Privacy Preserving in Association Rules Using Genetic Algorithm

Submitted By
Rahat Ali Shah
(471-FBAS/MSCS/F08)

Supervised By
Dr. Sohail Asghar
Director at UIIT
PMAS- Arid Agriculture University, Rawalpindi

Co-Supervised By
Dr. Ayyaz Hussain
Assistant Professor at DCS & SE
FBAS, International Islamic University, Islamabad

A thesis submitted to the
Department of Computer Sciences and Software Engineering
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE IN COMPUTER SCIENCE

Department of Computer Science and Software Engineering
Faculty of Basic and Applied Sciences
International Islamic University Islamabad

August 2012

August 2, 2012

Final Approval

This is to certify that we have read and evaluated the thesis entitled **Privacy Preserving in Association Rules Using Genetic Algorithm** submitted by **Rahat Ali Shah** under **Reg No. 471-FBAS/MSCS/F08** and that in our opinion it is fully sufficient in scope and quality as a thesis for the degree of Master of Science in Computer Science.

External Examiner
Dr. Mureed Hussain
Program Manager MISNA,
Shaheed Zufikar Ali Bhutto
Institute of Science and Technology (SZABIST), Islamabad

Dr. Mureed Hussain

Internal Examiner
Dr. Ali Daud
Assistant Professor,
Department of Computer Science & Software Engineering,
International Islamic University, Islamabad

Dr. Ali Daud

Supervisor
Dr. Sohail Asghar
Director,
University Institute of Information Technology,
PMAS- Arid Agriculture University, Rawalpindi

Dr. Sohail Asghar

Co-Supervisor
Dr. Ayyaz Hussain
Assistant Professor,
Department of Computer Science & Software Engineering,
International Islamic University, Islamabad

Dr. Ayyaz Hussain

In the Name of

ALLAH,

The most merciful and compassionate, the most gracious and beneficent

Whose help and guidance we always solicit at every step and every moment.

A Dissertation Submitted to

Department of Computer Science,

Faculty of Basic and Applied Sciences,

International Islamic University, Islamabad

As a Partial Fulfillment of the Requirement for the Degree of

Master of Science in Computer Science

DECLARATION

It is stated that this is an original piece of my own work, except where otherwise acknowledge in text and references. This work has not been submitted in any form for another degree or diploma at any university or other institution for tertiary education and shall not be submitted by me in future for obtaining any degree from this or any other university or institution. If any part of the system is proved to be copied out from any source or found to be reproduction of any project from any training institute or educational institutions, I shall stand by the consequences.

Rahat Ali Shah
471-FBAS/MSCS/F08

ACKNOWLEDGMENTS

The period I spent on my thesis was the most exigent and hard, yet very interesting and gratifying time of my life. I dealt with a range of research problems in the field of Privacy Preserving in Data Mining (PPDM) and Association Rule Mining.

- I would like to express my solemn gratefulness to my supervisor Dr. Sohail Asghar and co-supervisor Dr. Ayyaz Hussain for their persistent guidance and support, and encouragement on following my thesis on priority basis. In fact, I choose the Data Mining as research area because of first lecture of Dr. Sohail Asghar during my Master Program.
- I am thankful to my brother who helped and supported me in implementation of proposed technique.
- I would also like to thank anonymous examiners for reviewing this thesis.
- I am also thankful to all of my friends for their constant support and encouragement.

ABSTRACT

Association rule mining is one of the data mining techniques used to extract hidden knowledge form large datasets. This hidden knowledge contains most of the times confidential information that the users want to keep private or do not want to disclose to public. Therefore, privacy preserving data mining (PPDM) techniques are used to preserve such confidential information or restrictive pattern from unauthorized access. Furthermore, a rule or pattern is marked as confidential and need to hide if its revelation risk is above some given threshold. Numerous techniques are used to hide sensitive association rules by performing some modification in the original dataset. Most of the existing techniques are based on support and confidence framework. In addition, we identified that most of the techniques are suffering from the side effects of lost rules, ghost rules and other side effect, such as number of transaction modified and hiding failure. These effects play an important role in the motivation of proposed architecture. In current research work, genetic algorithm (GA) is used to triumph over the above mention side effects. Proposed research work can be divided into three phases. In phase 1, k-frequent itemsets are generated and then association rules are generated from these itemsets. Privacy Preserving Genetic Algorithm PPGA is applied to release a sanitize database in order to hide sensitive association rules in second phase. In phase 3, the original database is compared to sanitize database, to find the number of lost rules and ghost rules. In order to test the performance of the PPGA based framework, experiments were conducted on Zoo [75], Synthetic [76] and Extended Bakery [77] datasets. Experimental results show that the proposed framework gives better results than the existing state of the art techniques based on rule hiding distance, no of lost and ghost rules.

Table of Contents

List of Tables

List of Figures

CHAPTER 1: INTRODUCTION

Chapter 1

INTRODUCTION

In this chapter, we discuss the background study of Privacy Preserving in Data Mining (PPDM). In the same direction, motivations and research objectives will be discussed in a concise manner. Finally, we will formulate the problems occur, in order to achieve privacy of confidential information. At the end of this chapter, the outline and flow of the thesis is described.

1.1 Overview

Mining association rules is one of the data mining techniques which is used to extract useful or hidden knowledge from large dataset. Such extraction provides information to unauthorized user that organization wants to keep private or do not disclose to public (i.e., name, address, age, salary, social security number, type of disease and the like). The process of PPDM is used to hide confidential information from any type of mining algorithm [1,2,3,7,11,12]. Moreover, the basic objective of PPDM is to protect data against serious adverse effect. The privacy regarding data mining is divided into two types. The fist type of privacy, called output privacy, is that the data is altered so that the mining result will conserve certain privacy. Many modification techniques such as perturbation, blocking, aggregation, swapping and sampling are used for this type of privacy [4,5,8,9,14,17,18,20,23]. The second type of privacy, called input privacy, is that the data is manipulated so that the mining result is not affected or less affected. The cryptography based and reconstruction based techniques are used for this type of privacy [10,15,16,19,21].

Mining association rule is a two step process. In step 1, Apriory algorithm is used to mine frequent k-itemsets [28] from huge amount of data. In step 2, association rules are derived from the frequent k-itemsets. Furthermore, a rule is called sensitive if its discloser risk is above a user specified threshold. In addition, sensitive rules contain confidential data that we do not want to disclose to public. Example: consider two retailers Bob and Ali in a supermarket. Bob is the older one and Ali has newly joined the market. Now, Ali wants to place those items or products which the customers purchase more or whose purchase ratio is high. For this purpose, he wants to see the Bob association rules. Suppose any customer of Bob who buy milk as well as tea. We call that this rule is sensitive for Bob. Similarly, if Ali knows that any customer of Bob who buy milk as well as tea, he started a coupon scheme that offer some discount on milk with purchase

of tea. Gradually, Bob sale of milk with tea decreased and Ali sale increased. Consequently, Ali monopolizes the market as shown in Figure 1.1.

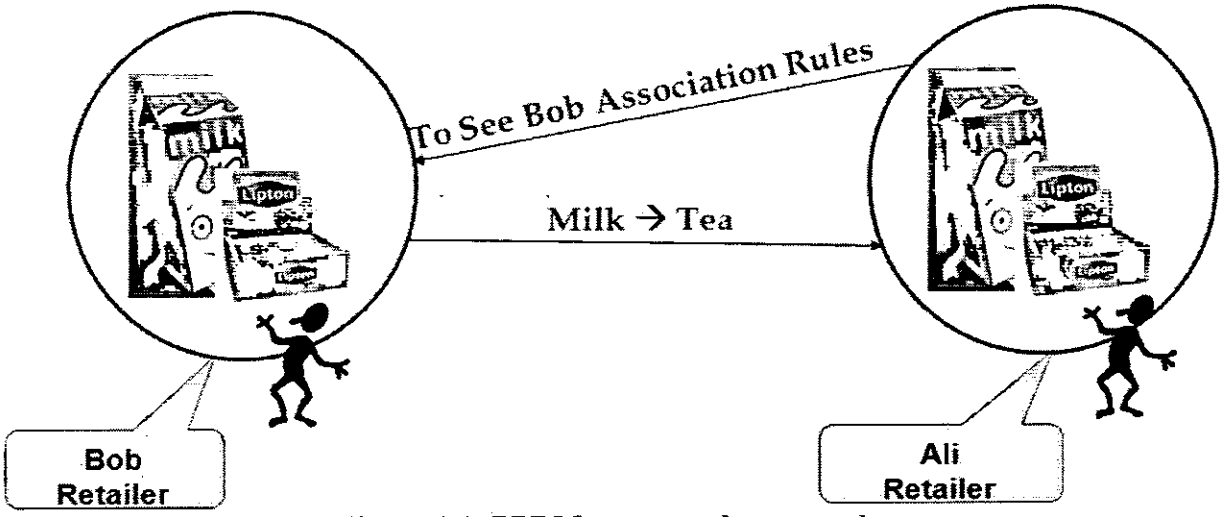

Figure 1.1: PPDM supermarket example

Additionally, association rules (ARs) are divided in to two sub category; weak association rules (WARs) and strong association rules (StARs) as shown in Figure 1.3. A rule is called weak association rule if its confidence is lower than user specified threshold. Similarly, a rule is marked as strong association rule if its confidence is greater than or equal to user specified threshold. Moreover, strong association rules are further divided in to sensitive (SARs) and non-sensitive association rules (NSARs). Furthermore, two strategies are used to hide sensitive association rules (SARs) [13].

- Increase support of the antecedent.
- Decrease support of the consequent.

This work is based on support and confidence framework. The support is the measure of the occurrences of a rule in a transactional database while the confidence is a measure of the strength of the relation between sets of items. An association is an implication of the form $X \Rightarrow Y$ where $X \subseteq I$, $Y \subseteq I$, and $X \cap Y = \emptyset$. Where $I=\{i_1,i_2,i_3.....I_m\}$ be set of literals, called items. X is called body or antecedent (tail) of the rule and Y is called head or consequent of the rule. An example of such a rule is that 60% of customers buy bread also buys butter. The confidence of the rule will be 100%, which means that 60% of records that contain bread also contain butter. The confidence of the rule is the number of records that contain both left hand side *(X)* and right hand side *(Y)*, divided by number of records that contain left hand side *(X)*, which is calculated with the following formula

$$Confidence\ (X \Rightarrow Y) = \frac{|XUY|}{|X|} \quad (1.1)$$

The support of the rule is the percentage of transaction that contain both left hand side *(X)* and right hand side *(Y)*, which is calculated with the following formula

$$Support\ (X \Rightarrow Y) = \frac{|XUY|}{|N|}, \text{ where } N \text{ is the number of transaction in } D. \quad (1.2)$$

Numerous techniques are used in the literature to hide sensitive association rules by performing some modification in the original dataset. The modification causes the problem of lost rules and ghost rules side effects. In current research work, we are trying to improve the existing PPDM in the domain of lost rules and ghost rules side effects by using genetic algorithm. Here binary dataset is passed as initial population to privacy preserving genetic algorithm PPGA. Similarly, the PPGA modifies the database recursively until the support or confidence of the restrictive patterns drop below the user specified threshold. In this work distortion is used as a modification technique, i.e. replacing 1□s to 0□s and vice versa. The proposed technique hide sensitive pattern successfully by reducing the lost rules and ghost rules side effects to zero in best case. Moreover, the technique can be applied for small as well as for large dataset such as medical, military and business dataset.

1.2 Problem Statement

Numerous techniques are used in the literature to hide SAR by reducing the support or confidence. Normally, this is done by modifying some transaction or item in the dataset. This process causes the problem of hiding failure, lost rules and ghost rules as shown in Figure 1.2.

