Max-Min Approach for Hiding Sensitive Association Rules

Submitted by:

Umm-e-Asma

Reg. # 576-FBAS/MSCS/F09

Supervised by:

Mr. Muhammad Imran Saeed

Department of Computer Science & Software Engineering

Faculty of Basic and Applied Sciences

International Islamic University, Islamabad

2012.

In The Name of

Allah Almighty

Most Beneficient, The Most Merciful

Department of Computer Sciences & Software Engineering
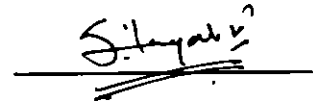## International Islamic University, Islamabad, Pakistan.

Dated: 26ᵗʰ March 2012

Final Approval

It is certified that we have examined the thesis titled **"Max-Min Approach for Hiding Sensitive Association Rules"** submitted by *Ms. Umm-e-Asma*, **Registration No.** *576/FBAS/MSCS/F09*. It is our judgment that this project is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for Masters Degree in Computer Science (MSCS).
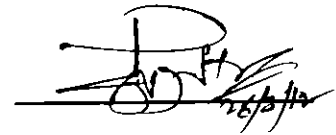
Committee

External Examiner:
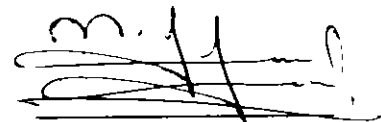Dr. Sikandar Hayat Khayal,
Head of Acad (ES), APCOMS,
Rawalpinidi.

Internal Examiner:
Dr. Ayyaz Hussain,
Assistant Professor,
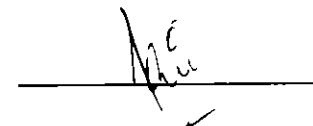Department of Computer Sciences & Software Engineering,
IIU, Islamabad.

Supervisor:
Mr. Muhammad Imran Saeed
Assistant Professor,
Department of Computer Sciences & Software Engineering,
International Islamic University, Islamabad.

Co- Supervisor:
Ms. Zakia Jalil
Lecturer,
Department of Computer Sciences & Software Engineering,
International Islamic University, Islamabad.

Department of Computer Sciences & Software Engineering

International Islamic University, Islamabad.

Dated: .. . .

Final Approval

It is certified that we have examined the thesis titled "Max-Min Approach for Hiding Sensitive Association Rules" submitted by *Ms Umm-e-Asma*, Registration No. *576/FBAS/MSCS/F09*. It is our judgment that this project is of sufficient standard to warrant its acceptance by the International Islamic University, Islamabad for Masters Degree in Computer Science (MSCS).

Committee

External Examiner:

_____

Internal Examiner:

_____

Supervisor: _____

Mr. Muhammad Imran Saeed
Assistant Professor,
Department of Computer Sciences & Software Engineering
International Islamic University, Islamabad.

A dissertation submitted to the

Department of Computer Science & Software Engineering,

International Islamic University, Islamabad

As a partial fulfillment of the requirements

For the award of the degree of

Masters in Computer Science

Declaration

I hereby declare that this software, neither as a whole nor as a part thereof has been copied out from any project. It is further declared that I have carried out this research entirely on the basis of my personal efforts made under the sincere guidance of my teachers and supervisor. No portion of the work presented in this report has been submitted in support of any application for any other degree or qualification of this or any other university or institute of learning. If any plagiarism is found in this report, my supervisor/examiners will not be responsible.

Umm-e-Asma

576/FBAS/MSCS/F09

Dated: _____

Acknowledgement

All praise to Almighty Allah, who gave me the understanding, courage and patience to complete this project.

I express my gratitude to my kind project supervisor **Mr. Muhammad Imran Saeed,** who provided me opportunity to learn and enhance my knowledge. As my project supervisor, he had been ready to help and guide me throughout the research work.

I would also like to thank my teachers in the department especially **Ms. Zakia Jalil,** for the moral support of my efforts.

And last of all I would like to acknowledge the support of my family members and friends. I would like to admit that I owe all my achievements to my truly, sincere and most loving parents, whose prayers are a source of determination for me throughout my educational career.

In the end I want to present my unbending thanks to all those hands that prayed for my betterment and serenity.

Umm-e-Asma

(576/FBAS/MSCS/F09)

Abstract

The past decade has seen an enormous growth in database technology and the amount of data gathered. This massive propagation of databases in majority areas of human endeavor has fashioned a great demand for novel, powerful tools for converting data into useful and task-oriented knowledge. In the efforts to fulfill this need, researchers have been trying to explore ideas and devise new methods and techniques in statistical data analysis, machine learning, neural nets, pattern recognition, data visualization, etc. These efforts have led to the emergence of a novel research area, routinely called data mining and knowledge discovery (extraction).

Data mining is basically concerned with picking out hidden associations present in business data to let the businesses construct predictions for future use. It is the process of data-driven mining of not so palpable but valuable information from large databases. Basic aim of this field is to dig out hidden, previously unknown and useful (or actionable) knowledge patterns from the data.

Privacy Preserving Data Mining is a new dimension in data mining, dealing with hiding confidential knowledge. When Statistical database is released the data owner may not want to publish all the data. Many techniques have been proposed in recent years to hide sensitive data. Researchers are concerned with changing the database in a way that all sensitive data or knowledge is hidden with minimum impact on insensitive data. The work here is a new technique for Hiding Sensitive Association Rules. The methodology is extension of already existing algorithm, Max-Min algorithm for hiding item sets and is based upon MAXMIN approach of decision theory. The proposed technique will be checked for the better performance in Privacy Preserving data mining.

Comparison of the existing techniques with the proposed one is given in this work.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1. Introduction

Data mining has attracted a great deal of attention in the information industry in current years, due to the availability of huge amounts of data and the imminent need for turning such data into useful information and knowledge. Intense work has been done in the field of data mining to develop the new techniques for the efficient analyses of the large data sets The knowledge gained after analyzing the data is being used for decision making in different domains like marketing, fraud detection, scientific discoveries, reading social behaviors of a community and much more Thus data mining has a lot of advantages, but large repositories of data which are used to extract patterns useful for decision making may contain data that is private for an individual or an organization. There is strong need to protect this private data because it becomes an ethical issue if the decision making affect someone's settlement.

It gives rise to a new research domain in data mining known as Privacy Preserving Data Mining (PPDM), according to which one must be able to analyze the data for its own interest without affecting other's. Privacy in data mining can be attained at two levels. [13]

- Data hiding

- Knowledge Hiding

In data hiding researchers are hiding private data directly; and thus data is not involved in decision making at any level.

And knowledge hiding private data may be used for extracting knowledge and then protecting that knowledge from being published. In data mining researchers are mainly concerned with knowledge hiding.

Knowledge in data mining can be derived in different forms depending upon the technique used to extract knowledge. in the proposed technique, concern is with the Association Rules which are formulated from frequent items. Since the introduction in 1993[26], the association rule mining has received an enormous attention. It is still one of most popular knowledge discovery methods in the field of data mining. Various techniques and algorithms have been

designed for it in recent years for efficient extraction of knowledge through association rules and also for hiding sensitive rules

While designing any technique for hiding the sensitive knowledge most important issues which must be resolved is the side effects produced as a result of hiding technique i.e. non-sensitive data should remain impassive in technical term it is said that while preserving privacy quality of data is also preserved. In this work an extension to already existing work is given. In already existing technique sensitive frequent item sets are hide. here algorithm is proposed which will directly hide association rule while maintaining data quality.

1.1. **Motivation**

Privacy has natural tradeoff with accuracy, if someone wants to ensure the privacy he has to compromise the accuracy of the data and vice versa. Most of the work done in the field is focused on hiding the data while reducing the side effects like generation of new knowledge, lost of existing one, more CPU usage and hiding failure.

Most of the existing techniques work with selecting frequent items sets as sensitive and hiding them, little work is done in the area of hiding association rules directly. The techniques given so far for hiding association rules has many limitations like some techniques are not scalable according to the number of items in the rule or technique fails to hide the data if certain constraints like local size of the data or size of the data to be hidden.

Work is needed in the field which promises privacy of our data with limited side effects on knowledge. The aim of this study is to recognize a technique for hiding association rules with zero hiding failure and least side effect on remaining data.

1.2. **Objectives and Contributions**

The main objective of this thesis is get insight into the issues related to privacy preserving data mining and to propose a system which hide the sensitive data with limited side effects. The main contributions of this thesis are that it overcomes the

problem new rules generated during the sanitization process and it also ensures zero hiding failure.

Hiding process of sensitive association rules can be done by decreasing confidence of the rules, the objective can be achieved in two ways, either by decreasing the support of frequent item set as whole or by increasing support of L.H.S. elements of the rule. Here in the thesis data screening is achieved using the later technique but in slightly different way.

No new rules are generated during the process as privacy preservation is achieved by introduction of new transactions in the data base instead of modifying the existing one.

1.3. Goals and Challenges

The main goal of the research is to suggest a new technique of data sanitization which can further be used to develop a real time-time application in data mining.

The main challenges are to find out such rule hiding method which gives higher accuracy of statistical data and also ensure privacy.

1.4. Key Points

Important terms and concepts which are necessary to understand the work done in this research are given in this chapter.

1.4.1. Data Mining

Data is collected everyday with amount of data doubling after very short time. Data mining is a technique that helps us to extract useful knowledge from a large database By useful knowledge we mean those patterns present in our data which were unknown before; and can be useful to us in making further decisions

Data mining has its applications in almost every field because with this enormous data some statistical measures are needed which help us to analyze the data. Most important application areas of data mining as presented in [25] are:

- Prediction and description

- Relationship Marketing

- Customer Profiling

- Outlier identification and detecting Frauds

- Customer segmentation

- Web site design and promotion

Complex data mining algorithms and techniques are used for extracting useful information. These techniques may vary according to the nature of data and user requirements. Mainly data mining techniques are categorized as follows:

- Association rule mining also known as market basket analysis

- Supervised Classification

- Cluster Analysis

- Web data mining

Data mining technique used in system proposed in this study is association rule mining.

1.4.2. Association Rule Mining

The aim of the association rule mining is to find out which database values are always associated with each other. This can be best understood by an example. Suppose transactional database of the super mart; the association rule mining can be used to find

out those items which are most frequently bought together. This can be helpful for the mart owner in deciding further business strategies.

Association rule mining has many other applications other than market basket analysis·

- Electronic Commerce

- Marketing

- Social Networking

- Health

- Bioinformatics

Association Rules are derived from frequent item set.

