An Efficient Algorithm for Classification Rule Hiding

S.Vijayarani
Assistant Professor
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Coimbatore, Tamil Nadu, INDIA

M.Divya
Research scholar
Department of Computer Science
School of Computer Science and Engineering
Bharathiar University
Coimbatore, Tamil Nadu, INDIA

ABSTRACT
Data mining is the extraction of the hidden information from large databases. It is a powerful technology with new great potential to analyze important information in the data warehouse. Preserving privacy against data mining algorithms is a new research area. It investigates the side-effects of data mining methods that derive from the privacy diffusion of persons and organizations. Privacy preserving data mining is the emerging field that protects sensitive data. Classification is one of the popular techniques of data mining. Classification is a data mining technique used to predict group membership for data instances. Classification involves finding rules that partition the data into disjoint groups. Many classification rule algorithms are used to generate the classification rules such as OneR, Ridor, and Conjuctive Rule. In this paper, we focus on the problem of privacy preservation in classification rules. The rule based classification algorithms namely C4.5, Ripper and Part algorithms are used for generating rules. The privacy is preserved by hiding the sensitive rules and the new dataset is reconstructed from the non sensitive rules. In this paper the experimental results shows the effectiveness of each algorithm.

General Terms
Data Mining, Classification, Algorithm.

Keywords
Rules, C4.5, Ripper, Part.

1. INTRODUCTION
Data mining is the process of extracting hidden patterns from large data sets. This can be achieved by combining methods from statistics and artificial intelligence with database management. Data mining gives its great role in the analysis of collections of observations of behavior [10]. Data mining is an automatic search for hidden patterns databases which is large. Data mining is an iterative process in which the progress is defined by discovery, either automatically or manually. Data mining is very useful in an exploratory analysis scenario in which there are no predetermined notions about what will constitute an "interesting" outcome. Data mining is one of the fastest growing fields in the industry of computer [9]. Data mining is reflected in its wide range of methodologies and techniques in which it is applied to a connection of problem sets. It is a natural activity to be performed on large data sets and can be performed on data represented in quantitative, textual, or multimedia forms [17].

The field of privacy has seen rapid advances in recent years because of the increases in the ability to store data. In particular, recent advances in the data mining field have lead to increased concerns about privacy [6]. Privacy Preserving Data Mining provides a comprehensive overview of available approaches, techniques and open problem in privacy preserving data mining. It is designed for a professional audience composed of practitioners and researchers in industry [2]. Privacy Preserving Data Mining (PPDM) concentrates in combining and mining databases to preserve the privacy of the individual parties' data. Privacy concerns for individuals are rapidly gaining attention [15]. Research has focused on Privacy Preserving Data Mining (PPDM). It uses various techniques, statistical, cryptographic [16] and others, to facilitate cooperative data mining. It also protects the privacy of the organizations or individuals involved. PPDM methods may address individual privacy of specific customers or collective privacy, the privacy of information about an organization's records overall [5]. Privacy preserving data mining is an important property that any mining system must satisfy. Privacy preserving protocols are designed to preserve privacy even in the presence of adversarial participants.

Privacy preserving protocols are designed in order to preserve privacy even in the presence of adversarial participants that attempt to gather information about the inputs of their peers [4]. There are different levels of adversarial behavior. Privacy preserving data mining techniques clearly depend on the definition of privacy [8]. Privacy captures the sensitive information in the original data and it is protected from either direct or indirect disclosure. Privacy preserving data mining aims at providing a trade-off between sharing information for data mining analysis. It also protects information to preserve the privacy of the involved parties [7].

The rest of the paper is organized as follows. In Section 2 concept of classification is given. In Section 3 classification rule hiding and the related works are discussed. Section 4 gives the general problem formulation. In Section 5, the proposed solution gives the details of the dataset, generation of classification rules, privacy for the sensitive rules etc. The effectiveness of the algorithm is evaluated and the experimental results of the proposed technique are discussed in Section 6. Conclusions are given in Section 7.