Problem 1 occurs when some sensitive patterns remains after the hiding process, we call this **Hiding Failure,** is the percentage of sensitive patterns that is discovered from sanitize dataset. It can be measured by formula as shown in equation 1.3.

$$HF = \frac{\#Sp(D')}{\#Sp(D)} \quad (1.3)$$

Where *# Sp (D)*, denotes the number of sensitive pattern discovered from database *D*. *D`* denotes sanitize dataset while *D* is the original dataset.

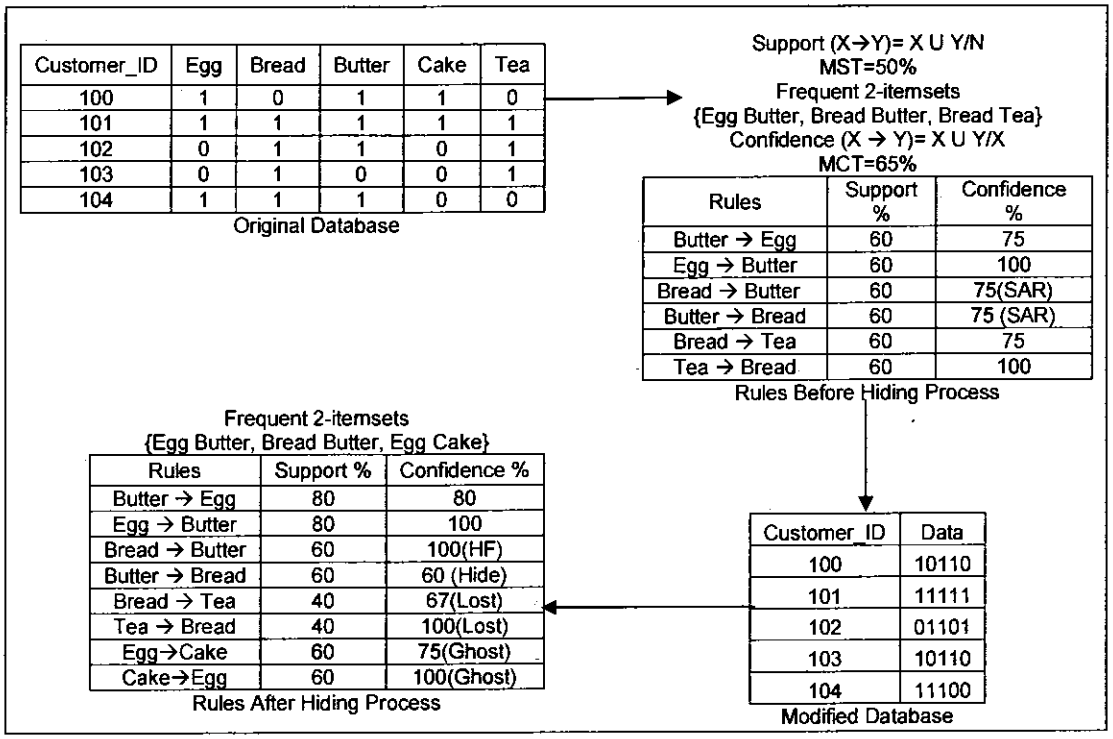

Figure 1.2: Problems causes by PPDM

Problem 2 occurs when some non-sensitive patterns falsely hidden during the hiding process, we call this **Lost Rules,** is the percentage of non-sensitive patterns that is not discover from sanitize dataset $D`$ and can be measured by formula as shown in equation 1.4.

$$LRs = \frac{\#\sim Sp(D) - \#\sim Sp(D')}{\#\sim Sp(D)} \quad (1.4)$$

Where $\# \sim Sp\ (D)$, denotes the number of non-sensitive association rules or pattern discovered from database D. Moreover, the lost rules and hiding failure is directly proportional. Similarly, the more sensitive pattern we hide, the more non-sensitive pattern we loss.

Problem 3 occurs when some unwanted patterns discover during hiding process, we call this **Ghost Rules,** is the percentage of artificial pattern that is discovered from sanitize dataset $D`$ but not discover from original dataset D. It is measured by formula as shown in equation 1.5.

$$GRs = \frac{|P| - |P \cap P'|}{|P'|} \quad (1.5)$$

Where $|P|$, denotes the number sensitive patterns discovered from D and $|P|$, denotes the number of artificial patterns discovered from $D`$.

1.3 Motivation

Organizations such as customer relationship management (CRM), telecommunication industry, financial sector investment trends, web technologies, demand and supply analysis, direct marketing, health industry, e-commerce, stocks & real estates, understanding consumer research marketing, e-commerce and product analysis generate huge amount of data. This huge amount of data contains useful information that organizations do not want to disclose to public. Through data mining, we are able to extract useful information. Agarwal et al. [28], mine associations rules between sets of items in large databases. Moreover, privacy preserving data mining PPDM techniques are used to preserve such confidential information or restrictive patterns from unauthorized access [1,2,3,7,11,12]. However, hiding confidential information causes side effects. The side effects may be in the form of lost rules, some non restrictive patterns lost and ghost rule, some new rules are falsely generated, not support by the original database as shown in Figure 1.3.

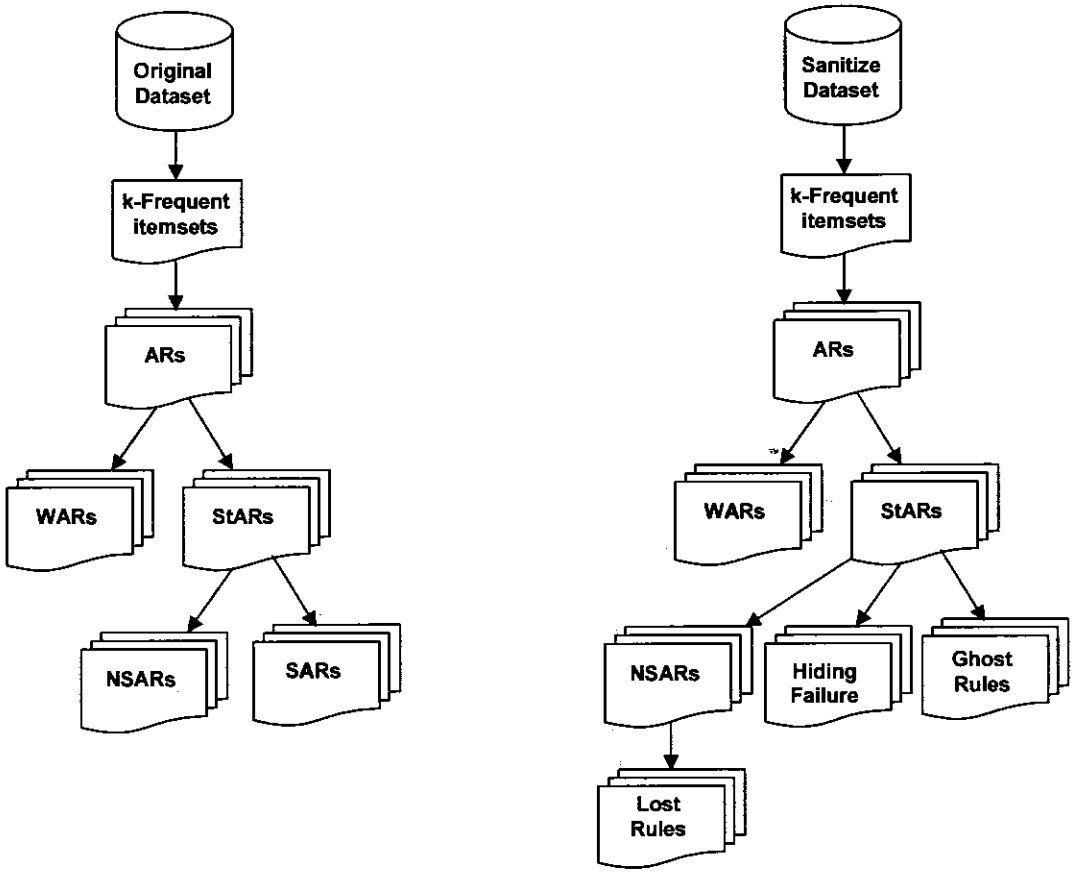

Figure 1.3: Association rules hierarchy before and after sanitization

PPDM is an extremely complex domain and need to standardize [54]. Such standardization in PPDM refers as NP-hard problem [6]. However, to provide an optimal solution to such a hard problem is an important motivation. In current research work, we are trying to improve the existing PPDM in the domain of lost rules and ghost rules side effects.

1.4 Research Objective

A lot of research has done in the area of PPDM (Privacy Preserving in Data Mining).The primary objective of PPDM to preserve confidential data from serious adverse affect (do not disclose to public) [11]. Association analysis is a powerful and popular tool for discovering relationship hidden in large dataset [25]. Moreover, the relationship can be represented in form of frequent itemsets or association rules. Furthermore, a rule is marked as sensitive if its discloser risk is above some given threshold.

The objectives of this research are:

- To achieve hiding failure should be null.
- To minimize the number of transaction modified.
- To minimize side effect in term of lost rules.
- To reduce the ghost rule side effect to zero.

1.5 Thesis Outline

The remaining thesis is organized as outlined below. Figure 1.4 describes the visual representation of thesis outline.

- **Chapter 1** describes the overview of Data Mining (DM), Association Rules (AR), Privacy Preserving Data Mining (PPDM) in association rules, the objective of PPDM, problems with PPDM, motivation and the objectives of the thesis.
- **Chapter 2** presents the literature review related to PPDM in association rules. In this chapter we discussed numerous PPDM techniques. Moreover, we found limitation in literature. Furthermore, we presented the research contribution of their work. At the end this chapter, we presented analysis of literature in term of table.
- **Chapter 3** defines the proposed model base on the identified limitation in the literature. In this chapter, we presented flow of proposed architecture. Moreover, we discussed the different component of the proposed framework. We also discussed the Genetic Algorithm (GA) and the different operators of GA. Furthermore, we presented algorithm

for the proposed model that how we program the proposed model. At the end of this chapter, we presented analysis of in term of table.