Frequent item sets can be stated as:

Frequent Item Set X is the item set that has support above user defined minimum support threshold (MST). [26]

$$Support\ (X) = \frac{|X|}{N} \geq MST \dots\dots\dots (2.1)$$

Association rules are represented as an implication of $X \rightarrow Y$ where both X and Y are frequent item/item set. X is referred to as rule's antecedent and Y as consequent. **Interesting/Strong Rules** are those rules which have confidence above user defined minimum confidence threshold (MCT). [26]

$$Confidenc\ (X \rightarrow Y) = \frac{Support\ (XY)}{Support\ (X)} \geq MCT \dots\dots\dots (2.2)$$

Many algorithms have been proposed for mining interesting knowledge in the form of association rules; some important mostly used algorithms are:

- Naïve Algorithm

- Apriori Algorithm

- Direct Hashing and Pruning Algorithm

In this work *apriori algorithm* is used, proposed in [27] to mine association rules and then hide selected sensitive rules from them.

1.4 3 Apriori Algorithm

Apriori[27]was proposed by Agrawal and Srikant in 1994. The algorithm works by finding the frequent set in the database. The algorithm is in fact a bottom up search in which an item set lattice is formed. In the lattice frequent item set are marked while moving upward in each level. It prunes many of the sets which are unlikely to be frequent sets, thus saving any extra work.

1 4 4. Downward closure property

According to the algorithm the subset of a frequent item set are always frequent and make a sub lattice of the original lattice. As shown in the Figure 2.1. And opposite to it superset of infrequent item set are always infrequent.
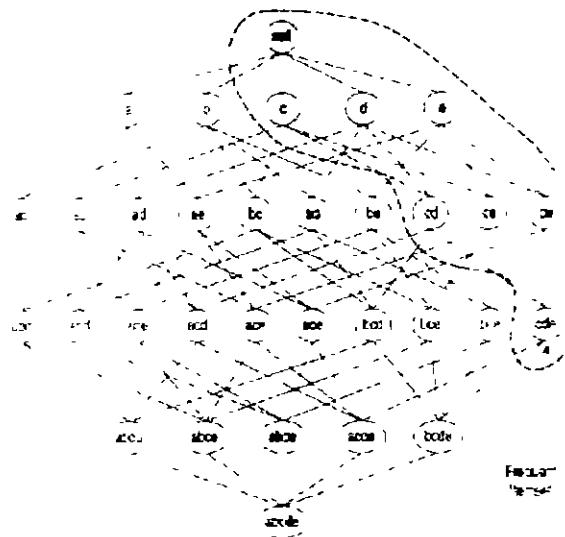
Figure 2.1: Apriori Lattice with frequent item set [31]

1.4.5 Maximal Frequent Item sets

It is support threshold that divide the frequent item set lattice to frequent or infrequent. The frequent item set which cannot be extended to the next level as their super sets are infrequent are called Maximal frequent item sets. Maximal frequent item set can be used to derive all its frequent subset using downward closure property. As marked with square in the Figure 2.2.

1.4.6. Minimal Item sets

Infrequent item sets which have all of its subsets as frequent are called minimal frequent item sets. Minimal item sets are encircled in Figure 2 2
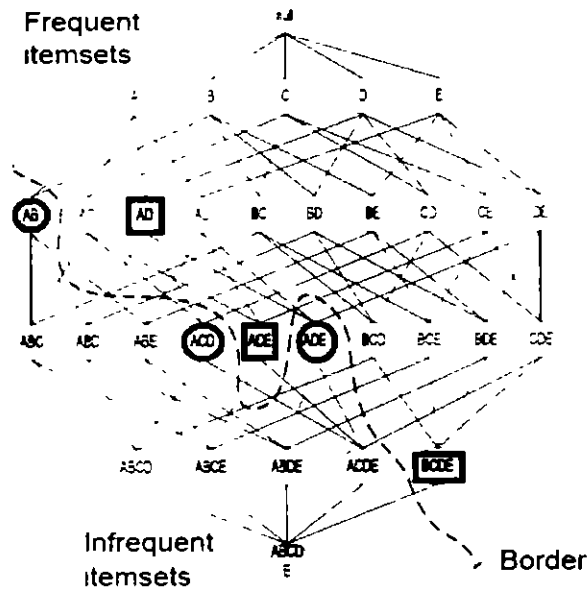
Figure 2.2: Borders in Apriori lattice [31]

1.4.7. Border Theories in Frequent item set

The formation of border in item set lattice due to maximal property was used by H. Mannila and H Toivonen [24]. According to the border theory item set lattice formed during apriori algorithm can be divided into positive and negative border. All the

maximal elements are said to be in positive border and all minimal item sets are said to in negative border.

1.4.8. Privacy preserving data mining

Protection of sensitive data against unauthorized access has always been a goal for the database. Current advancement in data mining technologies has increased the security risks of sensitive and statistically important data. Hence, the security issue has become a much more important area of research

This new dimensions of data mining is termed as Privacy Preserving Data Mining (PPDM). In PPDM sensitive knowledge is preserved from being discovered by unauthorized users.

Different PPDM algorithms have been proposed so for depending upon the type of technique used to mine the data. This study is going to deal with hiding sensitive association rule hiding. Algorithms presented in the area can be classified [13] as given in Figure 2.3
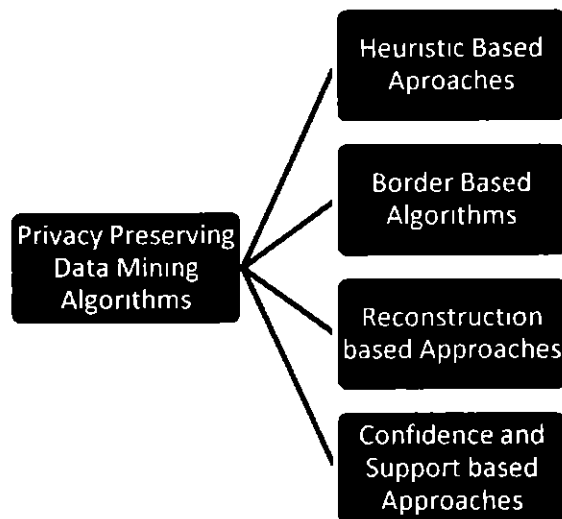
Figure 2.3: Classification of PPDM Algorithm

1.4.9. Sensitive Association Rule

Sensitive Association Rule `S` is a set of rules which is strong enough to be interesting but data owner does not want it to publish so that it may not be used for decision making by some third party. Criteria for identifying sensitive rules depend upon user's scenario in statistical database and also on user's social environment and ethical issues.

- **Association Rule Hiding**

It is a procedure to transform a database D into D' such that user is not able to mine 'S' from D'.

"Let R be the set of all the interesting rules mined from database D under certain condition of MCT and MST and there is a set of sensitive rules S and S ⊆ R We need to transform D to D' such that S cannot be mined from D' under same conditions of MST and MCT We say that S is safely hidden if we are able to discover R-S from D' i e no extra rules (ghost rules) are generated and no rules other than S (lost rules) are hidden under same MST & MCT"

1.4.10   Border Revision Theories

Border theories were further used to hide sensitive knowledge by Sun et al. [22] and Moustakides et al. [23]. The technique introduced uses the concept of border revision theory to discover positive and negative border elements from frequent item set lattice and then applies decision theory for maximizing the minimum gain while hiding sensitive knowledge.

1.5. Overview of the Manuscript

In this section overview of the remaining contents of this manuscript which is structured into five main chapters.

Chapter 2 presents an overview of different key points related to frequent item set association rules and borders in frequent item set lattice

Chapter 3 discusses previous work of researchers and the problems in state-of-art, it recalls the previous work related to knowledge hiding. The problems in the state-of-art are also discussed in this chapter.

Chapter 4 detail scheme of the proposed system is discussed. It contains the description of technique followed to hide association rules.

Chapter 5 discuss in detail about the datasets used for experimentation and which performance measures are used to evaluate the results. The evaluated results are also discussed in detail.

Chapter 6 consists of a conclusion where a review of the applications and future work are also presented.

1.6. Summary

The chapter review basic concept of association rule mining which will be used in technique proposed here for hiding purpose. It also formulates the problem of hiding association   rules.

Chapter 2

Literature Review

2. Literature Review

The Association Rule Hiding is done by either hiding frequent item sets or association rules directly. A lot of work has been done over the past few years which made improvements by making use of different methods but some of the flaws still exist in these techniques. Some of the previous efforts of the researchers in area also discussed below:

2.1. Association Rule Mining

Different techniques have been proposed to efficiently mine association rule from statistical databases, some of these are discussed below which are further used in proposed system.

2.1.1. Discovering Rules

The problem of mining association rules from a large database of customer transactions was first time introduced by Agerwal et al[26]. Support and confidence were used to extract interesting knowledge. The problem of knowledge extraction was decomposed into two sub problems:

- Finding frequent item sets that have support greater than threshold support

- Generating association rules that have confidence greater than threshold confidence.

Feasibility problem of finding frequent item sets was explained as it is computationally impossible to create all the possible item sets and then calculate its support. The algorithm proposed in this paper runs judgment process to decide what item set should be measured in a pass so it keeps stability between number of passes and number of item sets that are generated in a single pass. Moreover to ensure completeness pruning technique is used that help to neglect those item sets which are not going to be frequent so in this way algorithm helps to avoid extra pass over the database. The efficiency of algorithm is tested by applying it on a retail data the algorithm proves to be successful and efficient.

2.1.2. Apriori Algorithm

Two new algorithms (*Apriori* and *AprioriTid*) were proposed by Agerwal et al. [27] in 1994 for generating/discovering association rules named. Experiments show that performance of these algorithms is better than existing association rule mining algorithms The algorithm assumes maximal closure property of item sets which says that all subsets of a frequent item sets must also be frequent. Keeping this property in view algorithm avoid extra passes over the database All those item sets which have infrequent sub sets are ignored and not considered for calculating support.

Best features of these two algorithms were used to introduce new algorithm called *AprioriHybrid*. *AprioriHybrid* outperforms the performance of *Apriori* and *AprioriTid* but implementation of hybrid algorithm is much more complex so performance of Apriori algorithm may be considered a suitable tradeoff in certain situation. Another big contribution of this paper that properties of association rule finding are analyzed

2.2. Privacy preserving Data mining

Data mining has a lot of advantages, but large repositories of data which are used to extract patterns useful for decision making may contain data that is private for an individual or an organization. There is strong need to protect this private data because it becomes an ethical issue if the decision making affect someone's settlement. It gives rise to a new research domain in data mining known as Privacy Preserving Data Mining (PPDM). Large amount of work for attaining privacy in statistical databases, in all different dimensions of data mining has been done. Some of the research work related to technique proposed in this study is as follows:

2.2.1 Security and Privacy Implications of data mining

For first time in 1996 C. Clifton and D. Marks[18] suggested that as data mining is useful for extracting knowledge but it can be a possible security threat in future The

paper does not come up with some algorithmic solution or some strategy but it suggested possible solutions like giving limited access to the data or altering the data in some ways. But paper emphasizes that these measures may cost us with limiting the benefits of data mining.

