. This FURIA algorithm is a novel fuzzy rule-based classification. The experiment results show that FURIA outperforms the original RIPPER.

III. Problem definition

Given a dataset D, a set of classes C, a set of classification rules R over D through the algorithms C4.5, RIPPER and PART, find the best algorithm using some the performance factors.

IV. Proposed system

In the proposed system a clear view of the three algorithms is given. This section discusses a brief description of the three classification rule algorithms.

A. C4.5 algorithm

C4.5 is a program that inputs a set of labeled data and generates a decision tree as output. This resultant decision tree is then tested against unseen labeled test data to quantify its generalization. C4.5 is a program used for creating classification rules using decision trees from a set of given data.

C4.5 algorithm is an extension of the basic ID3 algorithm and it was designed by Quinlan. C4.5 is one of widely-used learning algorithms. C4.5 algorithm builds decision trees from a set of training data similar to the ID3algorithm, using the concept of information entropy. C4.5 is also known as a statistical classifier.

1. Check for base cases.
2. For each attribute a.

    1.1. Find the normalized information gain from splitting on a

3. Let a_best be the attribute with the highest normalized information gain.
4. Create a decision node that splits on a best.
5. Recurse on the sub lists obtained by splitting on a_best, and add those nodes as children of node.

Fig. 1: Algorithm for C4.5

At each node of the tree, C4.5 chooses one attribute of the data that splits its set of samples into subsets enriched in one class or the other.

B. Ripper algorithm

RIPPER is abbreviated as Repeated Incremental Pruning to Produce Error Reduction. RIPPER is especially more efficient on large noisy datasets .There are two kinds of loop in Ripper algorithm. This algorithm was designed by Cohen in 1995 namely, Outer loop and Inner loop. Outer loop adds one rule at a time to the rule base and Inner loop adds one condition at a time to the current rule. The information gain measure is maximized by adding the conditions to the rule. This process is continued until it covers no negative example. The algorithm is shown in the following fig.

1. Ripper(Pos, Neg, k)
2. Rule Set ← LearnRuleSet(Pos, Neg)
3. For k times
4. RuleSet ← OptimizeRuleSet(RuleSet, Pos, Neg)
5. LearnRuleSet(Pos, Neg)
6. RuleSet ← ∅
7. DL ← DescLen(RuleSet, Pos, Neg)
8. Repeat
9. Rule ← LearnRule(Pos, Neg)
10. Add Rule to RuleSet
11. DL` ← DescLen(RuleSet, Pos, Neg)
12. If DL` > DL + 64
13. PruneRuleSet(RuleSet, Pos, Neg)
14. Return RuleSet
15. If DL1 < DL DL ←DL`
16. Delete instances covered from Pos and Neg
17. Until Pos = ∅
18. Return RuleSet

Fig. 2: Algorithm for Ripper

The time complexity of this algorithm is O ($N\log^2 N$). The description length of rule base is termed as DL.

C. Part algorithm

PART stands for Projective Adaptive Resonance Theory. The input for PART algorithm is the vigilance and distance parameters [13].

1. Initialization

Number m of nodes in F_1 layer:=number of dimensions in the input vector. Number m of nodes in F layer:=expected maximum number of clusters that can be formed at each clustering level.

Initialize parameters L, ρ_o, ρ_h, σ, α, θ, and e.

1. Set $\rho = \rho_o$.
2. Repeat steps 3 – 7 until the stopping condition is satisfied.
3. Set all F_2 nodes as being noncommitted.
4. For each input vector in dataset S, do steps 4.1-4.6.
   a. Compute h_{ij} for all F_1 nodes v_i and committed F_2 nodes v_j. If all F_2 nodes are non committed, go to step 4.3.
   b. Compute T_j for all committed F_2 nodes V_j.
   c. Select the winning F_2 node V_J. If no F_2 node can be selected, put the input data into outlier 0 & then continue to do step 4.
   d. If the winner is a committed node, compute r_J, otherwise go to step 4.6.
   e. If $r_J >= \rho$, go to step 4.6, otherwise reset the winner V_J and go back to step 4.3.
   f. Set the winner V_J as the committed and update the bottom-up and top-down weights for winner node V_J.
5. Repeat step 4 N times until stable clusters are formed(i.e. until the difference of output clusters ay N-th and (N-1)-th time becomes sufficiently small)
6. For each cluster C_j in F_2 layer, compute the associated dimension set D_j. Then, set S= C_j and set $\rho = \rho + \rho_h$ (or $\rho = | D |= \rho_h$), go back to step 2.
7. For the outlier O, set S = 0, go back to step 2.