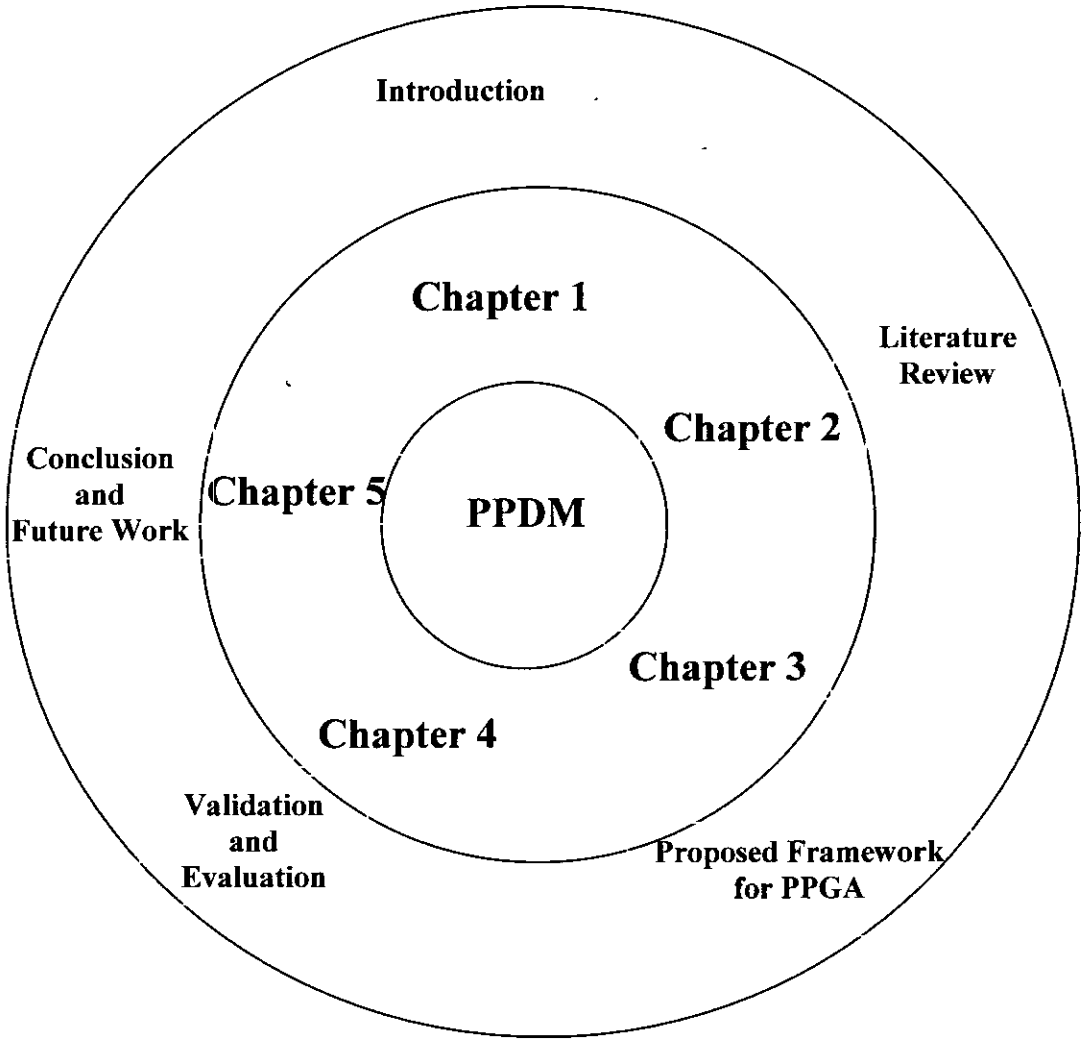

Figure 1.4: Visual Representation of Thesis Outline

- **Chapter 4** presents a detail overview of the results obtained after implementation of the proposed architecture. In this chapter, we presented the implementation of privacy preserving genetic algorithm PPGA in NetBean IDE 6.9.1 as development tool and jdk 6.0 as programming language. Additionally, we discussed the results obtained from Zoo dataset [75], Synthetic dataset [76] and Extended Bakery dataset [77]. Finally, the claim is validated by comparing the proposed model with other techniques in the literature.
- **Chapter 5** provides conclusion of the current research work. This chapter also presents the future work direction to carry out further work in such an important research area.

CHAPTER 2: LITERATURE REVIEW

Chapter 2

LITERATURE REVIEW

Sharing data is often beneficial but sometimes disclose confidential information. Privacy preserving data mining (PPDM) techniques are used to preserve confidential information from unauthorized access. In this chapter, we focus on issues regarding privacy preserving in association rules (PPARs). In this context, we review the literature in order to analyze and find limitation in existing literature. In this direction, we classify privacy preserving data mining techniques into three major classes: border-based approach, exact approach and heuristic approach as illustrated in Figure 2.1.

2.1 Border-Based Approach

This class of approach preserve the privacy of confidential knowledge, by modifying only a selected portion of itemsets which belong to the border in the lattice of the frequent (i.e. statistically significant) and the infrequent (i.e. statistically insignificant) patterns of the original dataset. Particularly, the revised borders (which hold the privacy of confidential data) enforced to preserve restrictive patterns in modified database. An analysis concerning the use of this approach in association rule mining can be found in the work of Sun et al. [53], Sun et al. [55] and Mannila et al. [56].

Suppose F is the set of all frequent itemset in D. We define the *negative border* of F, denoted as $B^{-}(F)$, to be the set of all infrequent itemsets from D in which all proper subsets appear in F.

$$B^{-}(F) = \{X \subseteq I : X \notin F \wedge \forall Y \subset X : Y \in F\}$$

ex:acd: infrequent ac, cd, ad: frequent

$$\text{acd} \in B^{-}(F)$$

Symmetrically, we define the *positive border* of F, denoted as $B^{+}(F)$, to be the set of all maximally frequent itemsets appearing in F,

$$B^{+}(F) = \{X \subseteq I : X \in F \wedge \forall Y \supset X : Y \notin F\}$$

ex: ac: frequent ac#: infrequent (#: any item)

$$\text{ac} \in B^{+}(F)$$

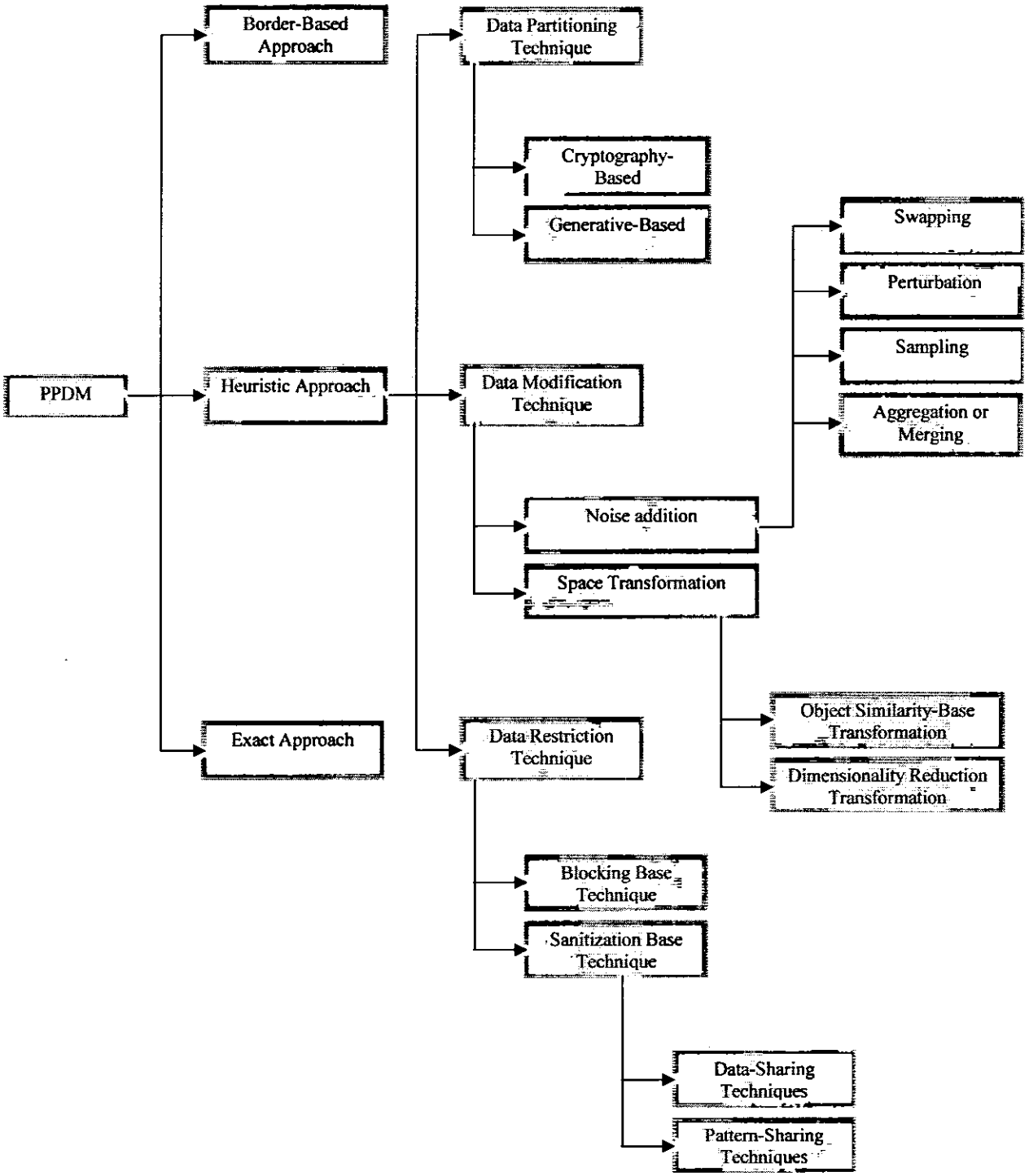

Figure 2.1: Literature Review Hierarchy

Example [56]. Consider the discovery of frequent sets with attributes $R = \{A, \ldots\ldots, F\}$. Assume the collection F of frequent sets is

$$F = \{\{A\}, \{B\}, \{C\}, \{F\}, \{A, B\}, \{A, C\}, \{A, F\}, \{C, F\}, \{A, C, F\}\}$$

The positive border of this collection contains the maximal frequent, i.e.,

$$Bd^{+}(F) = \{\{A, B\}, \{A, C, F\}\}$$

The negative border, in turn, contains sets that are not frequent, but whose all subsets are frequent, i.e., minimal non-frequent sets. The negative border is thus

$$Bd^{-}(F) = \{\{D\}, \{E\}, \{B, C\}, \{B, F\}\}$$

2.2 Exact Approach

This class of approaches involves non-heuristic algorithms. These algorithms consider the privacy of restrictive patterns as a constraints satisfaction problem (an optimization problem). Moreover, these approaches are targeted towards integer or linear programming to solve the optimization problem. Typically, the methodologies use in these approaches can guarantee that an optimal hiding solution exists, if there is optimality in the computed hiding solution or that an optimal hiding solution does not exist, if there is very good approximate solution. More generally, these approaches are usually slower than the heuristic ones. Similarly, the runtime that is required for the solution of the optimization problem will limit the scope of these approaches. An analysis concerning the use of this approach in association rule mining can be found in the work of Verykios et al. [50] and Menon et al. [57].

2.3 Heuristic Approach

These approaches are targeted toward efficient algorithms which preserve the privacy of confidential knowledge by heuristically select a portion of the transactions to modify. Due to their efficiency and scalability, the majority of researchers investigated these methodologies in the domain of privacy preserving data mining PPDM, in order to preserve the privacy of confidential knowledge. However, in knowledge hiding process the approaches of this class take locally best decision which may not always be globally best. Hence, in most of the time these approaches suffer from undesirable side-effects to find optimal hiding solution. An analysis concerning the use of these approaches in association rule mining can be found in the work of Atallah et al. [6], Chih et al. [26], Modi et al. [34] and Naeem et al. [31]. Heuristic approaches can be further classified into three broad categories namely, data partitioning [19, 21], data modification [12, 34] and data restriction [23].

2.3.1 Data Partitioning Techniques

Data partitioning techniques have been applied to some scenarios in which the databases available for mining are distributed across a number of sites, with each site willing to share only data mining results, not the source data. In these cases, the data are distributed either horizontally or vertically. Horizontal partition discussed by Kantarcioglu et al. [19] and vertical partition discussed by Vaidya et al. [21]. In horizontal partition the transactions are distributed in multiple partitions (different data base records placed on different places or sites) while in vertical partition the attributes are split across multiple partitions (different attributes or columns placed on different places or sites). Data partitioning techniques can be classified into two sub category; Cryptography-Based Techniques and Generative-Based Techniques.

2.3.1.1 Cryptography-Based Techniques

Cryptography-Based techniques are used to solve the *secure multiparty computation (SMC)* problem, presented by Du et al. [43], Goldreish et al. [44], and Pinkas et al. [45]. Similarly, secure multiparty computation (SMC) exist; when two or more party want to communicate but neither party want to disclose confidential data to third one.

2.3.1.2 Generative-Based Techniques

The idea behind this approach was first introduced by Veloso et al. [58].In this approach, each party shares just a small portion of its local model that is used to construct the global model. The existing solutions are built over horizontally partitioned data. Meregu et al. [59] discussed privacy preserving distributed clustering using generative model.

2.3.2 Data Modification Techniques

In this approach, some of the values in original database are modified, in doing so, privacy preservation is ensured. . In these techniques, the dataset chosen is binary transactional dataset and the entry value is flipped only. The data is altering by replacing 1's to 0's and vice versa until the support or confidence of association rules is drop below certain threshold. The technique is further divided into noise addition techniques and space transformation techniques.