2. CLASSIFICATION
In classification mining, a set of database tuples act as a training sample and it is analyzed to produce a model of the data that can be used as a predictive classification method for classifying new data into classes. Goal of the classification process is to build a

model that can be used to further classify tuples being inserted and that represents a descriptive understanding of the table content. One of the most popular classification mining techniques is represented by decision trees. Each internal node of a decision tree is associated with an attribute on which the classification is defined. Each outgoing edge is associated with a split condition representing how the data in the training sample are partitioned at that tree node [2]. The form of a split condition depends on the type of the attribute. Each node contains information about the number of samples at that node and how they are distributed among the different class values. The quasi-identifier attributes correspond to internal (splitting) nodes in the tree, edges are labeled with attribute values instead of reporting the complete split condition, and nodes simply contain the number of respondents classified by the node values.

Classification refers to the problem of categorizing observations into classes. Predictive modeling uses samples of data for which the class is known to generate a model for classifying new observations. Classification deals in generating rules that partition the data into disjoint groups. It is a technique in data mining used to predict group membership for data instances. 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 [20]. Test data are used in the classification rules to estimate the accuracy. Classification is the identification of new patterns.

A classification rule is a procedure in which the elements of the data 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.One of the essential tasks of data mining and machine learning research is building accurate classifiers. Given a set of training instances with known class labels, classifiers aim to predict the target classes for a set of test instances for which the class labels are not known [3]. At first, a classification model is developed from training data and then it is used to classify unseen instances. There are various methods for building classifiers such as decision trees, naive-Bayesian methods, statistical approaches, support vector machines etc. Such classifiers are CBA (Classification Based on Association), CAEP (Classification based on Aggregating Emerging Patterns and CMAR (Classification based on Multiple Association Rules). These approaches have higher accuracy compared to that of the decision tree classifier since decision-tree classifier examines one variable at a time.

3. RELATED WORK

Classification rule hiding algorithms considers a set of classification rules as sensitive and the protection is given either by suppression-based or reconstruction-based techniques [11]. The aim of suppression-based technique is to reduce the confidence of a sensitive classification rule by distorting some attributes in the dataset. A blocking technique, called parsimonious downgrading, and blocks the inference channel that leads to the identification of the sensitive rules by selectively sanitizing transactions so that missing values appear in the released dataset. This has an immediate consequence in the lowering of the confidence for the holding of the sensitive rules. The proposal of heuristic approach achieves to fully eliminate all the sensitive inferences, while effectively handling overlapping rules. The algorithm identifies the set of attributes that influence the existence of each sensitive rule and removes them from the supporting transactions that affect the non-sensitive rules. The reconstruction-based approaches targets at reconstructing the dataset by using only supporting transactions of the non-sensitive rules. These approaches are advantageous over the heuristic data modification approaches, since they hardly introduce any side-effects .This approach first performs a rule based classification of the original dataset to enable the owner of the data to identify the sensitive rules. Then, it proceeds to construct a decision tree that is constituted only on non-sensitive rules approved by the data owner. The constructed dataset remains similar to the original one, except from the sensitive part. The difference between the two datasets is proved to reduce as the number of rules increases. The quality of the reconstructed dataset can be further improved. Following papers which are based on classification rule hiding are reviewed.

The paper “Hiding Classification Rules for Data Sharing with Privacy Preservation” [18] has introduced a method of hiding sensitive classification rules from data mining algorithms for categorical datasets. In this approach the dataset is reconstructed according to the classification rules that have been checked and agreed by the data owner for releasing to data sharing. This method reconstructs a new dataset by using only the non-sensitive rules. A rule-based classification algorithm was used to extract classification rules of an original dataset. This approach archives high usability of reconstructed datasets.

The paper “A data perturbation approach to sensitive classification rule hiding” [1] focuses on privacy preservation in classification rule mining. The subject at hand is approached through the proposition of a data perturbation approach for hiding sensitive classification rules in categorical datasets. Such a methodology is absolutely necessary in case the data needs to be published on the web so that it is amply available for public use as opposed to other approaches like output perturbation or cryptographic techniques that restrict the usability of the data in different ways. It is ensured that not only the sensitive rules are hidden but also that the current structure of the rule set, thus the information value of the dataset, is preserved. Moreover a modification of the basic method which exhibits an alternative distribution procedure is also presented. Finally, a series of experiments are executed in order to evaluate the validity and effectiveness of the proposed approaches against existing similar ones.