The threat that data of any organization may have if they are working with data mining was focused by J. Vaida and C. Clifton [12]. The article explains that data mining requires a data warehouse where all the data necessary for statistical analysis is kept. Any leak to data ware house is the leak of information which may be very crucial e.g. in case of medical data one may not wishes his personal data to be released or transactional store data that may used for marketing purpose etc. so all necessary measure must be taken to avoid data theft.

The article also suggested some measures as keeping the data in a centralized data ware house may not be considered favorable when data privacy is needed. Moreover there may be an implication of cryptographic techniques or any other data hiding techniques to save our information from data terrorists.

2.2.2. Detecting Privacy in data mining

While giving solution to the possible threat in data mining J. Vaida and C. Clifton [12] stated that someone may be able to get a better solution if he/she knows which type of data will consider sensitive? So the parameters according to which we are going to differentiate between sensitive and non sensitive information must be known.

Useful work was done by P. Fule and J. F. Roddick [5] to provide the answer to our question. The paper suggests that there may not be a general way using which sensitive knowledge can be selected in all fields and in all regions. When talking about regions different regions may have different social and ethical values and we need to choose sensitive information according to those, similarly different types of data has different requirement for example while publicizing market basket data

customer may not have the issues but organization will definitely have while in case of medical data patient may never want his data to be publicized.

The proposed techniques consider sensitive knowledge according to the measure of interestingness of the rule. The rule which is more interesting is considered as more sensitive.

2.3. Privacy Preserving Data Mining Algorithms

Some of the algorithms which are used to hide statistical data from the data base are discussed below:

2.3.1. State-of-the-art Technique

The state-of-the-art techniques in Privacy preserving Data Mining, were recapitulate by V.S. Verykios et al [2] [13] and also propose classification hierarchy that can be used as the basis for analyzing the work done so for in PPDM. Comprehensive review of the work done in this area is also given along with its coordinates in hierarchy.

PPDM algorithms have been divided into number of classes in[2] depending upon the type of database on which algorithm is going to work, whether distributed or centralized; the working strategies of algorithm, whether heuristic based, cryptographic based or reconstruction based. In[13] a slightly different analysis has been done in which PPDM algorithms have been classified as heuristics, exact, border based approaches.

Different performance measure to evaluate PPDM algorithms were presented in the papers [2] and [13] Some of which are *overall complexity of algorithm, the data utility* during the modification of database in hiding process, the *level of uncertainty* that sensitive information can be inferred again, hiding failure, misses cost, Dissimilarity and recovery factor .

According to the survey none of the techniques presented so far has outperforms all the existing technique in all evaluation criteria. Researchers still are unable to propose a general solution to the problem.

Privacy preserving data mining techniques were surveyed and classified by Evfimievski and T. Grandison [17]. The survey includes only those techniques which were also used for classical data/database security techniques where no data mining is involved. The techniques viewed under the classes Suppression, Randomization, Cryptography and summarization. Advantages and disadvantages of each technique are given in the survey.

The paper also includes different application scenarios of data mining which can be more affectively used if data security is ensured. Also paper sketches some future trend in Privacy Preservation Data mining.

2.3.2. Limiting Disclosure of Sensitive Rule

The problem of limiting disclosure of sensitive rule was presented by M. Atallah et al. [4] .The paper presented that with progress in data mining algorithms security risks for the data also increases. The more data is subjected to statistical analysis the greater is the risk. In the paper selective sensitive frequent item sets were subjected for hiding with minimum impact on rest of the data The approach used in the paper is heuristic and it also prove that optimum solution (the solution in which sensitive data is hid and there is no affect on insensitive data) for this problem is NP hard problem.

The algorithm presented in this paper runs iteratively number of iteration depends upon number of sensitive item sets. It is also proposed that for hiding the sensitive information the first need is to explore different selection criteria for sensitive rules selection and then apply these on data while keeping in view the time and memory constraint

Heuristic algorithm is evaluated using *Cyclic Algorithm* also presented in the paper for the first time.

2.3.3. Association Rule Hiding using Support and Confidence

Five different algorithms (1 a, 1.b, 2.a, 2.b, 2 c) were presented by *Verykios* et al[19] In which they hide association rules by decreasing support or confidence of the rule below minimum support threshold or minimum confidence threshold respectively Sensitivity level is also assigned to each sensitive rule depending upon the impact of the rule; the impact of the rule is the degree by which a certain rule can affect other item sets. The techniques presented in this paper depends upon the hypothesis that sensitivity level of only frequent and strong rules can be considered as interesting

The hiding strategies that are proposed in the paper depends upon discovering those item set that partly or fully support the sensitive rule, so that support of the supporting item set or confidence of the sensitive rule is decreased. Out of the five algorithms presented first three are rule oriented and last two are item set oriented techniques.The algorithms presented tries to minimize the impact of hiding procedure at every step.

Algorithms are evaluated keeping two constraints in mind i.e., time and side effect on sanitized database The limitations of algorithms are that new association rules that are generated if confidence based hiding strategy is used and non-sensitive rules may be lost while using the other algorithms i.e. hiding by decreasing the support. So it is determined that none of the method gives the optimum solution however the choice of the suitable algorithm can be made according to relevant scenario.

2.3.4. Distortion based Frequent Item Set Hiding Algorithm

Presented distortion based technique was presented by *Pontikakis* et al [20]. In this technique security of the data is ensured by first converting the data set to binary form and then blocking some of the values. Blocking of the value is achieved by simply conversion of 1's to 0's. As database is presented in binary form so

converting a value from 1 to 0 is the deletion of item from data set so decrease in the support of selected item set.

The algorithm is efficient enough without any new rules generated but large numbers of insensitive information lost when using this technique. The technique used may leave the sanitized database without any privacy breach; the information which is hid can never be retrieved again. This type of technique could be very efficient for hiding information from the data which is not very critical like in case of medical data it would not be recommended to use the technique as deleting the value may become dangerous.

2.3.5. Introducing Unknown to Hide Association Rule

The idea of introducing unknown values in the data was given by Saygin et al[21] to hide a particular rule or item set The algorithm works on data by first converting it to binary form and then converting appropriate 1's or 0's to some unknown value like'?'. In this way support or confidence of the sensitive items is decreased below the minimum threshold values and the data is no more available for statistical analysis.

The paper discussed in detail the possibility that sanitized database can be retransformed into the original database by converting unknown value which is '?' in this case by 0's or 1's, but '?' is either replaced by 1 or 0 as one does not know the exact combinations with which data was transformed. Now whether sensitive data could be mined again or not depends upon confidence of our sensitive rules.

The main side affect while using the technique is that in sanitized database confidence and support calculated are marginal due to the presence of some unknown values in item sets which can be quite hectic. Moreover some rules may be lost.

The paper evaluated the hiding strategies in terms of CPU usage and side effects produced by each technique on varying degree of confidence value. Side effects

produces is considered as actual performance criteria in the case It is obvious from the results that higher the percentage of confidence lesser is the side effect and vice versa.

2.3.6. Hiding Sensitive Rules with Limited Side Effects

The framework for hiding association rules was presented by Y. H. Wu et al. [14] . The paper works on the idea that cost of hiding sensitive rules should minimize to the best possible level. Here the rules that are lost during sanitization and new rules generated are considered as the cost.

The Algorithm gave five schemes for hiding association rules, all these schemes are based upon distortion and blocking based techniques. The system avoids the unwanted affect on database by creating separate template and action table. Template table contains the sensitive data on which sanitization process is applied after assuring that modification may not have worse affect. This modified data is added to action table from where original database is modified.

The results show that method is scalable in terms of database size In tests conducted the algorithms shows no or minimum side effects. It is discovered that overlapping sensitive rules have adverse affect on performance of algorithms.

2.3.7. Reconstruction based Association Rule Hiding Technique

Reconstruction based framework was proposed by Y. Guo[8] in which FP tree based method is used for inverse frequent set mining as was proposed by Y.Guo et.al. [8]. The algorithm in [8]directly removes the sensitive information whether in the form of frequent item sets or rules, which may be selected on the basis of nature of data or scenario. FP tree is constructed on the basis of insensitive data left and from that FP tree sanitized database is reconstructed The paper also describes different categories of existing data hiding algorithms in data mining. The idea presented here is not still tested on real dataset for hiding information; an expected evaluation plan is presented in the paper.

The algorithm is efficient in terms of zero hiding failure but when database is reconstructed it does not include items which were not frequent in original database and thus it is needed to guess the presence of those items in different transactions and it may take memory and time overhead. The algorithm is laborious and expensive.

2.3.8. Hiding Frequent Item Sets with Limited CPU Usage

Techniques that emphasis not only on hiding the sensitive knowledge but also on accuracy of algorithms and time complexity of algorithms was given by S.R. Oliveira and O.R Zaiane in [29][30]. Algorithm is proposed known as Item Grouping Algorithm (IGA)[30], that focus on hiding sensitive knowledge in only two database scans irrespective of the size of database or number of sensitive item sets that are needed to hide. one scan is required to build index and the second scan is done to process the sanitization process. Another similar algorithm proposed a new technique for hiding interesting knowledge. The focus of the algorithm is not only accuracy but also limited database scan. Algorithm is known as Sliding window Algorithm (SWA)[29]. It works by copying all the non sensitive item sets from the database D to sanitized database D' and sensitive elements are subjected to sanitization process and then added to D'.

Algorithm is compared with current techniques and proved to be more effective in terms of less CPU usage and fewer misses cost.

2.3.9. Hiding Strategies while maintaining Data Quality

The idea of maintaining data quality of data while applying PPDM algorithms was introduced by E. Bertino and I.N. Fovino [9]. Data quality is formally defined and also an evaluation model for data quality is proposed, known as information quality model. The model explains how one can find the impact on quality of data and thus it helps to select the best PPDM algorithm according to our data. The model is applied on PPDM algorithms that work on association rule hiding.

A new methodology for performing PPDM operations was proposed by E. Bertino and I.N. Fovino [11], which also preserve the data quality. The algorithm is known as DQDB. The DQDB algorithm proposed in this paper is distortion based, and data quality is maintained by first calculating affect of each alteration on database and then limiting the use of those alterations which has maximum affect on data quality

2.3.10. Maintaining Privacy and Data Quality in PPDM

A novel technique to hide sensitive association rules was given by Modi et al [3], the algorithm along with preserving the quality of database. The algorithm presented is known as Decrease Support of R.H.S. item of Rule Clusters (DSRRC). The algorithm works by clustering the sensitive rules which have similar consequent, the right hand side item of an association rule. Now the consequent from the cluster having highest sensitivity is deleted from transaction supporting that particular item. Sensitivity of each transaction is calculated on the bases of degree of presence of items from sensitive rules in each transaction.