2.3.2.1 Noise Addition Techniques

In this approach, some noise (e.g., information not present in a particular record or transaction) is added to the original data to prevent the discovery of confidential data or to preserve the privacy of confidential information. In other cases, noise is added to confidential attributes by randomly shuffling the attribute values to prevent the discovery of restrictive patterns that are not supposed

to be discovered. The technique is further divided in to four sub categories: perturbation, aggregation or merging, sampling and swapping.

- **Perturbation [1]:** Modify the original value of attributes, by changing 1 to 0 or by adding noise.
- **Aggregation or Merging [51]:** The combination of several values into a broad category.
- **Sampling [5]:** Modify data for only sample of a population.
- **Swapping [52]:** Interchange values of individual records (transaction).

2.3.2.2 Space Transformation Techniques

This technique is targeted toward privacy preserving clustering. Moreover, these techniques must not only meet privacy requirements but also guarantee valid clustering results. A new technique, hybrid geometric data transformation method was introduced by Agrawal et al. [4]. Similarly, this technique not only meets privacy requirements but also guarantee valid cluster result. In the same direction, Oliveira et al. [62] introduces a new hybrid geometric data transformation method for privacy-preserving clustering, called Rotation-Based Transformation (RBT). Oliveira et al. [63], two new space transformation techniques were described, called object similarity based-representation and dimensionality reduction-based transformation.

- **Object Similarity Based-Representation:** The idea behind this approach is the similarity between objects. In this technique, if the data owners want to share data, first they compute dissimilarity matrix (matrix of distances) between object and then share such a matrix with third party. Many clustering algorithms in the literature operate on a dissimilarity matrix [64]. For instance, matrix of distances (similarity between objects) discloses the confidential knowledge if one party know all the coordinate of a few points. Moreover, we refer this approach for privacy preserving clustering over centralized data.
- **Dimensionality Reduction-Based Transformation:** This technique is applicable when the attributes of object reside either in a central location or split across multiple site. Similarly, we refer this approach as privacy preserving clustering over partition data.

2.3.3 Data Restriction Techniques

The prime objective of data restriction technique is to limit access to mining results. Particularly, the techniques can be classified as generalization, suppression of information or by blocking the access to some pattern that are not hypothetical (imaginary) to be discovered. In the same direction, the work done by Saygin et al. [23, 48] prevents the discovery of sensitive association

rules by modifying the values from known towards unknown. Moreover, this technique decrease the confidence of the rule blow the minimum confidence threshold by placing question mark "?" in place of original value. Furthermore, the technique produces uncertainty of the support and confidence of the rule without distorting the database. This technique is further divided in to two sub category: Blocking-based techniques and Sanitization-based techniques.

2.3.3.1 Blocking-Based Techniques

This technique modifies the original value of attributes by an unknown or question mark "?". Moreover, the technique is useful to hide confidential information if data are share for mining. Furthermore, the technique is applicable to preserve privacy in association rule and classification rule. It means that the private information remains private after hiding process. The technique was first introduced by Johnsten et al. [65, 66] to preserve privacy in classification. Later on, this technique was extended by Johnsten et al. [67] to preserve privacy in association rules. In this work a new methodology was introduced to hide confidential information in relational database and to control the unauthorized access to private data. In the same direction, the work done by Saygin et al. [23,48], introduced a set of algorithms which hide the sensitive information by replacing certain attributes of data items with question mark "?" or unknown, instead of deleting the items.

2.3.3.2 Sanitization-Based Techniques

These techniques are applicable to hide sensitive information or to preserve privacy in classification by purposefully suppression some items in transactional or relational databases, or even by generalizing information. It is further divided into data-sharing techniques and pattern-sharing techniques.

- **Data-Sharing Techniques:** The idea behind this approach was first introduced by Atallah et al. [6]. Such techniques hide the restrictive patterns that contain confidential information by performing some modification in the original data. In doing so, only a small number of transactions that contain restrictive patterns will be modified by removing some item or by adding noise. Moreover, the author proved that optimal sanitization is an NP-Hard problem. In the same direction, the work done by Dasseni et al. [14], introduced new algorithm which hide sensitive association rules by modify original data values and associations by changing some items from 0 to I in some transactions.

- **Pattern-Sharing Techniques:** These techniques insure privacy preserving in association rules by removing the restrictive patterns before sharing the data. In addition, the technique acts on sensitive rule instead of the data itself. A cohesive structure was introduced by Oliveira et al. [68] in the domain of preserving the privacy of restrictive patterns.

2.4 Privacy Preserving of Association Rules

Association rule is one of the data mining techniques used to extract hidden knowledge from large dataset. Sometime this hidden knowledge leak-out confidential information. In this section we review literature on privacy preserving in association rules.

Clifton et al. [40] discussed the security and privacy implication of data mining in a broad scale in order to preserve the privacy of confidential information. They presented the idea of limiting access to the database, eliminate unnecessary grouping, augmenting data, audit and fuzzy data. In this research they did not propose any specific algorithm.

The problem of Privacy Preserving in Data Mining (PPDM) was first presented by Atallah et al.[6]. They proved that optimal sanitization is NP-Hard problem. Moreover, they proposed a heuristic to exclude sensitive frequent itemsets, by deleting item from the transaction in the database.

Verykios et al. [22] discussed the issues regarding privacy preserving association rules. In this research, the author introduced five techniques namely algorithm 1.a, 1.b, 2.a, 2.b, 2.c. These algorithms base on support and confidence framework. Generally, algorithm 1.a hide association rules by increasing the support of the rule antecedent until the rule confidence below the minimum confidence threshold. Algorithm 1.b preserve privacy of association rules by decreasing the support of the rule consequent until either the support or confidence drop below the user specified threshold. Similarly, algorithm 2.a preserves the privacy of confidential information by decreasing the support of rule until their support drop below the minimum support threshold. The lost two algorithms hide restrictive patterns by decreasing the support of their generating itemset until their support is drop below the minimum support threshold. More generally, we can say that algorithm 1.a, 1.b and 2.a are rule oriented while algorithm 2.b and 2.c are itemset oriented. All of these algorithms use distortion (by replacing 1's by 0's and vice versa) as a modification technique. Moreover, the performance of these algorithms is measured on two aspect: efficiency; the time needed by each algorithm to hide a set of rules, and side

effects in term of lost rules; the number of rules that falsely hidden during the hiding process, and ghost rules; the number of unwanted rules (not support by original database) generate during the hiding process. Concisely, the time required for these algorithms is leaner or directly proportional to the volume of dataset and the cardinality of the hiding rules. The side effect of algorithm 1.a in term of lost rules and ghost rules decreases if the cardinality of the hiding rules decreases, otherwise, generate high ghost rules side effects. Similarly, algorithm 1.b, 2.a and 2.b gives mandatory result for lost rules side effects. In addition, these algorithms minimize ghost rule side effects for large databases. The algorithm 2.c, generate high side effect in term of lost rules and generate zero side effect in term of ghost rules. More precisely, none of these algorithms is best for all measure.

In the context of privacy preserving in association rules, Chih-Chia et al. [26] proposed a novel algorithm, FHSAR, for Fast Hiding Sensitive Association Rules. The technique hide sensitive association rule successfully by scanning the database only once. In doing so, the execution time minimize. The goal of the technique is to released database D', such that none of the sensitive association rule is derived and also to minimize lost rule and ghost rule side effects. The proposed technique is a two step process. In step 1, the algorithm established a relationship between transaction and restrictive patterns as shown in Figure 2.2.

Original database

ID	Transaction
1	1,2,4,5,7
2	1,4,5,7
3	1,4,6,7,8
4	1,2,5,9
5	6,7,8

Sensitive association rules

R.id	X→Y	X∪Y	X
1	1,2→5	1,2,5	1,2
2	1,4→7	1,4,7	1,4
3	1,5→7	1,5,7	1,5
4	6→8	6,8	6

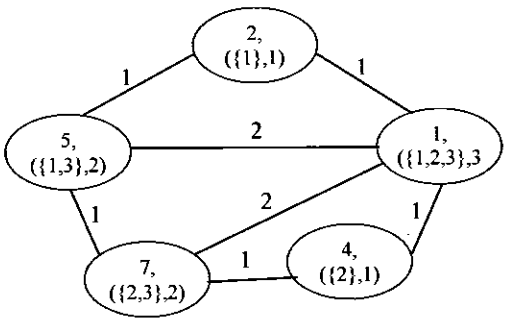

Figure 2.2: The correlation between t1 and sensitive association rule SAR [26]

Each transaction is assigning a prior weight W_i by using formula as shown in equation 2.4.1.

$$W_i = MIC_i / 2^{(|t_i|-1)} \qquad (2.4.1)$$

Where $MIC_i = max(|R_k|)$.

The prior weighted transaction (PWT) can be achieved by organizing table in decreasing order by W. According to heuristic, the selected item is removed. In stage 2, transactions are

modified one by one repeatedly until SAR {} =Ø. The comparison result of FSHAR shows that it is more efficient than other in term of CPU time required. It generates less new rules than previous work done. The bottleneck of FSHAR is the number of lost rules and performance in term of W, which is computed again after each item modified and the transaction is inserted in to the PWT by decreasing the order of W.

In the same direction, a new method was introduced by Remesh et al. [32] in the domain of PPDM. In this research, no modification or editing is performed in the original dataset to reduce the support and confidence of association rule. According to this approach, a rule is considered to be sensitive, if there is a sensitive item in the left hand side of the rule. Moreover, the author investigates new terms or variables to preserve the privacy of association rules. These are: $M_{confidence}$ (modified confidence), $M_{support}$ (modified support) and Hiding Counter. Furthermore, the support and confidence is modified by using the hiding counter. The conventional definition of support and confidence is defined as in equation 2.4.2 and 2.4.3

$$Confidence = \frac{|XUY|}{|X|} \tag{2.4.2}$$

$$Support = \frac{|XUY|}{|N|} \tag{2.4.3}$$

The modified confidence and support is depicted in equation 2.4.4 and 2.4.5 [32].

$$M_{confidenc}(X \rightarrow Y) = \frac{|XUY|}{|X| + hiding\ counter\ of\ rule\ x \rightarrow y} \tag{2.4.4}$$

$$M_{support}((X \rightarrow Y) = \frac{|XUY|}{|N| + hiding\ counter\ of\ rule\ x \rightarrow y} \tag{2.4.5}$$

Initially, hiding counter is set to zero. After that, it is incremented by one until $M_{confidence}$ $X \rightarrow Y$ drop below a Minimum Confidence Threshold MCT. As $M_{confidence}$ $X \rightarrow Y$ drops below MCT, the rule $X \rightarrow Y$ is said to be concealed. Generally, this approach did not mention about the modified database $D`$, from which the sensitive association rules may not derived. Similarly, the absence of release database is a question mark on this approach. Moreover, the technique is also unable to describe that the non-sensitive pattern may not be lost and also new pattern not be generated during hiding process. These side effects limit the scope of this approach.

In the context of privacy preserving association rule, Wang et al. [27] introduced two techniques, increase support of the LHS (ISL) and decrease support of RHS (DSR). In this

research, a blocking technique (replace a value with unknown ?) is used to hide sensitive predictive association rules. A predictive association rule set is the smallest rule set that make the same prediction as the whole association rule set by confidence priority. Similarly, a sensitive predictive association rule is a rule in the predictive association rule set that contains sensitive items on the left hand side of the rule [27]. Generally, the proposed technique based on support and confidence framework. In this work, the author used support and confidence interval. So that, the minimum support threshold MST of an itemset falls between minsup and maxsup of an itemset, and minimum confidence threshold MCT of the rule can be any value between minconf and maxconf of the rule. Moreover, the support of an item decreased by modifying the selected item in any transaction from 1 to?. Similarly, the support of an item increased by modifying the selected item in any transaction from 0 to ?. Comparatively, the performance of the proposed algorithms is compared with Saygin et al. [23]. Typically, the proposed technique required less number of database scanning. Moreover, the approach shows high side effect in term of lost rules. Furthermore, the opponent can easily obtain the SAR by replacing ? to 1 or by replacing ? to 0 and mine the database. In addition, the technique also has not shown the experiment on large dataset. Consequently, these side effects limit the scope of the proposed technique.