The paper “Data reduction approach for sensitive associative classification rule hiding” [12] explores an alternate approach for sensitive pattern hiding problem, data reduction, i.e. removing the whole selected tuples. By the reduction in data, every tuple in modified data sets is real data in which there is no change in it. The focused pattern type is classification rule in which everything is associative. The number of false-dropped rules and ghost rules denotes the impact on data quality. The experiments are conducted for the evaluation of the approach and the results have shown that data reduction approach can produce data sets with high data quality, thus it is applicable to the problem.

In this paper “A Heuristic Data Reduction Approach for Associative Classification Rule Hiding” [13] the problem of sensitive classification rule hiding by using data reduction

approach is addressed. It focuses on a specific type of classification rules, i.e. associative classification rules. It has given some of the observations on the data quality with regard to the data reduction processes. From the observations, the impact by each reduction precisely without any re-applying the classification algorithm is represented. Subsequently, a heuristic algorithm to hide the sensitive rules based on the observations has been proposed. Experimental results are shown to denote the effectiveness and the efficiency of the proposed algorithm.

The paper "Associative classification rules hiding for privacy preservation" [14] addresses a problem of sensitive classification rule hiding by a data reduction approach. This focuses on an important type of classification rules, i.e., associative classification rule. In this paper, the impact on the generation of data quality by data reduction processes is represented by the number of false-dropped rules and ghost rules. To address the problem, proposals of few observations on the reduction approach are given. Subsequently, a greedy algorithm is proposed for the problem based on the observations. Also, two-bitmap indexes are applied to improve the efficiency of the proposed algorithm. Experiment results are presented to show the effectiveness and the efficiency of the proposed algorithm.

4. PROBLEM FORMULATION

4.1. Formulation of Classification Rule

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 sensitive rules R^1 from the set of rules R. Reconstruct a new dataset D^1 using the non sensitive rules R^2, find performance factors such as Hidden failure, Misses Cost, Side effect and Efficiency.

5. PROPOSED SOLUTION

The following steps are required for the proposed solution.

Step 1: Consider a dataset of Wisconsin Breast Cancer and Heart Disease with set of items.
Step 2: C4.5, Ripper and Part algorithm are used to create rules.
Step 3: From the decision tree, the set of classification rules can be generated based on the threshold value.
Step 4: Select and hide the sensitive rules from the set of classification rules.
Step 5: Reconstruct new dataset with the non sensitive classification rules.
Step 6: Verify (i) Hidden, (ii) Misses Cost, (iii) Side effect, (iv) Efficiency.

5.1. Dataset

Dataset is collected from the website www.ucirepository.com. There are various types of datasets in this website namely Diabetes, Hepatitis, credit card approval etc Chess etc. In this work two dataset are used namely,

- Wisconsin Breast Cancer dataset
- Heart Disease dataset.