The performance is compared with existing technique that also works on similar parameter. The performance is measured in terms of internal and external parameters. The main emphasis of the algorithm is to preserve quality of the database while hiding the sensitive information. The algorithm has limitations as it works only for association rules having single element in its consequent, moreover as it works with deletion of consequent from the database it may leads to more number of rules lost.

2.3.11. Levelwise Search and Border of Theories in Knowledge Discovery

Concept of Border Theories by H. Mannila and H Toivonen [24]open a new gateway towards PPDM algorithms. The paper introduced a level wise search for finding the frequent knowledge, while Knowledge discovery process the number of accesses to the database are bound by introducing positive and negative borders The maximal frequent item sets (the item set which does not have a frequent item

super set) are said to be positive border and minimal non-frequent (non frequent item sets but their sub sets are frequent) as negative border.

2.3.12.   Border Theories to Hide Sensitive Knowledge

Border theories were further used to hide sensitive knowledge by Sun et al. [22]. Main purpose of the algorithm presented in the paper is to not only to hide the sensitive knowledge but also preserve the quality of non-sensitive information. Considering the apriori property that subset of a frequent item set are also frequent, sensitive item sets are hid in a way that it preserve the quality. During the hiding process at every step greedy approach is used and the affect of hiding element on non sensitive data is calculated so most appropriate decision is made. Borders are used as proper representation of non sensitive frequent item set and are used to make modification in the data base with as little affect as possible.

Algorithm is efficient as compared to contemporary techniques in terms of less hiding failure and minimal affect non sensitive data but a little more time consuming as compared to others.

2.3.13.   MAX-MIN Approach for Hiding Frequent Item Sets

A new approach for sanitizing the data is proposed by Moustakides et al. [23]. The technique introduced uses the concept of border revision theory to discover positive and negative border elements from frequent item set lattice and then applies decision theory for maximizing the minimum gain while hiding sensitive knowledge.

Borders in this paper are found using the technique introduced by Sun et al.[22]. Positive border elements in the frequent item set are discovered by deleting all the sensitive elements and their supersets from the lattice. Now frequent item sets which do not have their superset in the lattice are said to be in positive border while item sets in original lattice which are not present in revised lattice but all their sub sets are present are said to belong to the negative border elements. Here is the main idea

of the technique that positive borders are all those elements whose supersets are either sensitive elements or are super set of sensitive elements so any change for hiding sensitive elements will affect the elements in positive border.

To reduce the number of iterations for hiding process intersection between negative border set and sensitive elements set is subjected to hiding process. Now the sensitive item set whose super sets are also sensitive are removed so again maximal closure property of apriori lattice is used here; as if subset of an element is hidden, its superset would never be extracted.

So technique says positive border elements are non sensitive elements which are more vulnerable to change during sanitization process. A decision theory approach is defined to find a MAXMIN element. Two algorithms named MAXMIN-1 and MAXMIN-II are introduced In both algorithms the idea that positive border elements in the revised frequent item set lattice are more subjected to change when sanitizing the data base is used and MAXMIN element is found by considering the number of positive border elements that form the negative border item sets and then keeping a check on support of negative border element. MAXMIN-II works by finding the affect of alteration that is going to be applied on the database.

The two algorithms are evaluated and compared with other border based PPDM technique as given by Sun et al. [22] Parameters used are number of rules subjected for hiding, number of lost rules and number of new rules generated. Algorithm MAXMIN-II shows best results as compared to the other two in most of the cases and computationally much less demanding.

2.3.14. PPDM using Alternative Interest measure

The idea extracting useful knowledge using statistical approach other than confidence and support was proposed by E. R. Omiecinski et al.,[15]. The idea was used by M. Naeem et al . [16] to hide sensitive association rules using central tendency another standard statistical measure. The paper claimed that other statistical methods can be equally effective as confidence and support. Weighing

technique is used to weigh the sensitive rules to decide which sensitive rule may be used first to hide. Weighing methodology is based on Central tendency mean, median, mode and sum

The technique proved to be effective as compared to contemporary techniques in terms of no ghost rules (new rules generated during hiding) and hiding failure. While side affect of generation of lost rules still exist in the case.

2.4. Problem Statement

In the literature survey existing techniques were discussed for extracting useful knowledge, sanitization of databases and hiding sensitive data Problems related to every technique is mentioned which is needed to be solved in future. In our work we focused on the techniques applied to hide association rules directly or indirectly.

Optimal sanitization of database for the purpose of hiding sensitive data is an NP-hard problem [4]. Many approaches have been proposed for the purpose in all different dimensions of data mining but all those were unable to avoid the side effects of sanitization completely. The major side effects which are addressed in existing techniques are:

- New rules generation

- Lost of non sensitive Rules

- Hiding Failure

The techniques proposed so far are working either by hiding sensitive rules or hiding sensitive item sets. Following issues are identified in the existing problems which are needed to be resolved:

- The actual knowledge elements in association rule mining on the basis of which statistical analysis is performed are association rules and frequent item sets. So Association rules are actually needed to hide.

- If we are hiding sensitive item sets then we are in fact hiding association rules indirectly. For example if we hide item set *ab* then we are hiding rules *a→b and b→a*. Now there is a possibility that only *a→b* is sensitive so doing this we are of course hiding extra knowledge as well. So all the association rules associated with an item set may not be sensitive, so along with sensitive rules, insensitive associations rules get hidden which is not desirable in case of data mining, because it leads to wrong patterns derivation from data set and hence wrong decision.

So in this work sensitive rules are selected to hide directly. The techniques proposed for hiding sensitive rules directly have following issues:

- Little work has been done in rule hiding with minimum concern of data quality

- Data quality is considered by [3] while hiding rules but has limitation that it can only hide rule with one element in its consequents, which is not practical approach in real world data mining.

2.5. Summary

In this chapter techniques concerned with data sanitization for the purpose of hiding sensitive knowledge are discussed. A thorough literature survey is done regarding the need of the privacy in data mining and concerned sanitization algorithms. The techniques studied so far may be classified as Heuristic approaches as discussed in [3],[16], Border theories based approaches [23], [22], Reconstruction based techniques [8], Support and Confidence based algorithms to hide sensitive knowledge [19], and Blocking and distortion techniques [20], [21] The problem related to techniques is discussed and the area concerned to work done here is narrowed down at the end of chapter.

Chapter 3

Proposed Solution

3. Proposed Solution

The objective of this chapter is to propose architecture for hiding sensitive association rules. The proposed technique can effectively used to reduce the limitations present in the previous techniques

All the discussion in literature review emphasis on the need of technique for hiding association rules successfully without affecting data quality. The technique must be scalable in terms of number of elements/items present in sensitive rule's antecedent and consequent. The algorithm proposed here works on direct hiding of association rules by decreasing rule's confidence below user defined confidence. Data quality is maintained by adopting border theory and weighing the sensitive rule's element.

3.1. Proposed System Framework

The proposed Architecture for our approach/algorithm is portrayed in Figure 4.1 that consist of three phases given below

- Preprocessing

- Finding the MAX-MIN element

- Hiding sensitive elements and database updation

All these phases are further divided into number of step as shown in Figure 4.1 and explained with the help of an example.
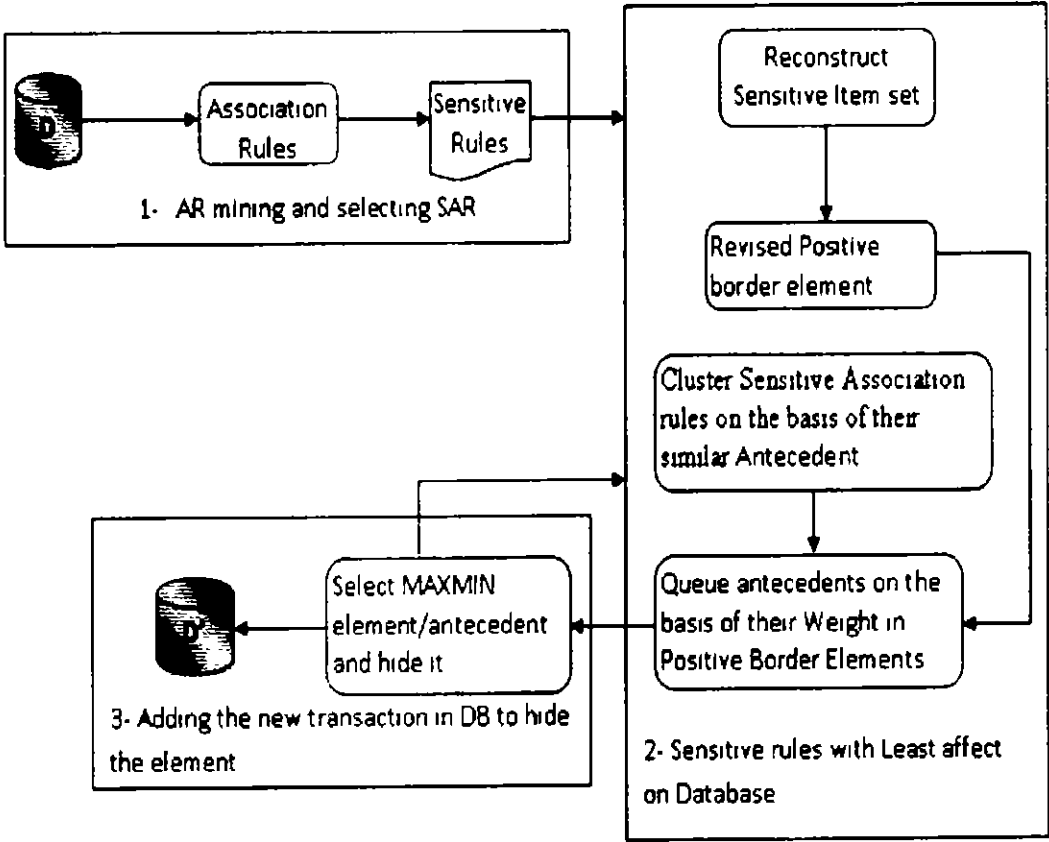
Figure 4.1: Proposed Framework

3.1.1. Preprocessing

The database is loaded and is converted into binary forms before any processing because in binary form computations can be performed more easily. For example **Given** item set $I= \{a,b,c,d,e\}$ and database D consist of transactions $\{T1, T2, T3, T4, T5, T6, T7\}$, as shown in table 4.1.