Fig. 3: Algorithm for Part

V. Experimental results

The above three algorithms are compared using dataset namely Heart Disease and Breast Cancer Wisconsin (Original). These dataset are collected from UCI Repository in the website www.ucirepository.com.The breast cancer dataset contains 699 instances and 10 attributes. The heart disease dataset contains 303 instances and 76 attributes. The following charts are based on the comparison made of three algorithms namely C4.5, RIPPER and PART for heart disease dataset. The Matlab tool is used to evaluate the performances of the each algorithm. The Matlab version is 7.8.0(R2000a).MATLAB is a high-level technical computing language and it has a good interactive environment for algorithm development, data visualization, data analysis, and numeric computation.MATLAB provides a high-level language and development tools to develop and analyze the algorithms and applications.

A. Performance measures

In this paper the comparison is based on the number of rules and sensitive rules created by the three algorithms and the time each algorithm takes for creating rules.
The charts are given to denote the accuracy of the algorithms, number of rules created by each algorithm, number of sensitive rules created by each algorithm and time taken by each algorithm to create the rules. The performance measures are, time and No of rules.

B. Time

This is an effective performance measure that estimates the time taken by each algorithm for generating the classification.

Table 1: Time Factor

Dataset	C4.5	Ripper	PART
Breast Cancer	0.1044	0.1108	0.0777
Heart Disease	0.1869	0.1905	0.0840

The chart fig. 4, represents the time complexity of C4.5, Ripper and Part algorithm using Wisconsin Breast Cancer Dataset and Heart Disease Dataset. By analyzing the results, the PART algorithm generates more number of rules compared to the other algorithm.

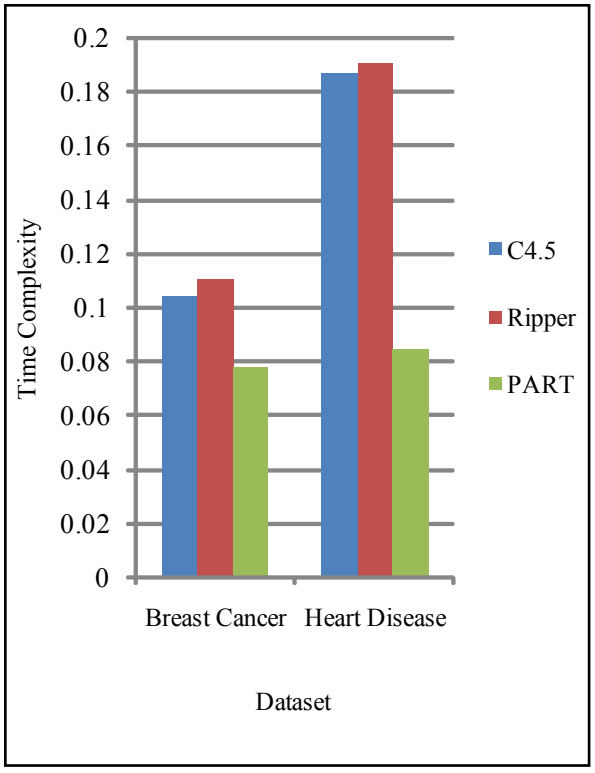

Fig. 4: Time Complexity

C. No of Rule Generation

This factor gives the counts of the total number of rules generated by each algorithm.

Table 2: No of Rule Generation

Dataset	C4.5	Ripper	PART
Breast Cancer	1789	1345	2316
Heart Disease	277	359	765

The chart fig. 5, represents the number of rules generated by C4.5, Ripper and Part algorithms using Wisconsin Breast Cancer Dataset and Heart Disease Dataset. By analyzing the results, the PART algorithm generates more number of rules compared to the other algorithm.