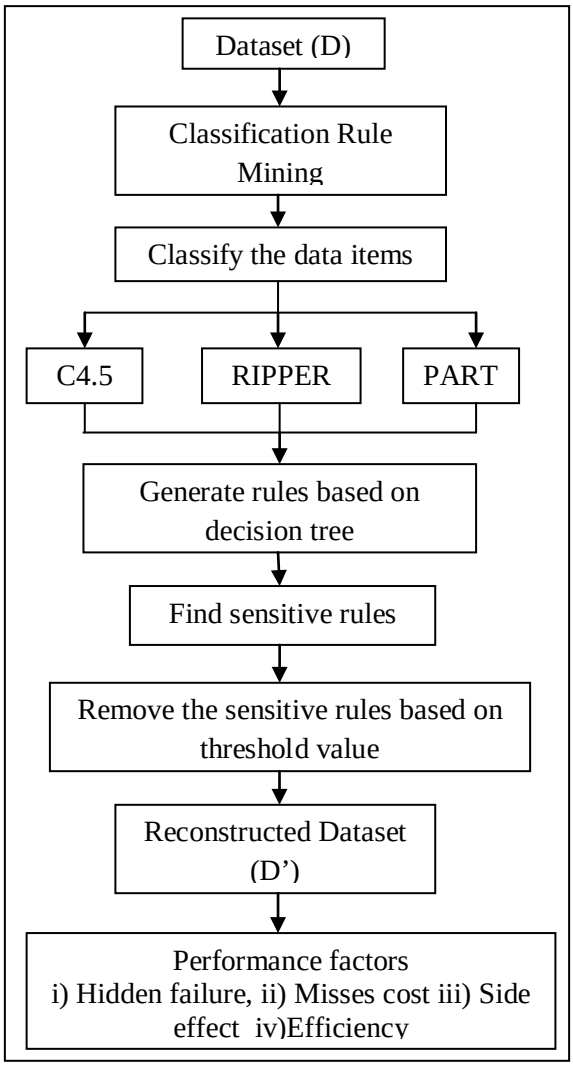

Fig 1: System Architecture

5.1.1. Wisconsin Breast Cancer dataset

This dataset contains 699 instances and 10 attributes plus 1 class attribute. The attributes are as follows: Sample code number, Clump Thickness, Uniformity of Cell Size, Uniformity of Cell Shape, Marginal Adhesion, Single Epithelial Cell Size, Bare Nuclei, Bland Chromatin, Normal Nucleoli, and Mitoses. The characteristic of attribute is Integer.

5.1.2. Heart Disease dataset

This dataset contains 303 instances and 10 attributes plus 1 class attribute. The attributes are as follows: id, ccf, age, sex, pain loc, painexer, relrest, pncaden, chest pain type, smoke, years,cigarettes per day, xhypo, etc.

5.2. Classification Rule Algorithm

5.2.1. C4.5 Algorithm

C4.5 is an algorithm that takes a set of labeled data as input and creates a decision tree as a result. 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 in the form of decision trees from a set of given examples.

C4.5 is a software extension of the basic ID3 algorithm and it was designed by Quinlan. C4.5 is one of best known and most widely-used learning algorithms. Algorithm C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy. The training data is a set S = s1, s2, of already classified samples. C4.5 is often referred to as a statistical classifier [19]. The most important part of the C4.5 algorithm is the process of generating an initial decision tree from the set of training samples.

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 2: Algorithm for C4.5

The skeleton of the C4.5 algorithm is based on Hunt's CLS method for constructing a decision tree from a set T of training samples. To make a decision-tree model more readable, a path to each leaf can be transformed into an IF-THEN production rule.

5.2.2. RIPPER

RIPPER stands for 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.

1. Ripper(Pos, Neg, k)
2. RuleSet ← 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 ← Learn Rule(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 3: Algorithm for Ripper

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

5.2.3. PART

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

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 noncommitted, 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 4: Algorithm for Part

5.3. Classification Rules

The classification rules are generated using an important concept known as decision tree. Decision tree gets the input as the dataset items. The decision tree has nodes namely root node and leaf nodes. The attributes of the data items are placed in the appropriate nodes by entropy calculation.

The class attribute is the leaf nodes. Entropy is a formula to calculate the homogeneity of a sample.A completely homogeneous sample has entropy of 0.An equally divided sample has entropy of 1. .

The formula for entropy is

$$\text{Entropy}(S) = \sum_{i=1}^{c} p_i \log_2 p_i$$

The example of sample dataset is as follows:

Table 1. Sample dataset

Sample code number	Bare Nuclei	Bland Chroma tin	Mitoses	Class
1	3	1	1	2
10	3	2	1	2
2	3	1	1	2
4	3	7	1	2
1	3	1	1	2
10	9	7	1	4
1	3	1	1	2

After the process of generating decision tree, the rules are created from the decision tree using if-then rules. Following rule is generated with the data item present in the above sample dataset.