Zhang et al. [38] introduced a new technique in the domain of privacy preserving association rule. In order to preserve privacy of association rule, the proposed technique uses two processes, adding weak association transaction (WAT); a transaction that partially support association rule, and removing strong association transaction (SAT); a transaction that provide strong devotion to the mine rule. Moreover, the support of association rule is decreased by adding WAT to a transactional database or removing SAT from transactional database. In this work, the author investigated four modification strategies for modification of weak association transaction. These strategies are: null substitution, unknown substitution, data substitution and direct usage. Comparatively, the performance is not compared with other approaches in the literature. Moreover, the proposed technique fails to hide sensitive association rules successfully. Furthermore, the author has not specified the database on which experiment were performed. The experimental results show us, that the lost rules and ghost rules side effect is high.

The technique proposed by Duraiswamy et al. [24], investigated a solution to preserve confidential information from unauthorized access. According to this approach, a rule is called sensitive, if it has sensitive item in the Right Hand Side (RHS). This approach add together

sensitive rule in to a cluster. Typically, a rule is said to be hidden if support of the sensitive item reduces from MST (Minimum Support Threshold). Moreover, the technique uses distortion method to transform source database D into release database $D`$, so that the sensitive rule that contain sensitive item in right hand side may not derived using any kind of mining algorithms. The efficiency of the algorithm is, the time taken to search sensitive rule in the database, is reduced because the sensitive rules are clustered. Additionally, the technique fail to hide sensitive association rule that contain sensitive item in both side. Furthermore, the lost rules side effect is also high.

Krishna et al. [33] proposed a novel method to derive statistical and fuzzy association rules from quantitative data. In real world the data is always available in quantitative values. In order to generate Booleanized Association Rules (BARs), the quantitative data will be first converted in to booleanized data and then passed this booleanized data to Apriori algorithm to generate BARs using support and confidence framework. After that, Statistical Association Rules (SARs) and Fuzzy Association Rules (FARs) are generated from quantitative data using other relationship measure instead of support and confidence. These measures are: Mean and Standard as represented in equations 2.4.6 and 2.4.7 [33].

$$X = \frac{1}{n}\sum_{i=1}^{n} Xi \tag{2.4.6}$$

$$\sigma = \sqrt{\frac{1}{n}(Xi - \mu)^2} \tag{2.4.7}$$

Let D be a transactional database that contain set of attributes $\{A, B, C, ...P\}$ and n transaction. Hence, the statistical association rules can be presented in the form of association rules (ARs) as depicted in equation 2.4.8 [33].

$$A(\overline{X}_{1a}, \sigma_{1a}), B(\overline{X}_{1b}, \sigma_{1b}) \Rightarrow C(\overline{X}_{1c}, \sigma_{1c}) \tag{2.4.8}$$

Where X, σ denote the Mean and Standard Deviation and A,B,C denote attributes in database D respectively.

The fuzzy logic is used to mine Fuzzy Association Rules (FARs) from quantitative data. Therefore, the membership function is used to find the membership value that will be either 1 or 0 for an element as shown in equation 2.4.9 [33].

$$mf_x(x) : D \rightarrow [0,1] \tag{2.4.9}$$

Generally, the FARs is presented in pair form such as <attribute, linguistic term> which is easy to asses and understands by the user. However, the membership function can be built for any easy understanding with fuzzy set such as Low, Medium, High, and Very High. Hence, FAR is represented as below.

A(Low/0.8, Medium/0.2), B(Medium/0.7,High/0.3) → C(High/0.1, Very High/0.9) [39]

This rule can be interpreted in the form such as *A(Low/0.8), B(Medium/0.7) → C(Very High/0.9)*. Thus, the theoretical interpretation is based on higher values.

Consequently, the authors converted quantified data into Booleanized Association Rules BARs and generate association rules using the Apriori algorithm [33]. Furthermore, the quantified database is used for the generation Statistical Association Rules SARs and Fuzzy Association Rules FARs using the commodity dataset and the results are compared with the BARs. The main strength of the proposed approach is to describe the behavior of each attribute in form of association rules. Additionally, the clustering technique dependency is removed by using cross validation to cluster data in an optimal and automated way. Moreover, dissimilarity among values of clusters is calculated with the use of coefficient variation which is the ratio of σ to $\bar{x}$.In spite of reasonable benefits, the measures used in proposed approach are highly influenced either from the very low or very high value of commodities. Similarly, the proposed method does not hide the sensitive data and display all the patterns either interesting or not. Finally, the results are not very easy to interpret especially by choosing either on high values or low values of the measures in the rule.

Gupta et al. [25] discussed the problem of fuzzy association rule hiding derived from quantitative data. A lot of research has been done to hide boolean association rules, which is concerned whether an item is present in a transaction or not (discovered from binary dataset). But in real world the data is always available in quantitative values, which is concerned the quantity of an item e.g. weight in pounds, typing speed (discovered from quantitative dataset). In this research, a new hiding technique introduced, called Decrease Rule Support (DRS), to hide fuzzy association rules derived from quantitative data. This technique based on support and confidence framework. Generally, the input of the algorithm is source database *D*, Min-Support Threshold (MST) and Min-Confidence Threshold (MCT). The goal of the algorithm is to release a database *D`*, so that, the interesting fuzzy association rules cannot be derived. A rule is side to be hidden if its support drops below the minimum support threshold or its confidence decrease

than minimum confidence threshold. Moreover, two strategies are used to decrease the confidence of a rule $A \rightarrow B$. The first one is to increase the support count of left hand side or A and second strategy decreases the support count of right hand side or B. However, the technique used in this approach divides the transactions based quantitative dataset in to region as illustrated in Table 2.1.

Table 2.1: Fuzzification of Transaction Data [25]

Transaction	A			B			C			D		
Regions	A_1	A_2	A_3	B_1	B_2	B_3	C_1	C_2	C_3	D_1	D_2	D_3
T_1	0	1	0	1	0	0	0.4	0.6	0	0.6	0	0
T_2	0.6	0	0	0	0.8	0.2	0.8	0.2	0	0	0.2	0.8
T_3	0.8	0.2	0	0.6	0	0	0.2	0.8	0	0	0.4	0.6
T_4	0.6	0.4	0	1	0	0	0.4	0.6	0	0	0.6	0.4
T_5	0	0.8	0.2	0.8	0	0	0.6	0.4	0	0	1	0
Count	**2.0**	**2.4**	**0.2**	**3.4**	**0.8**	**0.2**	**2.4**	**2.6**	**0**	**0.6**	**2.2**	**1.8**

Typically, a membership function is used to transform quantitative values in to fuzzy values (between 0 and 1). After that, Apriori process is applied [33] to generate fuzzy association rules based on fuzzy count as described in Table 2.2 and Figure 2.3. Later on, PPDM technique is applied to hide useful fuzzy association rules. Comparatively, the performance of this approach is better in term of hiding failure and transaction modification. The technique generates less number of lost rule; non-restrictive pattern loss during hiding process, and ghost rule; new pattern generate during the hiding process. Moreover, the technique is used to hide single item rules i.e. $A \rightarrow B$. Typically, these side effects limit the scope of the proposed technique and cannot be generalized.

Table 2.2: Set of Quantitative Data [25]

	A	B	C	D
T_1	10	5	8	3
T_2	3	11	6	14
T_3	6	3	9	13
T_4	7	5	8	12
T_5	11	4	7	10

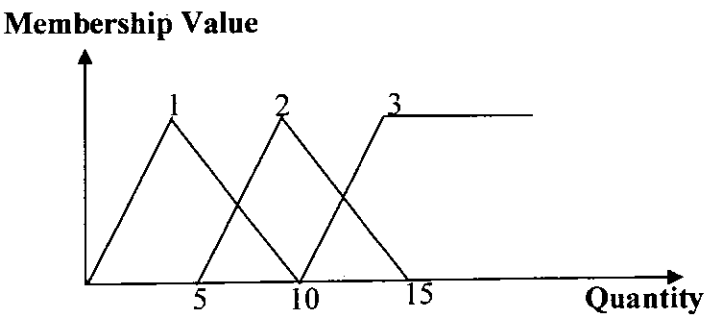

Figure 2.3: Member Function [25]

More recently, a new approach was presented by Dehkordi et al. [30], in the domain of Privacy Preserving in Data Mining (PPDM). In this research work, a novel method for hiding Sensitive Association Rules (SARs) using genetic algorithm was introduced. The technique used in this approach base on support and confidence framework. Generally, the work done in this research divide the database into two parts such as safe transactions; do not contain sensitive items and no need sanitization, and critical transactions; contain sensitive items and need sanitization. Moreover, the technique uses four fitness strategies to preserve privacy in association rules. These fitness strategies are: Confidence based fitness strategy, Support based fitness strategy, Hybrid fitness strategy and Min-Max fitness strategy for the specification of fitness function. All these fitness strategies based on weighted sum function. The solution presented in this approach, uses distortion (replacing 1s by 0s and vice versa) method to transform original database D in to release database $D`$ with minimal side effect, such that the SARs may not derived from the sanitized dataset (a dataset which is released after modification) using any mining algorithms. In doing so, some non-restrictive patterns may be lost, called lost rules, and also some new patters are generated, called ghost rules. Typically, the technique hides sensitive association rules successfully. Particularly, they did not perform experiment on any dataset. Moreover, the side effect in term of lost rules and ghost has not defined clearly. Furthermore, the technique was not compared to the previous techniques exist in the literature. Consequently, all these limitation limit the scope of this research work.

The method proposed by Modi et al. [34], addressed privacy preservation in association. In this approach, a new heuristic, called Decrease Support of Right Hand Side Item of Rule Clusters (DSRRC) were introduced in the domain of PPDM. This approach hide sensitive association rule having single item in right hand of the rule. Moreover, a rule is called sensitive, if it leaks out confidential information or information that individual or organization want to keep private not disclose to public. This approach uses distortion; replacing 1s by 0s and vice versa, as a modification technique. The idea behind this approach is that, the technique first select restrictive pattern and then these patterns are clustered base on common item in the Right Hand Side (R.H.S) of the rules. After this, find the sensitivity of each item and the sensitivity of each rule in the rule clustered. The rule clusters are then sorted in decreasing order base on their sensitivity. Later on, the heuristic used in this approach, calculate the sensitivity of each transaction for each rule cluster. The transactions are then sorted in descending order base on

their sensitivity. In order to preserve the privacy of association rules, the hiding process will start from the highest sensitive transaction and continues until all the sensitive rules in all clusters are not hidden. Consequently, the author compare his result with algorithm 1.b and claim that this technique is better than algorithm 1.b, in term of hiding failure; the technique fail to hide the entire association rules, misses cost; non-restrictive pattern falsely hide during hiding process, artifactual pattern; new pattern generate during hiding process may not support by the original database, and data quality of sanitized database. Particularly, the technique fail to hide sensitive association rule that contain more than one item in the Right Hand Side (R.H.S) of rule. Moreover, the proposed technique generates lost rule and ghost rule side effects.