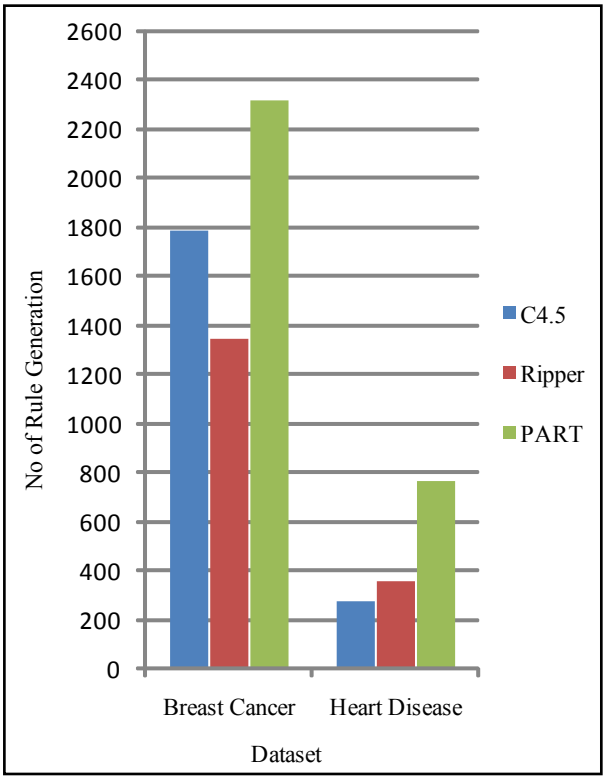

Fig. 5: Chart for No of Rule Generation

VI. Conclusion

In this paper, we have compared C4.5, RIPPER and PART algorithms which are very suitable for generating rules in classification technique. The classification rule generation algorithms generates classification rules which is both sensitive and non sensitive. From the experimental results it is concluded that in the case of time factor & number of rules generation, Part algorithm seems better than the other two algorithms for Breast Cancer Dataset and Heart Disease dataset. In future some privacy preserving technique can be induced for the rule generation in the classification technique.

References

[1] Hongze Qiu, Haitang Zhang, "Fuzzy SLIQ Decision Tree Based on Classification Sensitivity", I.J.Modern Education and Computer Science, 2011, 5, pp. 18-25.

[2] Jeffrey W. Seifert, "Data Mining An Overview", CRS Report for Congress.

[3] "Congressional Research Service", the Library of Congress December 16, 2004.

[4] Jens Hühn, Eyke Hüllermeier, "FURIA: An Algorithm for Unordered Fuzzy Rule Induction", Philipps-Universität Marburg, Department of Mathematics and Computer Science.

[5] Jianyu Yang, Rutgers, "Classification by Association Rules: The Importance of Minimal Rule Sets", the State University of New Jersey, New Brunswick, NJ 08903 USA.

[6] Juggapong Natwichai, Xue Li, Maria Orlowska, "Hiding Classification Rules for Data Sharing with Privacy Preservation", DaWaK 2005, LNCS 3589, Springer-Verlag Berlin Heidelberg 2005, pp. 468–477.

[7] Minoru SASAKI, Kenji KITA, "Rule-based Text Categorization Using Hierarchical Categories", Faculty of Engineering, Tokushima University, Tokushima, Japan, 770-8506.

[8] Murlikrishna Vishwanathan, Geoffrey I.Webb, "Classification Learning Using All Rules", Proceedings of the Tenth European Conference on Machine Learning (ECML '98), Springer, pp. 149-159.

[9] Srivatsan Laxman, P S Sastry, "A survey of temporal data mining", Sadhana Vol. 31, Part 2, 2006, India, pp. 173–198.

[10] Thair Nu Phyu, "Survey of Classification Techniques in Data Mining", Proceedings of the International MultiConference of Engineers and Computer Scientists Vol. I IMECS 2009, 18th–20th March, 2009, Hong Kong.

[11] Thales Sehn Korting, "C4.5 algorithm and Multivariate Decision Trees", Image Processing Division.

[12] Veronica S. Moertini, Jurusan Ilmu Komputer, "Towards the use of C4.5 algorithm for classifiying banking dataset", Fakultas Matematika dan Ilmu Pengetahuan Alam universitas, Katolik Parahyangan Bandung.

[13] Yongqiang Cao, Jianhong Wu, "Projective ART for clustering data sets in high dimensional spaces", Elsevier Science Ltd, Neural Networks 15, 2002, pp. 105-120.

Mrs. S.Vijayarani has completed MCA and M.Phil in Computer Science. She is working as Assistant Professor in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of research interest are data mining, privacy and security issues. She has published papers in the international journals and presented research papers in international and national conferences.

Ms. M. Divya has completed M.Sc in Computer Science. She is currently pursuing her M.Phil in Computer Science in the School of Computer Science and Engineering, Bharathiar University, Coimbatore. Her fields of interest are privacy in data mining.