Bare Nuclei = 9 ^ Mitoses = 1 ⟶ Class = 4

5.4. Sensitive Rules

Generally classification rules contain sensitive and non sensitive rules. The sensitive rules are identified using a threshold value. The identified sensitive rules are hidden by removing the sensitive rules from the non sensitive rules.

5.5. Reconstruction of new dataset

A new dataset is reconstructed using the non sensitive rules which remained after hiding the sensitive rules. After reconstruction the rules are generated for testing whether any sensitive rules has occurred leading to any side effect. The experimental results of this process are discussed in the following section.

6. EXPERIMENTAL RESULTS

6.1 Analysis of Results

The experimental results are analyzed based on the following performance factors.

i. Hiding Failure
ii. Misses Cost
iii. Side Effect
iv. Efficiency

6.1.1. Hiding Failure

This measure quantifies the percentage of the sensitive patterns that remain exposed in the sanitized dataset.

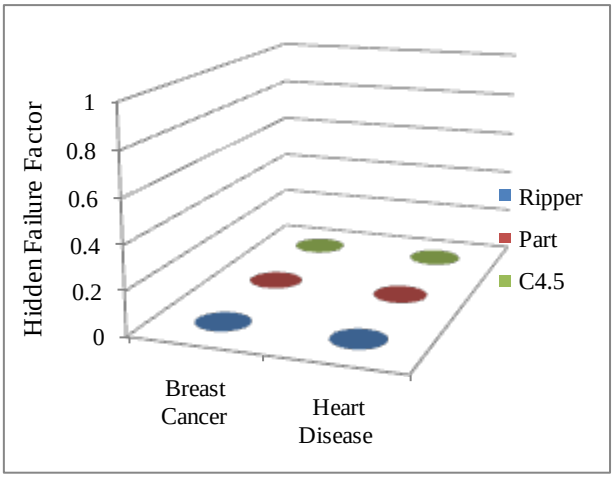

Fig 5: Hidden Failure

The above chart represents the Hiding Failure factor as 0% obtained as a result of the reconstruction based technique using Wisconsin Breast Cancer Dataset and Heart Disease Dataset.

6.1.2. Misses Cost

This measure quantifies the percentage of the nonrestrictive patterns that are hidden as a side-effect of the sanitization process.

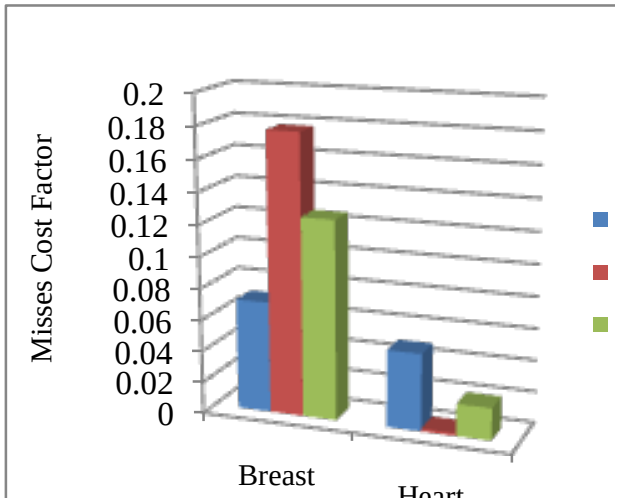

Fig 6: Misses Cost

The above chart represents the Misses Cost of the proposed technique using Wisconsin Breast Cancer Dataset and Heart Disease Dataset.

6.1.3. Side Effect

Similarly to the measure of misses cost, the side effect factor is used to quantify the amount of non-sensitive association rules that are removed as an effect of the sanitization process.