A unified framework in the domain of preserving the privacy of restrictive patterns was introduced by Chen et al. [35]. In this approach a novel algorithm, Advanced Decrease Support of Sensitive Items (ADSSI) was investigated, to preserve the confidential information from any kind of thread. Moreover, the goal of this technique is to transform the original dataset D into sanitized dataset $D`$, in such way that none of the sensitive association rule is derived. Typically, the technique used in this research complete in three stages. In stage 1, scan database and record useful information in term of Support Count Table (SC Table), item table (I_{ALL} table) and Weight Table (WT). Similarly, in stage 2, weight W_i for each transaction T_i is computed by formula as shown in equation 2.4.10.

$$W_i = \sum i_{sf} \in I_{SF} \bigcap T_i (SC(\{i_s, i_{sf}\}) - Min\_Support * N) / (|T_i \setminus I_{SF}| - 1) \quad (2.4.10)$$

Later on, the transactions in WT are sorted in decreasing order by W_i. Furthermore, the database D is modified; using Decrease Support of Sensitive Items (DSSI) algorithm introduced by Chang et al. [36], in order to completely hide sensitive association rules. In stage 3, the proposed technique is used to modify the released database, insert sensitive item i_s to transaction that do not contain i_s, in order to minimize the lost rule side effects. Moreover, the technique is silent that how we can minimize ghost rules.

Oliveira et al. [37] proposed a model in the domain of preserving the privacy of frequent itemset. In this research, taxonomy of algorithm: naïve algorithm, Minimum Frequent Item Algorithm (MinFIA), Maximum Frequent Item Algorithm (MaxFIA) and Item Grouping Algorithm (IGA) were introduced. Moreover, all these algorithms use inverted file and transactional retrieval engine. Generally, an inverted file consists of vocabulary and occurrences. In this research, the inverted file vocabulary consist the number of items present in transactional

database, their corresponding frequency and for each item there is a correspondence transaction IDs. Furthermore, the transactions IDs correspond to each item are sorted in ascending order as described in Table 2.3.

Table 2.3: Example of transaction modeled by document and the corresponding inverted file [37]

Tr-Id	Items/Terms
T1	ABCD
T2	ABC
T3	ABD
T4	ACD
T5	ABC
T6	BD

Items	Freq	
A	5	→
B	5	→
C	4	→
D	4	→

Vocabulary

T1, T2, T3, T4, T5
T1, T2, T3, T4, T5, T6
T1, T2, T4, T5
T1, T3, T4, T6

Transaction IDs

In addition, binary search is used to search for the transaction IDs of a particular item. The function of transactional retrieval engine is to accept query from algorithm, processes the query using a query language and return the result to algorithm. Furthermore, the inverted file and transactional retrieval engine speed up the searching process. These algorithms fall in two categories, item restriction based; hide sensitive rule by decreasing the support of the frequent itemset, and pattern restriction based; hide sensitive rule by decreasing the confidence of the sensitive pattern. The Naïve algorithm is pattern restriction based and MinFIA, MaxFIA and IGA are item restriction based. Unlike distortion these algorithms selectively remove individual items from sensitive transaction. In addition, these algorithms use discloser threshold ψ, controlled by the user. Consequently, when ψ=0, then the hiding failure will be zero and misses cost will be high in all cases. Similarly, when ψ=100, then the hiding failure will be high and misses cost will be zero. Moreover, the side effect in term of ghost rules is not mentioned.

Yuhong et al. [49] proposed a reconstruction base technique in the domain of privacy preserving association rule. In this research, a new method, called FP-tree, was introduced for inverse frequent set mining. They used two thresholds MST and MCT, to generate rules R. In R, sensitive rules R_h exist such that $R_h \subseteq R$. Hence, R_h represent sensitive association rule that leak out confidential data and need to be secured. Therefore, transformed original database D into release database D', such that none of the Rh is derived from released database. In doing so, the proposed framework considers the frequent itemset, generated from the original database with minimum support threshold and minimum confidence threshold. Moreover, sanitization algorithm is used to extract non-sensitive rules and convert back into the FP-tree. In doing so, modified database is obtained which contains infrequent items. The main advantage of the

model, preserve the confidentiality of restrictive patterns inversely without reducing the minimum support and minimum confidence thresholds. In addition, if multiple transaction have the same itemset then it is very complex to generate the sanitize database. The technique fails to hide all the sensitive association rules and also fail to control the ghost rules and lost rule side effects.

In the same direction, Wang et al. [13] introduced Pattern Inverse tree (PI tree). This technique is used to hide informative association rules. It can be define as the smallest association rules set that performs the same guess as the entire association rule set by confidence priority. Moreover, each node of PI-tree store three type of information: name of the item, the occurrences of items on the path from the current node to root and transaction ID that contain all items from current node to root. Additionally, the construction of PI tree complete in two steps. In step 1, a PI tree and frequency list is built. While in step 2, the rules $x \rightarrow y$ having sensitive item in the left hand side x is sanitized. The approach is used to transform D in to $D`$, so that, no informative association rules containing x on the left hand side is discovered. As a result, the technique fails to hide a rule having sensitive item on right hand side. The results shows us, that the technique loss non sensitive pattern falsely and also generate new pattern which may not support by the original database. These side effects limit the scope of the proposed technique.

Besides the support and confidence of association rules, Malik et al. have proposed other measures in the domain of PPDM [41]. In this approach they define five measures namely Correlation, Coefficient, Laplace, Kappa and J-Measure. They presented that these measures are better in result as compare to conventional support and confidence frame work. The technique use in this approach is completed in four steps. In step 1, signal and text features from images were extracted. In step 2, their frequencies were calculated. In step 3, popular dimensionality reduction techniques were applied to prune non-interesting features. In last step, association rules are generated.

More recently, Naeem et al. [31] proposed a novel architecture in the domain of PPDM. In this approach the author used five measures namely Confidence, All-Confidence, Conviction, Leverage and Lift in order to mine association rules. In addition, the author proposed weighting mechanism, in order to assign a weight to each transaction. Moreover, the weight shows the dependency of transaction on sensitive association rules. The weighting mechanisms used in this

approach are: Sum, Mean, Median and Mode. The database is sorted by using any of these weighting mechanisms.

$$q = \sum_{x=1, y=1, z=1}^{n, |SAR|, |I|} Iz, \forall\ SARy \subseteq TIDx \tag{2.4.11}$$

Equation 2.4.11 [31] represents the total count of single item in Sensitive Association Rule SAR while this single item must be a sub-set of a Transaction. x, y and z are counters, whose values are set by number of transaction, number of SAR, number of Items. n is the maximum number of transaction in dataset up to which x will be counted on. The upper limits of y and z are determined by total number of SAR and total number of items in a SAR. Once the set of count is determined, than the weight is calculated using weighted sum function. Moreover, the approach can be used for single value association rules as well as for multi values association rules. Furthermore, leverage is outperforming in all cases. Consequently, the technique is only applicable on dataset whose attributes not more than 26. The author claim that the technique do not create ghost rules side effects. The side effects in term of lost rules are still available.

2.5 Compare and Contrast

The main focus of rule hiding is association rules and frequent patterns. Aggarwal et al. [60] proposed that preserving the privacy of restrictive patterns refers to the process of modifying the original database in such a way that some restrictive patterns hide without seriously affecting the data and the non-restrictive patterns. The main goal here is to hide as many sensitive rules as possible, while keeping preserved as many non-sensitive rules as possible. Generally, the process of modification or sanitization can be divided in to data blocking and data distortion techniques. The major concept of data distortion techniques, are the replacement of selected values with "false" values (i.e., replacing I's by 0's and vice versa). Moreover, this technique is applicable, in order to reduce the support and confidence of the sensitive association rules from user specified threshold. An analysis concerning the use of this technique can be found in the work of Verykios et al. [22], Duraiswamy et al. [24] and Dehkordi et al. [30]. All of these approach adding false values to real transaction which causes so many side effect problems. Similarly, the major concept of blocking technique, are the replacement of an existing attribute value with "unknown" or "?". In blocking technique the algorithms do not add false value to the database. In addition, to restore a value by an unknown value instead of placing a false value is a little bit

more advantageous for specific application such as medical application. An analysis concerning the use of this technique can be found in the work of Dasseni et al. [14], Weng et al. [27] and Saygin et al. [23, 48]. The solution presented in these approaches, uses blocking method to transform original database D in to release database $D`$ by increasing support of the rule antecedent by changing 0s to ? or by decreasing support of rule consequent by changing 1s to ?. Hence, comparing to other techniques in the literature, these approaches do not distort the database but only change some known values to unknown. The main limitation of these approaches is the privacy violation of the modified database. For example, the opponent can easily leak out the information by replacing question mark by 1s or 0s.

Clifton et al. [40] discussed the security issues and implication of data mining. He did not propose any specific algorithm. Moreover, he investigated the idea of limiting access to the database; supplementing data, remove needless combination, audit and fuzzy data. Later on, Atallah et al. [6] proved that optimal sanitization is an NP-hard problem and need to standardization. In this research, they proposed a first heuristic based on support reduction, to exclude sensitive frequent itemsets. In the same direction, Verykios et al. [22] introduced five algorithms. Generally, these algorithms run on the strategy which is based on reducing the support and confidence of rules. Moreover, the proposed techniques used distortion as a modification technique. The distortion method simply changes the bit values of data items in transactions. Precisely, none of these techniques is best to overcome all the side effects caused by preserving the privacy of association rules. Similarly, the time taken by each algorithm to hide a set of rules is also high. Later on, Chih-Chia et al. [26] proposed a novel algorithm, *FHSAR*, for Fast Hiding Sensitive Association Rules. Typically, the technique hide sensitive association rule successfully by scanning the database only once. In doing so, the time is needed to hide a set of rules is minimized. Generally, the proposed technique assign a weight W to each transaction used a weighting mechanism. Comparatively, the comparison result of *FSHAR* shows that it is more efficient than other in term of CPU time required. Moreover, it generates less new rules than previous works. The bottleneck of *FSHAR* is the number of lost rules and performance in term of W, which is computed again after each item modified. In the same direction, Dehkordi et al. [30] used genetic' algorithm in the domain of privacy preserving in association rules. Moreover, the technique uses four fitness strategies to preserve the confidential information from unauthorized access. The solution presented in this approach, uses distortion (replacing 1s by 0s

and vice versa) method to transform original database D in to release database $D`$. Particularly, the author has not defined any dataset for experimentation. Moreover, the side effect in term of lost rules and ghost not define clearly. Furthermore, the technique was not compare to the previous techniques exist in the literature. More recently, Naeem et al. [31] used five measures namely Confidence, All-Confidence, Conviction, Leverage and Lift in order to mine association rules from large databases. In addition, the author proposed weighting mechanism, in order to assign a weight to each transaction. Moreover, the approach can be used for single value association rules as well as for multi values association rules. Consequently, the technique is only applicable on dataset whose attributes not more than 26. The technique generates zero ghost rules side effects. Furthermore, the technique generate high side effect in term of lost rule.

Additionally, table 2.4 present the overall summary of the critically reviewed literature.