Table 4.1: Database D

TID	Items
T1	abcd
T2	bcd
T3	cd
T4	bc
T5	abe

TID	Items
T6	abcd
T7	acde

Database is converted to binary form so that database took a form of matrix. Binary value represents the presence or absence of particular item in database.

Table 4.2: Database D in Binary Form

TID	a	b	c	d	e
T1	1	1	1	1	0
T2	0	1	1	1	0
T3	0	0	1	1	0
T4	0	1	1	0	0
T5	1	1	0	0	1
T6	1	1	1	1	0
T7	1	0	1	1	1

3.1.2. Association rule Mining

The first phase is to compute interesting rules from database and then to identify the sensitive association rules. Association rules are mined using apriori algorithm as was proposed in [27].

The algorithm states. *Given D be a set of transactions T in the data base let T be a transaction consisting of number of items i, so we can say that T = {i^1, i^2, i^3 ... i^n} A set of items (X subset T) is known as an item set If X is present above certain threshold support (minsup), then X is a frequent item set The apriori ought to find out all association rules of the form X$\rightarrow$Y (where Y subset T), holds if rule has confidence above threshold confidence (minconf)*

The algorithm works in two phases:

• **Candidate Generation**

The process of finding possible item sets that can be checked against *minsup* to undergo a pruning process is known as candidate generation. A frequent item set lattice known as *apriori lattice* is generated during the process. The process works in bottom up fashion, generating one item; frequent item set in the first step and then joining them to generate candidates with larger number of items; these candidates are checked against *minsup*, in this way all non frequent item sets are cut back The process continues unless no more candidates are discovered. Item set lattice generated in our system is shown in the Figure 4.2. Frequent itemsets with *minsup ≥ 3* are highlighted in the lattice.
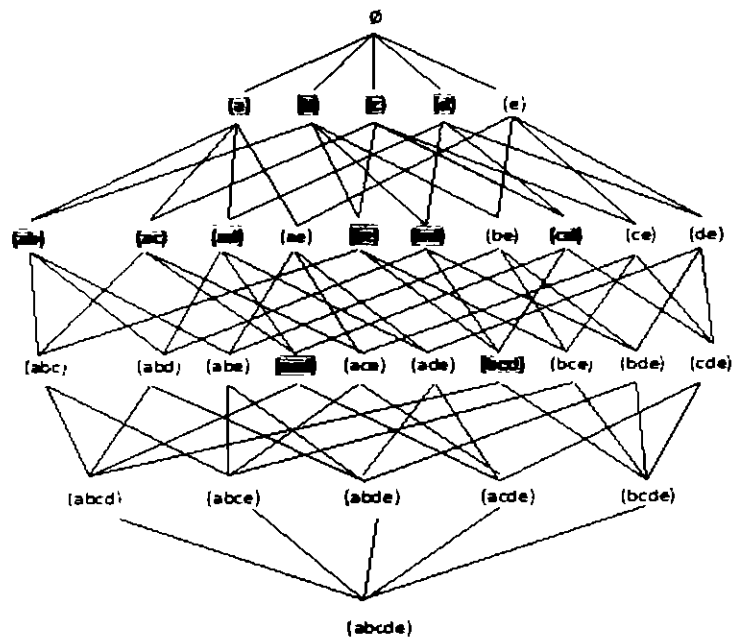
Figure 4.2: Frequent item set in Apriori Lattice

• **Rules Generation**

Large item frequent sets are taken one by one and all possible association rules of the form X→Y are generated and then are checked for *minconf*, the rules with confidence less than *minconf* are pruned off. *mincof* in this case is *0 5* All the rules with confidence ≥ minconf are shown in table 4 3. Confidence is calculated

using formula given in equation 2.2. Number of items in consequents and antecedent may vary depending upon frequent item set.

Table 4.3: Association rules with *minconf ≥0.50*

Association Rules	Confidence
a→b	0.75
b→a	0.60
a→c	0.75
c→a	0.50
a→d	0.75
d→a	0.60
b→c	0.80
c→b	0.6
b→d	0.6
d→b	0.6
c→d	0.83
d→c	1
a→cd	0.75
c→ad	0.50
d→ac	0.60
cd→a	0.6
ad→c	1
ac→d	1
b→cd	0.60
c→bd	0.5
d→bc	0.6
cd→b	0.6
bd→c	1
bc→d	0.75

3.1.3. Sensitive rule Selection

In the second part sensitive rules are identified. Criteria for identifying sensitive rules depend upon user's scenario in statistical database and also on user's social environment and ethical issues as discussed in [5]

Sensitive rules selected to hide in this example are {b→a, d→a, d→bc}.

3.1.4. Finding the MIN-MAX element

The element which has minimum affect on data while maximizing the possible gain is called MAX-MIN element according to decision theory In our case min-max element is discovered by weighig sesitive elements against positive border elemets and clustering rules according to their consequents.

- ## Reconstruct the sensitive item sets

In this step each sensitive rules is processed to form the frequent item set from which that particulare rule was originated Reconstruction involves the *antecedent* ∪ *consequent* of that rule. Reconstructed sensitive item sets in our example are {ab, ad, bcd}.

- ## Finding the revised positive border element

Border revision theory used for hiding sensitive knowledge in [22][23] The technique starts from the minimum level of apriori frequent item set lattice shown in the Figure 4.3 (a) given below and moves upward removing all the sensitive itemsets reconstructed in last step and also their super item sets the sensitive item sets and their supersets deleted are encircled in Figure 4.3 (b) The process continues until it reaches at upper limit of the lattice. Now in this modified lattice all those frequent item sets which do not have a super set are belong to revised positive border. Positive border elements are shown as underlind elemnets in figure 4.3 (b).

These revised positive border elements are the most vulnerable non sensitive frequent item set
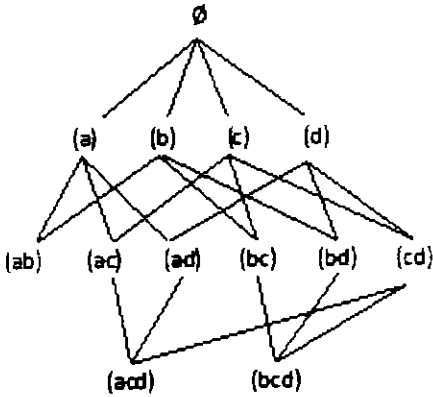
Figure 4.3 (a): Frequent item set lattice
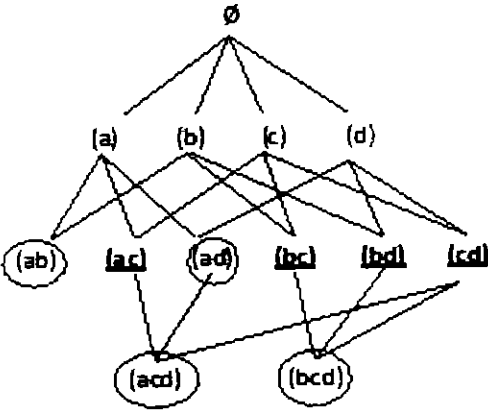
Figure 4.3 (b): Frequent item set lattice with Positive border

- **Cluster the sensitive association on the basis of similar antecedent**

Sensitive rules are clustered on the basis of similar L.H.S. element the antecedent, in contrast to the technique [3]. In [3] they are hiding rules by decreasing the support of R H.S. element while proposed technique works with increasing the support of antecedent as was also done in [19].

But here in this case each and every sensitive rule is not considered for hiding or taking one rule apply sanitization and then scan the database to check for all those rules which were hidden during the process. Instead the technique work by making clusters of sensitive rules; where each cluster represent all the sensitive rules with similar antecedent.

The idea behind this technique is that all those rules with similar antecedent having confidence *count* $\leq c$ will be hidden if we are able to hide one rule having confidence *count* $= c$ so instead of processing each rule from the clusters the technique select one rule with largest confidence. Number of Clusters generated in our example are two as given below:

C1 includes *{ b→ a'$_{0\ 60}$}* and C2 includes *({d→a}/$_{0\ 6}$,{d→bc}/$_{0\ 60}$)*

Antecedends of the sesitive rule from each with highest confidence is selected.

It can be formulated as:

qn = {antec|antec ∈ Cn AND minconf(Rantec) = maxconf(Cn)} .... (4.1)

Where Cn is the cluster n corresponds to the number of cluster here n={1,2}, minconf(R$_{antec}$) is confidence of the rule from which antecedent *antec* belongs to and maxconf(Cn) is the greatest confidence of the rule that belongs to the cluster Cn

In this example $Q = \{q1,q2\}$ where *q1 = b, q2 = d*

- **Weighing the antecedent**

In our very last step the element with maximum confidence of its rule is selected, all these antencedent are queued according to increasing order of their weight in positive border elements

As database is subjected to sanitization process on the basis of these antecedent elements, here in this step a measure has been taken to minimize the affect of sanitization.

In the example elements from Q are fetched one by one and are weighed against revised positive border element set and is queued in decreasing oreder of its weight.

Table 4.4: Antecedent Weight in Positive border

Antecedent	Weight in the Positive Border
b	2
d	2

3.1.5. Hiding the element and data base updation

In this step MAX-MIN elements are subject to hidig process and database D˙ is updated.

- **Hiding strategy**

  From the queue created in last step first element is selectes, it is MAXMIN element with maximum confidence value in its cluster and minimum weight in positive border elements.

  Now this MAXMIN is subjected to hiding process.

  Hiding strategy in our algorithm is adding new transaction in the database. MAXMIN element is selected from the queue one by one and corressponding elements are added in the database, and a transition table is maintaind. Transition table created in our example is shown below:

Table 4.5: Transition table

bd
bd
b

- **Database Updation**

  From the database we do not mean the original database here in the last step a transition table is maitained in which all the possible possible values may need to introduce in original database are added to avoid extra scans over the database.

  The values in the in the transistion table are append in the database. Here is our new updated data base D˙.

Table 4.6: Database D′

TID	Items
T1	abcd
T2	bcd
T3	cd
T4	bc
T5	abe
T6	abcd
T7	acde
T8	bd
T9	bd
T10	b

Now if database D′ is subjected for the rule mining process the resultant association rules are given in table 4.7; all the sensitive rules are hidden. Fortunately, no new rules are generated during the process, but some non sensitive rules are also lost.