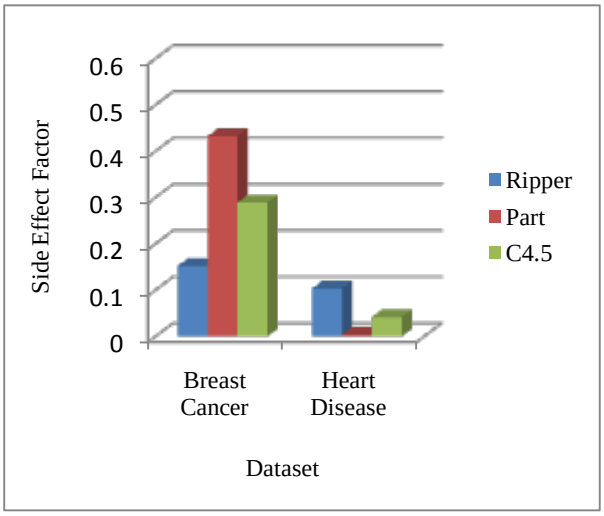

Fig 7: Side Effect

The above chart represents the percentage of the number of non sensitive rules removed as a result of sanitization process using Wisconsin Breast Cancer Dataset and Heart Disease Dataset.

6.1.4. Efficiency

This category consists of measures that quantify the ability of a privacy preserving algorithm to efficiently use the available resources and execute with good performance.
Efficiency can be measured using the two measures namely,

a. Time
b. Accuracy

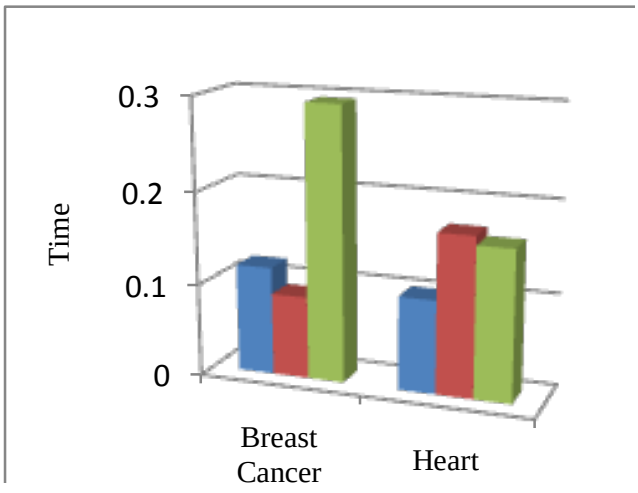

Fig 8: Time plot

The above chart represents the time complexity of C4.5, Ripper and Part algorithms using Wisconsin Breast Cancer Dataset and Heart Disease Dataset.

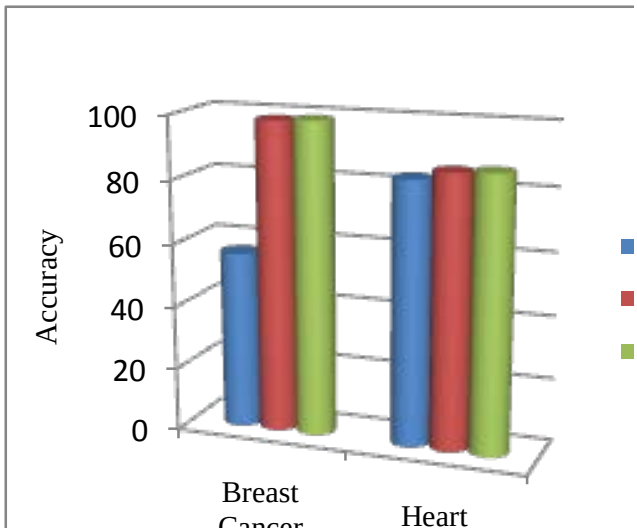

Fig 9: Accuracy plot

The above chart represents the accuracy plot of C4.5, Ripper and Part algorithms using Wisconsin Breast Cancer Dataset and Heart Disease Dataset.

7. CONCLUSION

Classification rule is one division of data mining technique. Other data mining techniques should also be considered for securing both data and knowledge. In this proposed work, rule based classification algorithms such as C4.5 algorithm, Ripper algorithm and PART algorithm are used for generating rules in classification rule mining.

The privacy is preserved by removing the sensitive rules from the rules to reconstruct the new dataset from the non sensitive rules. The performance of both the techniques are compared with the following factors such as Hidden failure, Misses cost, Side effect and Efficiency. The efficiency of each algorithm is determined based on the Accuracy and Time factor. The conclusion of this research work is privacy is preserved for the classification rules by reconstructing the new dataset by removing the sensitive rules. In future, other techniques can be used to reconstruct the new dataset.