Table 2.4: Summary of Literature Review on Privacy Preserving Association Rules

Author	Technique / Algo	Dataset	Hidden Rules	Lost Rules	Ghost Rules	Hidden Failure	Transaction Modified
Clifton et al. [40] 1996	- Did not propose any specific algorithm	×	×	×	×	×	×
Attallah et al. [6] 1999	-Heuristic approach	×	×	×	×	×	×
Desseni et al. [14] June 2000	-Algorithm 1.a -Algorithm 1.b -Algorithm 2.a	IBM synthetic data generator	2 hidden rules	×	×	×	×
Saygin et al .[48] 2002	- Reduce support and confidence by Safety Margin (SM)	Anonymous web usage data of Microsoft website	√	×	√	×	√
Oliveiraet al. [37] 2002	-Naïve algorithm -MinFIA -MaxFIA -IGA	IBM Synthetic data generator	×	√	0	√	×
Verykios et al. [22] Apr. 2004	-Algorithm 1.a -Algorithm 1.b -Algorithm 2.a -Algorithm 2.b -Algorithm 2.c	IBM synthetic data generator	×	√	√	×	√
Yuhong et al. [49] June 2007	FP-tree base method for inverse frequent set mining	1. BMS-POS 2. BMS-WebView-1 3. BMS-WebView-2	×	√	√	√	√
Wang et al. [61]Aug. 2007	ISL DSR	IBM Synthetic data generator	×	0% ISL 11% DSR	33% ISL 5% DSR	12% ISL 2% DSR	59% for 1item, 128% for 2 item ISL 14% for 1 item, 25% for 2 tiem DSR
Krishna et al. [33] July 2008	- Appriory algorithm was used to Generate *BARs* -Clustering technique to infer these rule - Generate *SARs* used X and SD or σ -Generate *FARs* used fuzzy logic	Commodity Export data available at Reserve Bank of India (RBI)	×	×	×	×	×
Duraiswamy et al. [24] Aug. 2008	SRH	Example dataset	√	×	×	√	×
Chih-Chia et al. [26] Nov. 2008	FHSAR	IBM data generator	×	4-8 for |SAR|=5, 19-22 for |SAR|=10	0-2 for |SAR|=5& 10	×	154-1414 for |SAR|=5, 589-5801 for |SAR|=10
Dehkordi et al. [30] Aug. 2009	-Used *GA* for *SAR* hiding	Example dataset	×	×	×	×	√
Gupta et al. [25] Oct. 2009	DRS	Wisconsin Breast Cancer dataset from UCI Machine Learning Repository	√	√	√	×	√
Chen et al. [35] Dec. 2009	ADSSI	IBM synthetic data generator	√	8%	0	√	×
Modi et al. [34] July 2010	DSRRC	Example dataset	√	36%	0%	0%	6.4%
Naeem et al. [31] Dec. 2010	- Weight generation algorithm	Zoo dataset Lymphography dataset Thyroid0387 dataset Hypothyroid dataset	×	×	0	0	×

- Purpose of √ means that the author presented it clearly.
- Purpose of × means that the author did not presented it clearly.

2.6 Summary

In this chapter we have reviewed the literature on privacy preserving of association rules. From the literature it is clear, that each and every technique has improved privacy preserving of association rules in a single dimension while ignoring the remaining dimension. The problem of optimal sanitization is an NP-hard presented by Atallah et al. [6]. In order to analyze the side effects, the following parameters are presented in the literature.

- **Lost Rules [25, 26, 49, 50, 51]:** Non-sensitive association rules which are falsely hidden during the hiding process, by transforming the original database into sanitize database.
- **Ghost Rules [25, 26, 48, 49, 50, 51]**: New rules which are falsely generated during the hiding process not support by the original database.
- **Hiding Failure [48, 50, 51]:** Some techniques do not hide the entire sensitive association rules.
- **Hidden Rules [24, 25, 26, 48, 49, 50]:** Association rules which are not generate after the hiding process.
- **Modified Transaction [24, 25, 26, 48, 49, 50, 51]:** The number of transaction which is modified during the hiding process.

In the next chapter, we are presenting a model for privacy preserving of association rules which is going to resolve the highlighted limitation in a more reliable way.

CHAPTER 3: PROPOSED FRAMEWORK

Chapter 3

PROPOSED FRAMEWORK FOR PPGA

In the previous chapter, we have discussed numerous Privacy Preserving Data Mining PPDM techniques. From the literature we found, that most of these techniques are based on support and confidence framework. From this review, we identified that most of the techniques are suffering from the side effects of lost rules, ghost rules and other side effect, such as number of transaction modified and hiding failure. The above mention side effects play an important role in the motivation of proposed architecture. In the proposed architecture, Genetic Algorithm GA is used to triumph over the above mention side effects. This work has a partially resemblance to the work done by Dehkordi et al. [30]. However, the difference is that we have defined our own fitness strategy. The question rises, "Why are we using GA in PPDM?" The answer to such question can be justified that PPDM is an extremely complex domain and need to standardize [54]. Such standardization in PPDM refers to be NP-hard problem [6]. Therefore, GA is used to provide optimal solution to hard problem. Such optimality of solution depends on the complexity of fitness function. The possible strength of fitness function ensures a desirable level of optimal solution. There are other evolutionary approaches also available in the literature. These approaches are Particle Swarm Optimization (PSO) [69], Ant Colony Optimization (ACO) [70], Simulated Annealing (SA) [71], Tabu Search (TS) [72], and Honey Bees Mating Optimization (HBMO) [74]. In this chapter, we will discuss in detail about genetic algorithm, its different operator and the use of genetic algorithm in sensitive association rules hiding.

3.1 Architecture for Privacy Preserving Genetic Algorithm (PPGA)

In the proposed archeticure, the GA is used to preserve the privacy of association rules. Genetic algorithms have been developed by john Holland [73]. Holland's GA is a method for moving from one population of "chromosomes" (e.g., strings of "bits" representing candidate solutions to a problem) to a new population. In terms of genetic algorithm the dataset is called population and transaction is called chromosome. Moreover, GA is evolutionary and meta-heuristic technique used to solve complex problem. Hence, preserving the privacy of association rules is a complex problem and need optimal sanitization. Therefore, GA is used to hide restrictive patterns, $X \rightarrow Y$, by decreasing support of Y or by increasing the support of X. Furthermore, it often requires a

"fitness function". Hence, the fitness function assigns a value to each transaction (chromosome) in the database (population). Additionally, the fitness of the transaction depends on how well that transaction solves the problem at hand. The fitness is calculated with the help of Equation 3.3.

Let D be a set of transaction in a dataset, denoted as $D=\{T_1, T_2, \ldots\ldots\ldots, T_n\}$ and R be a set of identifier, defined as $R=\{1, 2, \ldots\ldots\ldots, n\}$. Each record T_r is defined as a set of data items, $Tr=\{d_1, d_2, \ldots\ldots\ldots\ldots d_k\}$, where I represent a set of identifier, $I=\{1, 2, \ldots\ldots\ldots\ldots\ldots, k\}$.

Let S be a set of sensitive item or sensitive pattern, denoted as $S=\{s_1, s_2, \ldots\ldots\ldots, s_m\}$ and P be a set of identifier for elements of S, defined as $P=\{1,2, \ldots\ldots\ldots\ldots, m\}$.

$$\forall S_p \in T_r \vee S_p \notin T_r,\ 1/\sum_{i=1}^{k} Count(S_p)\ in\ T_r : S_{p \geq 1} \tag{3.1}$$

e.g. $D = \{T_1, T_2, T_3\}$

$T1 = \{Bread, , Butter\}$

$T_2 = \{Bread, Egg, ,\}$

$T_3 = \{Bread, , Butter\}$

$S_p = \{Bread, Butter\}$

Count (S_p) in T_2= (Bread)

$S_p \in T_2, \sum_{i=1}^{k} Count(S_p)\ in\ T_2 = 1$

$$\forall d_i \in T_r, \sum_{i=1}^{k} d_i : d_i \rightarrow [1] \tag{3.2}$$

e.g. $d_i \in T_2, \sum_{i=1}^{k} d_i = (1+1)=2$

Let F be a set of fitness values, defined as $F=\{f_1, f_2, \ldots\ldots f_h\}$, and V be a set of identifier for elements of F, denoted as $V=\{1, 2, \ldots\ldots h\}$.

$$\forall f_v \in T_r,\ where\ f_v = 1/\sum_{i=1}^{k} Count(S_p)\ in\ T_r + \sum_{i=1}^{n} d_i \tag{3.3}$$

e.g. $\sum_{i=1}^{k} d_i = 2$

$\sum_{i=1}^{k} Count(S_p)\ in\ T_2 = 1$

$f_v = (1/1)+2 = 3$

Equation 3.3, describes that the fitness value depend on the number of sensitive item in a transaction. It means that, the fitness function is rule oriented. Moreover, transactions are sorted in descending order on the base of fitness value. Furthermore, the transaction having lower fitness value will be selected for modification. Hence, fitness function goes toward maximization.

Let C be a set of chromosome, denoted as $C=\{O_1, O_2, \ldots\ldots\ldots, O_n\}$ and S be a set of identifier, defined as $S=\{1, 2, \ldots\ldots\ldots, n\}$. Each record O_s is defined as a set of items, $O_s=\{o_1, o_2, \ldots\ldots\ldots\ldots o_k\}$, where I represent a set of identifier, $I=\{1, 2, \ldots\ldots\ldots\ldots\ldots, k\}$.

$|D| = |C| \,|\, \forall\, |O_s| = |T_r|$

$D := C \wedge T_r := O_s$

$$\forall S_p \in O_s \vee S_p \notin O_s,\ 1/\sum_{i=1}^{k} Count(S_p)\ in\ O_s : S_{p \geq 1} \tag{3.4}$$

$$\forall o_i \in O_s,\ \sum_{i=1}^{k} o_i : o_i \rightarrow [0] \tag{3.5}$$

The fitness for each offspring can be calculated by Equation 3.6.

$$\forall f_v \in O_s,\ where\ f_v = 1/\sum_{i=1}^{k} Count(S_p)\ in\ O_s + \sum_{i=1}^{n} o_i \tag{3.6}$$

The question rises, "How do we justify the fitness function."? The answer of such question is justified that the fitness function is divided into two parts. The 1st part is called transaction sensitivity; it increases the priority by decreasing the value of those transactions which contain sensitive item as shown in Equation 3.1 and 3.4. On the base of this equation transactions having maximum number of sensitive items will be selected for modification. The 2nd part is called transaction priority; it increases the priority of selected transactions, contain same number of sensitive items as shown in Equation 3.2 and 3.6.

Table 3.1: Original Dataset

Transaction ID	Bread	Butter	Egg	Tea	Cake
1	1	1	1	0	0
2	1	0	1	0	0
3	1	0	0	1	0
4	0	1	0	0	1

Consider a rule *Bread→Egg*, in Table 3.1, transaction 1 and 2 have the same number of sensitive items such as $T_1=2$ and $T_2=2$. Put these values in Equation 3.1.

$T_1 = 1/2 = 0.5 = T_2$

Now Equation 3.2, is applied to increase the priority of sensitive transactions by counting the availability of data items in each transaction such as $T_1=3$ and $T_2=2$. Thus, put these values in Equation 3.3, to find the fitness of both transactions.

$f_1 = 0.5+3 = 3.5$

$f_2 = 0.5+2 = 2.5$

Similarly, transaction having lower fitness value will be selected for modification. However, T_2 will be selected. Hence, if the priority of sensitive transaction is changed such as $T_1 = 2$ and $T_2 = 3$, then T_1 will be selected which generate lost rule side effect. The fitness justification shows that the lost rule side effect is minimized.

Table 3.2: Sanitize Dataset

Transaction ID	Bread	Butter	Egg	Tea	Cake
1	1	1	1	0	0
2	1	0	0	0	0
3	1	0	0	1	0
4	0	1	0	0	1

Table 3.2, illustrates that transaction *T2* is replaced with offspring (transaction) using Equation 3.4 and Equation 3.5. Equation 3.4, increase the priority by decreasing the value of those offspring that contain sensitive items. Similarly, Equation 3.5, increase the priority of selected offspring, contain same number of sensitive items, by counting the non-availability of data items.

Definition 1: PPGA hide sensitive association rules successfully.

$$\therefore \exists S_p \in T_r \Delta T_r \mid \exists! \, S_p \in T_r \tag{3.7}$$

Equation 3.7, shows that the proposed technique hides sensitive association rules successfully. Because, the technique only modify those transactions in which sensitive items are present.

Definition 2: PPGA minimizes lost rule side effect.

$$\therefore \forall \; S_p \in T_r \wedge f_v \in T_r \mid T_r \Delta \, O_s \; \forall \, f_v \in T_r : f_v \leq f_{v+1} \tag{3.8}$$

Equation 3.8, describes that the lost rules is minimized because, the technique select those transactions to modify in which less number of data items are available.