Table 4.7: Association Rules in Database D′

Association Rules	Confidence
a→b	0.75
a→c	0.75
c→a	0.50
a→d	0.75
b→c	0.50
c→b	0.60
b→d	0.62
d→b	0.71
c→d	0.83
d→c	0.71
a→cd	0.75
c→ad	0.50
cd→a	0.60
ad→c	1
ac→d	1
c→bd	0.50
cd→b	0.60
bd→c	0.60

Association Rules	Confidence
bc→d	0.75

3.2. Proposed Algorithm

In this section algorithm used for hiding purpose is discussed. Time Complexity of the algorithm proposed here depends upon the number of association rules selected as sensitive and differnce between the confidence of the sensitive rule and threshold confidence below which rules will be hidden As the difference increases time complexity also increases. The algorithm used to hide the rule is given in Figure 4.4.

```
Function: hide_rule()
    Input: Database D, Threshold Confidence conf, Sensitive Rules S, Reconstructed Frequent item sets FISr,
           Revised Positive Border rpb
    Output: Database D'
    cluster   = cluster_senRule (S)
    q_antec   = weigh(cluster, rpb)
    while q_antec != Φ
        maxmin = {m | m ∈ q_antec}
            while temp_conf>conf
                if maxmin ∉ FISr then
                    if trans_table == Φ then
                        count = count+1
                        trans_table(count) = maxmin
                        supp_cons = S consequent.support+1
                    else
                        for (j=1 to length(trans_table)
                            if maxmin ⊆ trans_table[j] OR trans_table[j] ⊆ maxmin then
                                temp = union (trans_table[j],maxmin)
                                    if temp ∉ FISr then
                                        count = count+1
                                        trans_table(count) = maxmin
                                        supp_cons = S consequent support+1
                                    else
                                        count = count+1
                                        trans_table[count] = maxmin
                                        supp_cons = S consequent.support+1
                else
                    for (i = 1 to size(maxmin))
                        count = count+1
                        trans_table[count] = maxmin item.at(i)
                        supp_cons = S consequent support+1
                temp_conf = S antecedent support/ supp_cons
        q_antec = q_antec - maxmin
```

Figure 4.4: Proposed Algorithm

3.3. **Summary**

In this chapter a new technique for hiding association rules is presented. Different techniques used in the algorithm are discussed in detail and functionality of algorithm is also explained with the help of an example

The presented algorithm is expected to have limited side effect of sanitization process over the database. It is anticipated that no new rules will be generated during the process and hiding failure will also be zero but the problem of lost rules is still there, in this algorithm number of rules lost is reduced to minimum level. Overall performance of the algorithm is assumed to be better than other state-of-the-art techniques.

Chapter 4

Experimentation

4. Experimentation

The main focus of this chapter will be the validation of methods for hiding association rules on the selected set of data. This focus has been achieved by using dataset of [28]described in section 5.1. The section 5.2 will include the details of pre-processing on the selected data set Performance measures adapted during the course of this project and its comparison with other techniques and algorithms will be the spotlight of section 5.3. In the last section of this chapter, the evaluation and validation of hiding association rules through the chosen algorithm will be integrated.

4.1. Dataset

Our method for hiding association rule is validated using mushroom dataset of UCI machine learning repository. This data set includes descriptions of theoretical samples parallel to 23 species of gilled mushrooms in the Agaricus and Lepiota Family. Each species is recognized as definitely poisonous, definitely edible, or of unknown edibility and thus not recommended. This latter class was pooled with the poisonous one. The Guide to this dataset clearly states that there is no simple rule for determining edibility of a mushroom.

The dataset is multivariate with 8124 number of instances. The total number of attributes present in this dataset is 22 with missing values for attribute 11. All 22 attributes are nominally valued.

4.2. Pre-processing on dataset

The process started with basic transformations on the selected dataset for ease of processing. The pre-processing applied on the selected dataset involved mapping the values of each attribute to numbers from 1 to 126. These values are then mapped again to their binary counterparts to improve performance and efficiency. This resulted in leaving us with a binary dataset which was analyzed further to hide association rules. Original dataset values and their corresponding mapped values are shown in Table 5.1.

Table 5.1: Mapped Attributes

Attribute Value/ Mapped Value

Attribute												
Cap-shape	Bell=b / 1	Conical=c / 2	Convex=x / 3	Flat=f / 4	Knobbed=k 5	Sunken=s / 6	—	—	—	—	—	—
Cap-surface	Fibrous=f / 7	Grooves=g / 8	Scaly=y / 9	Smooth=s / 10	—	—	—	—	—	—	—	—
Cap-color	Brown=n / 11	Buff=b / 12	Cinnamon =c / 13	Gray =g/ 14	Green=r / 15	Pink=p / 16	Purple=u / 17	Red=e / 18	White= w / 19	Yellow=y / 20	—	—
Bruises?	Bruises=t / 21	No=f / 22	—	—	—	—	—	—	—	—	—	—
Odor	Almond=a / 23	Anise=l / 24	Creosote= c / 25	Fishy=y / 26	Foul=f / 27	Musty=m / 28	None=n / 29	Pungent =p / 30	Spicy=s / 31	—	—	—
Gill-attachment	Attached=a / 32	Descending= d / 33	Free=f / 34	Notched=n / 35	—	—	—	—	—	—	—	—
Gill-spacing	Close=c / 36	Crowded=w / 37	Distant=d / 38	—	—	—	—	—	—	—	—	—
Gill-size	Broad=b / 39	Narrow=n / 40	—	—	—	—	—	—	—	—	—	—
Gill-color	Black=k / 41	Brown=n / 42	Buff=b / 43	Chocolate= h / 44	Gray=g / 45	Green=r / 46	Orange=o / 47	Pink=p / 48	Purple= u / 49	Red=e / 50	White= w / 51	Yellow =y / 52

Attribute Value/ Mapped Value

Attribute									
Stalk-shape	Enlarging=e /53	Tapering=t / 54	—	—	—	—	—	—	—
Stalk-root	Bulbous=b /55	Club=c / 56	Cup=u / 57	Equal=e/ 58	Rhizomorphs =z/59	Rooted=r / 60	Missing=? / 61	—	—
Stalk-surface-above-ring	Fibrous=f/ 62	Scaly=y / 63	Silky=k / 64	Smooth=s / 65	—	—	—	—	—
Stalk-surface-below-ring	Fibrous=f/ 66	Scaly=y / 67	Silky=k / 68	Smooth=s / 69	—	—	—	—	—
Stalk-color-above-ring	Brown=n / 70	Buff=b / 71	Cinnamon =c / 72	Gray=g/ 73	Green=r / 74	Pink=p / 75	Red=e / 76	White=w / 77	Yellow= y / 78
Stalk-color-below-ring	Brown=n / 79	Buff=b / 80	Cinnamon =c / 81	Gray =g / 82	Green=r / 83	Pink=p / 84	Red=e / 85	White=w / 86	Yellow= y / 87
Veil-type	Partial=p / 88	Universal=u / 89	—	—	—	—	—	—	—
Veil-color	Brown=n / 90	Orange=o / 91	White=w / 92	Yellow=y / 93	—	—	—	—	—

Attribute	Attribute Value/ Mapped Value								
Ring-number	None=n / 94	One=o / 95	Two=t / 96	—	—	—	—	—	—
Ring-type	Cobwebby =c / 97	Evanescent= e / 98	Flaring=f / 99	Large=l / 100	None=n /101	Pendant= p / 102	Sheathing =s / 103	Zone=z / 104	—
Spore-print-color	Black=k / 105	Brown=n / 106	Buff=b / 107	Chocolate= h / 108	Green=r / 109	Orange=o / 110	Purple=u / 111	White=w / 112	Yellow= y / 113
Population	Abundant= a / 114	Clustered=c / 115	Numerous =n / 116	Scattered= s / 117	Several=v / 118	Solitary=y / 119	—	—	—
Habitat	Grasses=g / 120	Leaves=l / 121	Meadows =m / 122	Path=p / 123	Urban=u / 124	Waste=w / 125	Woods=d / 126	—	—

4.3. Performance measures adapted

The performance measures selected for the current study are the following

4.3.1. Hiding Failure (HF)[13]

This measure enumerates the percentage of the sensitive patterns that remain exposed in the sanitized dataset. It is stated as part of the restrictive association rules that emerge in the sanitized database divided by those that appeared in the original dataset. Formally,

$$HF = \frac{|RP(D')|}{|RP(D')|} \quad \text{.. (5.1)}$$

where R_P (D') corresponds to the sensitive rules discovered in the sanitized dataset D', R_P (D) to the sensitive rules appearing in the original dataset D and $|X|$ is the size of set X. Ideally, the hiding failure should be 0%.

4.3.2. Misses Cost (MC) [13]

This measure quantifies the percentage of the nonrestrictive patterns that are buried as a side-effect of the process of sanitization. It is calculated as follows.

$$MC = |Rp(D)| - |Rp(D')|/|Rp(D)| \quad \text{.. (5.2)}$$

Where R_P (D') is the set of all non-sensitive rules in the sanitized database D' and R_P (D) is the set of all those non-sensitive rules that are in the original database D There exists a compromise between the MC and the HF, because the more sensitive association rules one needs to hide, the more rightful association rules is expected to miss.

4.3.3 Artifactual Pattern(AF) [13]

AF measures the number of unwanted changes made in the database. It is calculated as:

$$AF = (P' - P \cap P')/P' \quad\text{..}\quad (5.3)$$

Where P identifies association rules is original database D and P' is association rules discovered from database D'

4.3.4. Dissimilarity (Diss) [13]

The measure of dissimilarity counts the difference between the original and the sanitized datasets, where the horizontal axis contains the items in the dataset and the vertical axis corresponds to their frequencies. It is calculated as follows:

$$Diss\ (D, D') = 1/\sum_{i=1}^{n} fD(i) * \sum_{i=1}^{n}[fD(i) - fD'(i)] \quad\text{...............................}\quad (5.4)$$

Where n is the number of distinct items in the original dataset D and *fD(i)and fD'(i)corresponds to* the frequency of the *i*th item in the dataset D or D' respectively.

4.4. Validation of results

The algorithm is tested using mushroom data set and efficiency of algorithm is considered against performance measures given above. All the parameters are tested against five different frequency threshold values to check the scalability factor in our algorithm.