8. REFERENCES

[1] Aggelos Delis , Vassilios S. Verykios , Achilleas A. Tsitsonis ,"A data perturbation approach to sensitive classification rule hiding", SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing, ACM New York, NY, USA ©2010.

[2] Agrawal, R. & Srikant, R. (2000), Privacy-preserving data mining, in 'Proceedings of the 2000 ACM SIGMOD international conference on Management of data', ACM Press, pp. 439–450.

[3] Atallah, M., Elmagarmid, A., Ibrahim, M., Bertino, E. & Verykios, V. (1999), Disclosure limitation of sensitive rules, in 'KDEX '99: Proceedings of the 1999 Workshop on Knowledge and Data Engineering Exchange', IEEE Computer Society, Washington, DC, USA, pp. 45–52.

[4] Chen, X., Orlowska, M. & Li, X. (2004), A new framework of privacy preserving data sharing, in 'Proceedings of 4th IEEE International Workshop on Privacy and Security Aspects of Data Mining', IEEE Press, pp. 47–56.

[5] Clifton, C. & Estivill-Castro, V., eds (2002), IEEE ICDM Workshop on Privacy, Security and Data Mining, Vol. 14 of Conferences in Research and Practice in Information Technology, ACS.

[6] Clifton, C. & Marks, D. (1996), Security and privacy implications of data mining, in 'Workshop On Data Mining and Knowledge Discovery', University of British Columbia Department of Computer Science, Montreal, Canada, pp. 15–19.

[7] Domingo-Ferrer, J. & Torra, V., Eds (2004), Privacy in Statistical Databases, Vol. 3050 of LNCS, Springer, Berlin Heidelberg.

[8] Estivill-Castro, V., Brankovic, L. & Dowe, D. L. (1999), 'Privacy in data mining', Privacy – Law and Policy Reporter 6(3), 33–35.

[9] Han, J. and M. Kambert, es, "Data Mining: Concepts and Techniques", Morgan Kaufmann, San Francisco, 2000.

[10] Jeffrey W. Seifert, "Data Mining An Overview", Analyst in Information Science and Technology Policy Resources, Science, and Industry Division.

[11] Juggapong Natwichai, Xue Li, Maria E. Orlowska "A Reconstruction-based Algorithm for Classification Rules Hiding", Seventeenth Australasian Database Conference (ADC2006), Hobart, Australia. Conferences in Re- search and Practice in Information Technology (CRPIT), Vol.

[12] Juggapong Natwichai, Xingzhi Sun, Xue Li, "Data reduction approach for sensitive associative classification rule hiding", ADC '08 Proceedings of the nineteenth conference on Australasian database - Volume 75, Australian Computer Society, Inc. Darlinghurst, Australia, Australia ©2007

[13] Juggapong Natwichai,Xingzhi Sun, Xue Li, "A Heuristic Data Reduction Approach for Associative Classification Rule Hiding", IBM Research Laboratory, Beijing, China.

[14] Juggapong Natwichai, Xingzhi Sun, Xue Li, "Associative classification rules hiding for privacy preservation", IBM Research Laboratory, Beijing, China.

[15] Jr., R. J. B. & Agrawal, R. (2005), Data privacy through optimal k -anonymization, in 'Proceedings of the 21st International Conference on Data Engineering', IEEE Computer Society, pp. 217– 228.

[16] Lindell, Y. & Pinkas, B. (2000), Privacy preserving data mining, in 'Proceedings of the 20th Annual International Cryptology Conference on Advances in Cryptology', Springer-Verlag, pp. 36– 54.

[17] Mary DeRosa. "Data Mining and Data Analysis for Counterterrorism".

[18] Natwichai, J., Li, X. & Orlowska, M. (2005), Hiding classification rules for data sharing with privacy preservation, in 'Proceedings of 7th International Conference on Data Warehousing and Knowledge Discovery', Lecture Notes in Computer Science, Springer, pp. 468–467.

[19] Quinlan, J. R. (1993), C4.5: Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, USA.

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