4.4.1. Hiding Failure

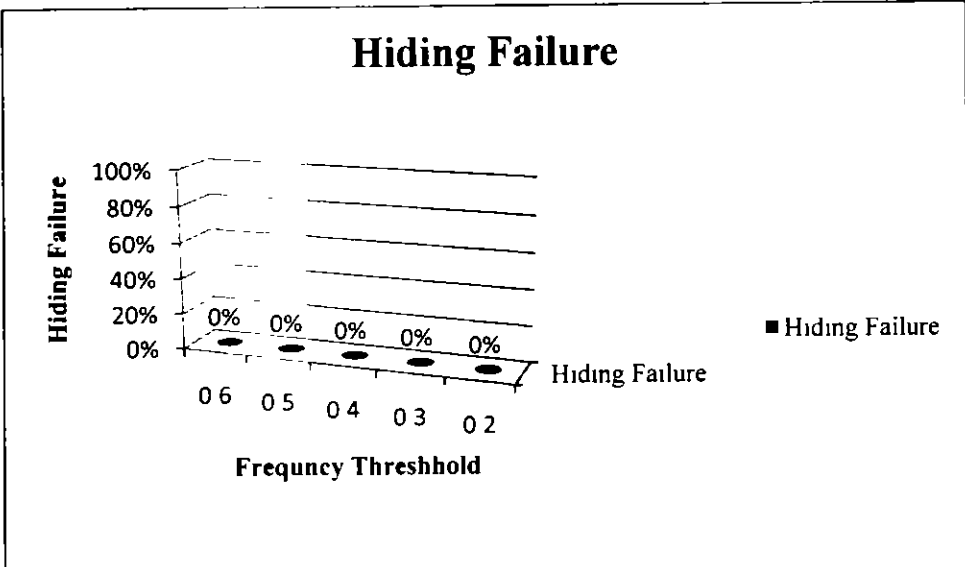Hiding failure against five different frequency threshold values is calculated and is shown in Figure 5.1.

Figure 5.1: Hiding Failure

Hiding failure appears to b 0% in all different frequency threshold values.

4.4.2. Misses Cost (MC)

Misses Cost is calculated again against different frequency threshold values is shown in Figure 5 2.
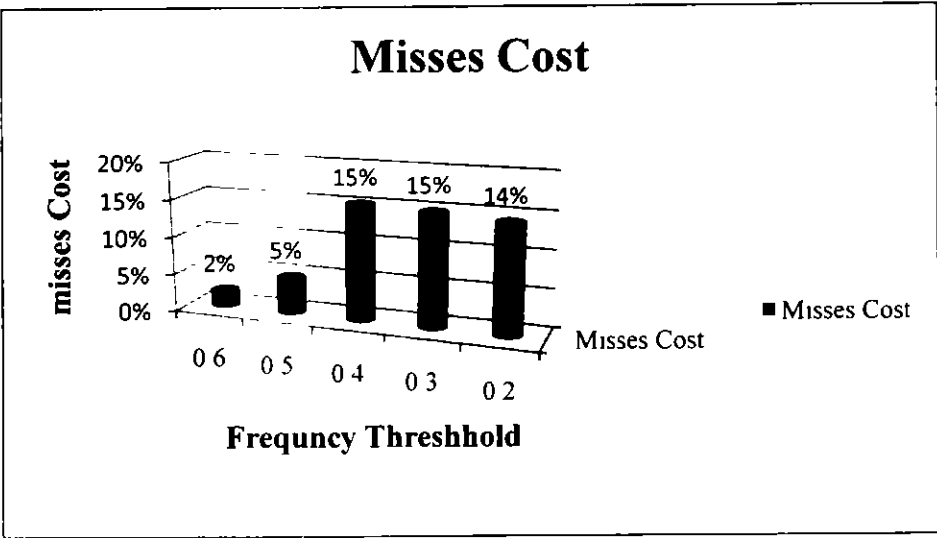

Figure 5.2: Misses Cost

Surprisingly misses cost appears to be decreasing as frequency threshold to hide rules decreases. The reason behind this is our strategy of inserting new transactions in database to decrease the confidence so lower the threshold at which we have to hide the data more transactions are added and these new transactions help the non sensitive rules to hide.

4.4.3. Artifactual Pattern

Artifactual pattern that corresponds to the unwanted rules generated in our data is tested against five different thresholds values the results are shown in Figure 5.3.
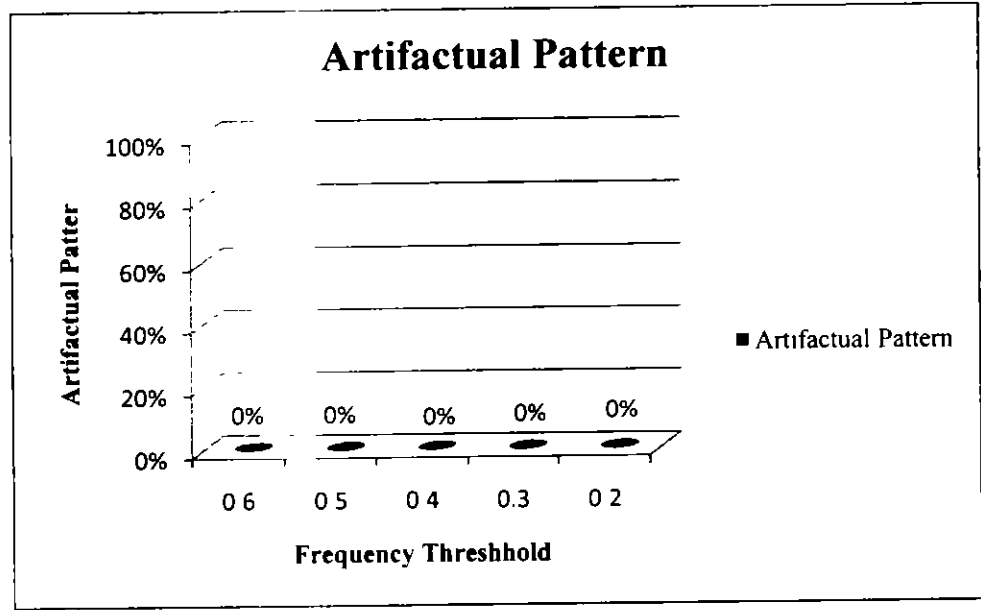

Figure 5.3: Artifactual Pattern

No new rules generated during the execution of algorithm so Artifactual Pattern is 0%.

4.4.4. Time required:

Time required to run algorithm against different frequencies is calculated and is shown in chart given below in Figure 5.4.
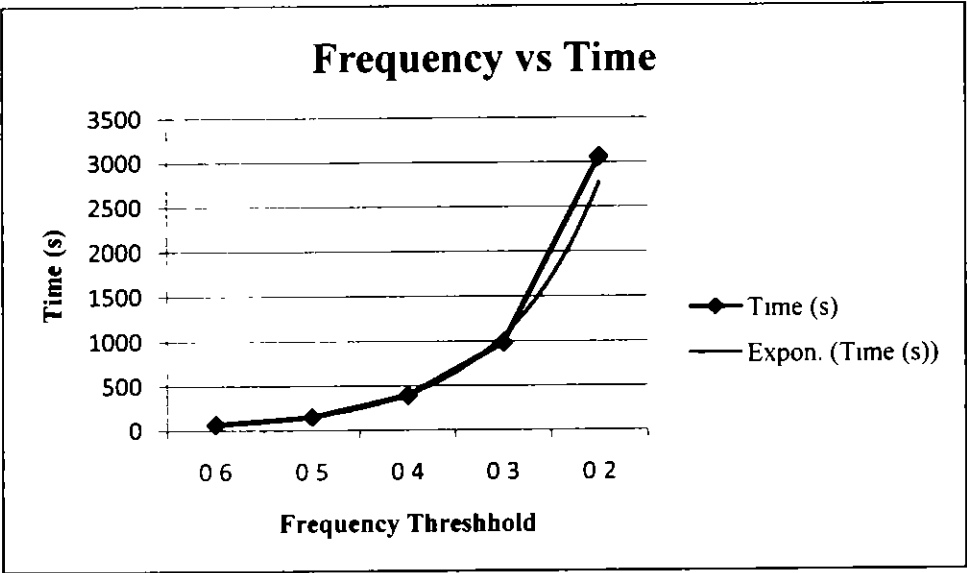
Figure 5.4: Frequencies and Time Graph

Graph shows that running time of the algorithm increases exponentially as frequency to hide rules decreases.

4.4.5. Dissimilarity (Diss)

Dissimilarity is the measure of change of frequency of items in database Dissimilarity of database D and D' is shown in Figure 5.5.
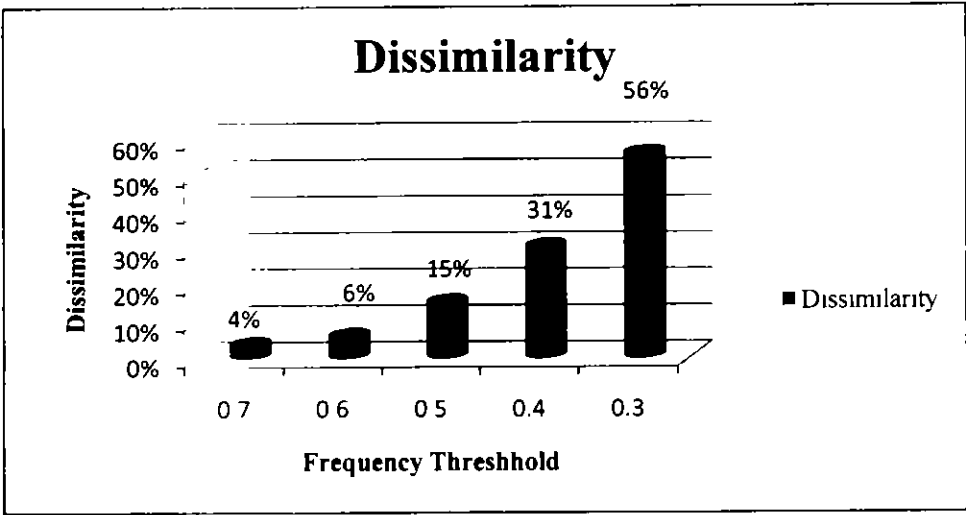

Figure 5.5: Dissimilarity at Different Frequencies

Dissimilarity of both databases is calculated at five different frequency threshold values as shown in graph.

4.5. Comparison with contemporary techniques

In this section performance of algorithm is compared with other contemporary techniques. The algorithm1b from [19] and algorithm DSRRC [3] both these algorithms hides association rules by decreasing the support of its antecedent in contrast to the technique proposed here in which support of consequent is increased to hide a association rule.

The parameters used for comparisons are Hiding Failure, Misses Cost. Artifactual Pattern and Dissimilarity All these parameters are calculated at *0 50 support threshold* and *0 75 confidence threshold*. Results as shown in table 5.2; it shows that performance of proposed algorithm is better than the existing one

Table 5.2: Performance Comparisons

Algorithms	Parameters			
	Hiding Failure	Misses Cost	Artifactual pattern	Dissimilarity
DSRRC	0%	36%	0%	6%
Algo1b	33%	45%	68%	10%
MAX-MIN	0%	3%	0%	3%

All these results are mapped in Figure 5.6.
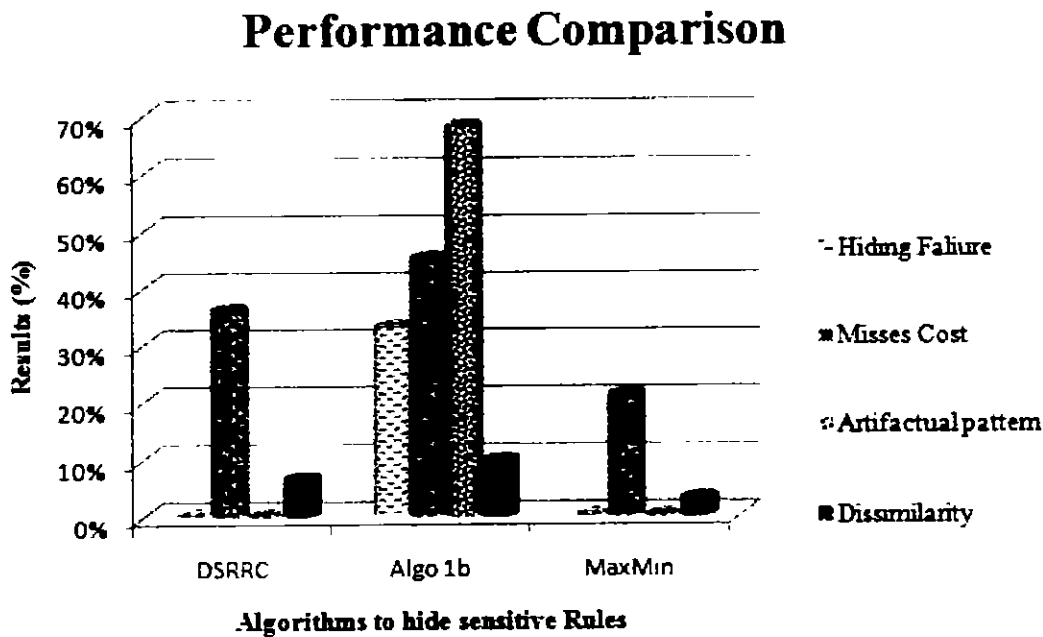
Performance Comparison

Figure 5.6: Comparative Performance Evaluations

4.6. Summary

In this chapter dataset used to evaluate our algorithm is discussed. Some preprocessing on dataset is done; attributes are mapped against numeric numbers Parameters to check the performance of algorithm are also discussed in chapter. Different experiments are performed on dataset to check working of our algorithm. Moreover algorithm is compared with contemporary techniques and results are presented in the form of charts.

Chapter 5

Conclusion & Future Work

5. Conclusion and Future Work

The work presented in this paper is direct hiding of association rule using MAX-MIN criteria from decision theory, as already discussed it is the extension of an existing work [23] (Moustakides et al. 2006), the technique developed here not only emphasis on hiding the association rule but also tries to enhance the quality of data in sanitized database.

While working with association rule mining actual breach to the privacy is association rules, because these are the association rules not frequent item sets on the basis of which decisions are made. Knowledge can be hidden, both by hiding sensitive rules and frequent item sets, but in frequent item sets it's similar to hiding knowledge at higher level of granularity. The existing MAX-MIN approach for hiding sensitive knowledge work by hiding frequent item sets so it hide all association rules associated with those item sets but the possibility is there that all rules associated with an item set may not be sensitive.

For decreasing the confidence of the rule; the support of antecedent is increased in proposed algorithm, as it is already used by *Verykios* et al (2004) [19]. But Unlike their proposed technique new transaction to database are added that contains antecedents/union of antecedents of sensitive association rules. In this way possible generation of large number of ghost rules is avoided.

Hiding strategy in the work presented is based upon increasing support of antecedents so all non sensitive rules with same left hand side elements are more vulnerable to changes. The basic aim here is to find more vulnerable element and minimize affect on them by weighing the antecedents and calculating their impact on revised positive border elements as was used by Sun et al (2005) [22].

Performance of algorithm is tested on mushroom dataset; the same dataset is used in base paper. Efficiency is compared against both types of techniques; the technique that uses border theories and technique that hide sensitive rules by lowering confidence. Experimental results prove that the technique proposed here in this study is better than the existing one. Comparison of side effects; hiding failure, new rule Generation and lost rules are done.

The algorithm DSRRC [3] worked on similar pattern i.e., sensitive rule is weighed for some predefined parameters and then subjected to hiding process but it is not scalable it can only work for sensitive rules with single item in its antecedent while proposed algorithm is scalable as it can work with multiple items; the proposed algorithm is tested for 10 elements in its consequent or antecedent and it gives results on the same pattern.

However, Efficiency of the algorithm can be affected by the confidence of the sensitive rule and number of sensitive rules subjected for hiding practice. Moreover if all sensitive rules chosen have different antecedents number of clusters will increase and it may increase the time needed for sanitization process. While working on the algorithm time or size of the database if it increases as result of sanitization process is not considered. The time range for execution of algorithm varies with frequency threshold same constraint works for raw data introduced in database.

In future Max-Min criteria can be used on other knowledge discovery methods like classification or clustering. Border theories can be applied for knowledge hiding when using other statistical measures. Moreover, the need of the hour is to find a general solution for knowledge hiding problem that gives optimum results for hiding knowledge. For this purpose other suitable statistical measures can be investigated for knowledge discovery.

Chapter 6

References

6. References

[1] M. Kantarcioglu, J. Jin, C. Clifton, When do Data Mining Results Violate Privacy. "Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining", USA, August 22-25, 2004

[2] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis, State-of-the-art in privacy preserving data mining, : ACM SIGMOD Record, Vol 33(1):50–57, 2004

[3] C.N. Modi, U.P. Rao, D. R. Patel, Maintaining Privacy and Data Quality in Privacy Preserving Association Rule Mining. "2nd International Conference on Computing Communication and Network Topology", Karur, July 29-31, 2010

[4] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and V. Verykios, Disclosure limitation of sensitive rules, : In Proceedings of KDEX'99, pages 45–52, 1999

[5] P. Fule, J. Roddick, Detecting Privacy and Ethical Sensitivity in Data Mining Results. "In Proceedings 27th Australian Computer Science Conference (ACSC)", Australia, 2004

[6] R. Agrawal, R. Srikant, Privacy Preserving Data Mining, : ACM SIGMOD, Volume 29 Issue 2, June 2000

[7] Delis, V. S. Verykios, A. A. Tsitsonics, A Data Perturbation Approach to Sensitive Classification Rule Hiding. "SAC '10 Proceedings of the 2010 ACM Symposium on Applied Computing, ACM", New York, USA, March 22-26, 2010

[8] Y. Guo, Y. Tong, S. Tang, D Yang, A FP-Tree_Based Method for Inverse Frequent Set Mining, pp. 152-163, 2006.

[9] Y Guo, Reconstruction-Based Association Rule Hiding "In Proceedings of SIGMOD2007 Ph.D. Workshop on Innovative Database Research (IDAR2007)", Beijing, China, June 10, 2007.

[10] E. Bertino, I.N. Fovino. Information Driven Evaluation of Data Hiding Algorithms, "7th International Conference on Data Warehousing and Knowledge Discovery", Lecture Notes in Computer Science, Vol 3589/2005, pp. 418–427, 2005.

[11] I.N. Fovino, A. Trombetta. Information Driven Association Rule Hiding Algorithms. "International Conference on Information Technology", Gdansk, May 18-21, 2008

[12] J. Vaidya, C. Clifton. Privacy-Preserving Data Mining: Why, How, and When, IEEE Computer Society, 2004.

[13] V. S. Verykios, A. Gkoulalas-Divanis, A Survey of Association Rule Hiding Methods for Privacy, Privacy Preserving Data Mining Models and Algorithms, Springer Publishing Company Inc., pp. 267-286, 2008

[14] Y.H.Wu, C.M. Chiang, and A. L. P. Chen,: Hiding Sensitive Association Rules With Limited Side Effects, IEEE Transactions on Knowledge and Data Engineering, Vol 19(1):29–42, Jan , 2007.

[15] E. R. Omiecinski, Alternative Interest measures for mining Association in databases,: IEEE Transactions on Knowledge and Data Engineering, Vol. 15(1):57-69, Jan-Feb, 2003.

[16] M. Nadeem, S. Asghar. S. Fong, Hiding Sensitive Association Rules Using Central Tendency. "Advanced Information Management and Service (IMS), 2010 6th International Conference", Islamabad, Pakistan, Nov 30-Dec 02, 2010.

[17] Evfimievski, T. Grandison, Privacy Preserving Data Mining, 2008.

[18] Clifton, D. Marks, Security and Privacy Implication of Data Mining. "In proceeding of ACM SIGMOD workshop on Data Mining and Knowledge Discovery", 1996.

[19] V.S. Verykios, A.K. Emagarmid, E. Bertino, Y. Saygin, E. Dasseni, Association Rule Hiding,: IEEE Transactions on Knowledge and Data Engineering, Vol 16 (4), April, 2004

[20] E. D. Pontikakis, A.Tsitsonis, and V. S. Verykios, An experimental study of distortion-based techniques for association rule hiding. "In Proceedings of the 18th Conference on Database Security (DBSEC)", pages 325–339, 2004

[21] Y. Saygin, V.S. Verykios, C Clifton,:Using Unknowns To Prevent Discovery of Association Rules, SIGMOD. Vol 30(4), 45–54, December, 2001

[22] X. Sun, P.S. Yu, A border-based approach for hiding sensitive frequent item sets. "In Proceedings of The Fifth IEEE International Conference on Data Mining (ICDM)", pp. 426–433, Brisbane, Austrila, Nov 27-30, 2005.

[23] G. V. Moustakides and V. S. Verykios, A max-min approach for hiding frequent item sets. "In Workshops Proceedings of the 6th IEEE International Conference on Data Mining (ICDM)", pages 502–506, 2006.

[24] H. Mannila and H. Toivonen, Levelwise Search and Borders of Theories in Knowledge Discovery Data Mining and Knowledge Discovery, Vol 1(3), 241–258, 1997.

[25] Introduction to Data Mining with Case Studies by G.K.Gupta.

[26] R. Agrawal, T. A. Imielinski, Swami, Mining Association Rules between Sets of Items in Large Databases "In Proceedings of ACM SIGMOD Conference On Management of Data". Washington DC. pp.207–216, June, 1993.

[27] R. Agrawal, H. Mannila. R. Sirikant, Fast Discovery of Association rule, 1994

[28] http://archive.ics.uci.edu

[29] S. R. Oliveira, O. R. Zaiane, Protecting Sensitive Knowledge by Data Sanitization."In proc. Of IEEE International Conference on Data Mining". 2003.

[30] S. R. Oliveira, O. R. Zaiane, Privacy Preserving Frequent Itemset Mining. " In Proceedings Of IEEE International Conference on Data Mining Workshop on Privacy, Security, and Data Mining" 2002.

[31] www.cs.helsinki.fi/group/bioinfo/teaching/dami_s10/dami_lecture4.pdf