Definition 3: PPGA minimizes ghost rule side effect.

$$\therefore \forall \; S_p \in O_s \wedge f_v \in O_s \mid T_r == O_s \; \forall \, f_v \in O_s : f_v \geq f_{v+1} \tag{3.9}$$

Equation 3.9, illustrates that the ghost rules is minimized. As the selected transactions are replaced to those transactions (offspring) in which maximum number of data items are unavailable.

In this step, different operators of genetic algorithm are applied. These operators are tournament selection, single point crossover, mutation, and inversion as shown in Figure 3.1. Table 3.3, describes the notation used in the proposed architecture.

Table 3.3: Notation and Definetion

Notation	Detail
D	Original Dataset
D`	Sanitize Dataset
T_r	Transaction ID
AR	Association Rule
SAR	Sensitive Association Rule
MCT	Minimum Confidence Threshold
MST	Minimum Support Threshold
f_v	Fittnes of Each Transaction (Cromosome)
TMG	Transaction Modified in each Generation
O_s	Offspring ID
LRs	Lost Rules
GRs	Ghost Rules

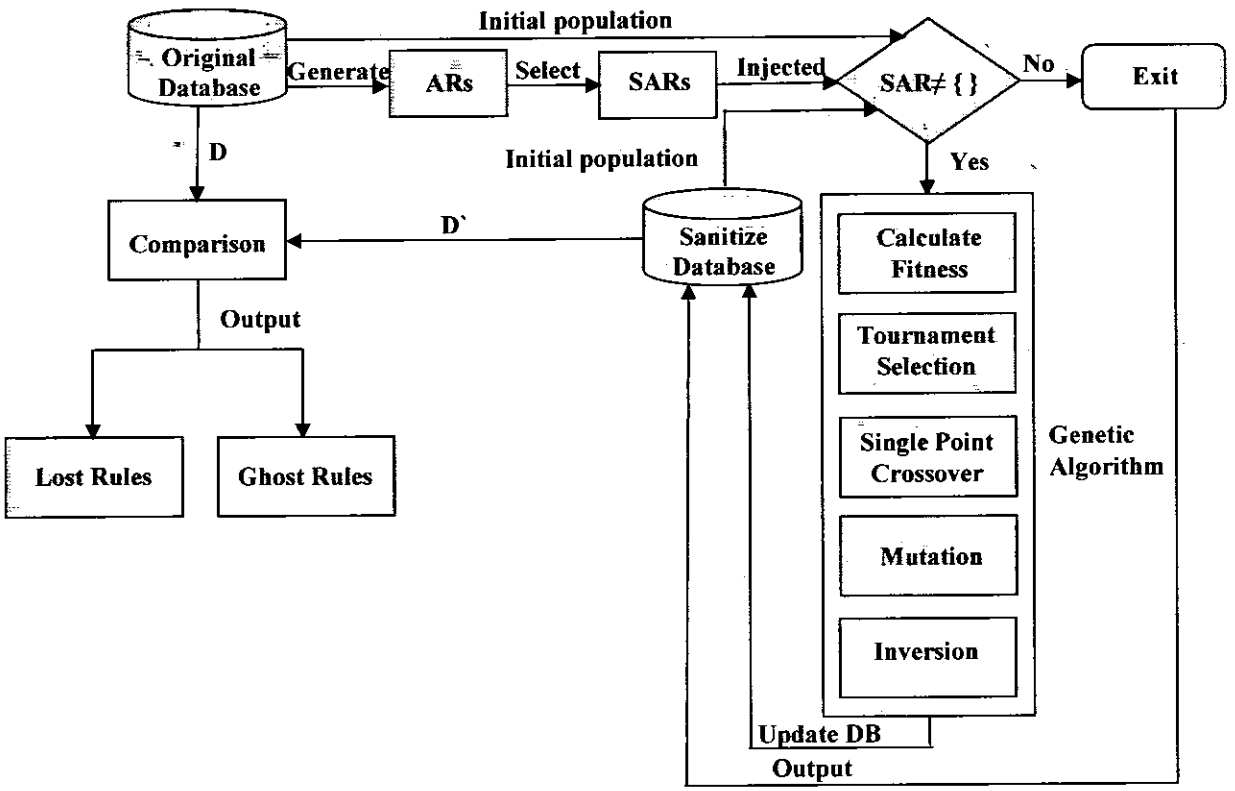

Figure 3.1: Framework of PPGA for Hiding Sensitive Association Rules

- **Tournament Selection:** In Tournament selection two chromosomes are selected randomly from population and more fit of these two is selected for mating pool as shown in Table 3.4.

Table 3.4: T-Selection

Population	Fitness	T-Selection	
11100	3.5	10010 10100	10010
10100	2.5		
10010	3.0		
01001	3.0		

- **Single Point Crossover:** In single point crossover the parent's chromosome is split in to two portions such as head and tail. Similarly, the head of one chromosome combine with the tail of another chromosome in the mating pool as illustrated in Figure 3.2.

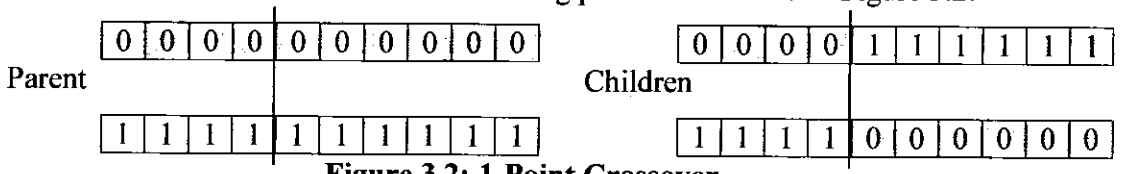

Figure 3.2: 1-Point Crossover

- **Mutation Operator:** Mutation operator randomly changes the values (1 by 0 or 0 by 1) of some locations in the chromosome as depicted in Figure 3.3.

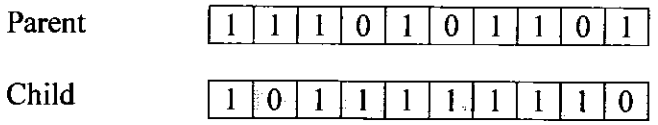

Figure 3.3: Mutation

- **Replacement or Inversion Operator:** In this operator some of the chromosome of initial population will replace with some of the chromosome of offspring as described in Table 3.5.

Table 3.5: Inversion Operation

Population	Offspring	Mutation	Fitness	Inversion
11100	10100	00100	5.0	11100
10100	11010	11110	1.5	00100
10010	11100	10100	3.5	10010
01001	11100	11000	4.0	01001

3.2 PPGA

The process of PPGA is completed in three phases as shown in Figure 3.4. The input values of the algorithm are original dataset *D*, sensitive association rules *'SARs'*, minimum confidince threshold *'MCT'*, minimum support threshold *'MST'* and number of transaction modified in each generation *'TMG'*.

In phase-1, apriori algorithm proposed by Agarwal et al.[28], is applied to mine k-frequent itemset. The association rules are generated from these frequent itemset. In phase-2,

diperent operators of genetic algorithm are applied to hide sensitive association rules by transforming original database D into sanitize dataset $D`$ (modified dataset). Finally, original database $D$ compare with modified database $D`$, to find lost rules and ghost rules side effect.

1. Input: Original Database D, SARs, MCT, MST, N, Replace
2. Output: Transform D into D'

Phase -1

3. FS → Frequent Itemset (D)
4. AR → Generate Association Rules (FS)
5. SAR → Select Sensitive Association Rules (AR)

Phase -2

6. WHILE SAR{} != Ø OR generation != N
7. Fitness: $f_v=1/\sum_{i=1}^{k} \text{Count(Sp)}$ in $T_r+\sum_{i=1}^{k} d_i$: $d_i \rightarrow [1]$
8. Selection: Base on f_v
9. Crossover: $T_r * T_{r+1}$
10. Mutation: Select T_r, Change 1 to 0 or 0 to 1 randomly
11. Fitness: $f_v= 1/\sum_{i=1}^{k} \text{Count(Sp)}$ in $O_s+\sum_{i=1}^{k} o_i$: $o_i \rightarrow [0]$
12. Replace: $T_r \Delta O_s$
13. Wend

Phase -3

14. D $\oplus$ D`

Figure 3.4: Privacy Preserving Genetic Algorithm

3.3 Components of PPGA

The components of the proposed model can be divided into three phases. In phase 1, k-frequent itemsets is generated. In phase 2, privacy preserving genetic algorithm PPGA is applied to release a sanitize database, by performing some modification in original dataset, in order, to hide sensitive association rules. In phase 3, the original database is compared to sanitize database, to find the number of lost rules and ghost rules. The components of PPGA are described in Figure 3.1.

3.3.1 Phase-1 of PPGA

In phase-1, the data which are in CSV file format as shown in Table 3.6 is first converted into Boolean format such as 1 and 0. Moreover, we need a format that describes the availability and non-availability of data items. Thus, 1 represents the availability of data item and 0 represents non-availability of data item as shown in Table 3.7. The items are separated by comma. The question rises, "Why we convert data into Boolean format? The answer of this question is that it is easy to implement. Next, Apriori algorithm is used with some Minimum Supporting Threshold (MST) to mine k-frequent itemsets and then to generate association rules. This process contain following steps.

Table 3.6: Data in CSV file format

Transaction ID	Items bought
1	Bread, Butter, Egg
2	Bread, Egg
3	Bread, Tea
4	Butter, Cake

Table 3.7: Boolean Data in CSV file format

Transaction ID	Bread, Butter, Egg, Tea, Cake
1	1, 1, 1, 0, 0
2	1, 0, 1, 0, 0
3	1, 0, 0, 1, 0
4	0, 1, 0, 0, 1

Step-1: To find frequent 1-itemset or L_1, compare the frequency or support of each 1-itemset to minimum supporting threshold and select those whose support is greater than user specified threshold. E.g., if the minimum support is 50%, then the only three frequent 1-itemsets are generated from Table 3.7.

C1= {Bread}, {Butter}, {Egg}

L_1= {Bread} support 75%, {Butter} support 50%, {Egg} support 50%

Step-2: Generate a set of candidate k-itemsets or C_1 by joining L_{k-1} with itself $(L_{k-1} * L_{k-1})$ e.g. the candidate 2-itemsets or C_2 is obtained by joining L_1 with itself $(L_1 * L_1)$.

C_2= {Bread, Butter}, {Bread, Egg}, {Butter, Egg}

Step-3: In order to find frequent k-itemset or L_k, scan the database to get the support of each candidate k-itemset and compare with minimum support threshold to prune unfrequented k-itemset from this set e.g. the frequent 2-itemset or *L2* is obtained by scanning the database to get the support of candidate 2-itemset or C_2.

C_2={Bread, Butter} support 25%, {Bread, Egg} support 50%, {Butter, Egg} support 25%

L2={Bread, Egg} support 50%

Step-4: Generate association rules from these frequent itemset e.g. only two association rules are generated from the frequent 2-itemset, as in step-3.

R1= Bread→Egg, support is 50% and confidence is 66%

R2= Egg→Bread, support is 50% and confidence is 100%

Step-5: In this step we remove all the duplicate association rules e.g. in the current example, there is no duplicate rule, to remove.

Step-6: Discard some of the association rules whose confidence is lower than Minimum Confidence Threshold (MCT), called weak association rule e.g. if the minimum confidence threshold is 75%, then the rule R_1 is discarded.

Step-7: Finally, select some of the association rules as sensitive association rules whose confidence are greater than or equal to minimum confidence threshold, e.g. in the running example, consider that the rule R_2 is sensitive.

3.3.2 Phase-2 of PPGA

In phase-2, the data which is in CSV file format will first convert to transactional database as shown in Table 3.1. Select some of the rule as sensitive association rule, whose confidence is greater than MCT as shown in Table 3.8. Initially three inputs namely MCT, Sensitive Association Rules (SARs), and initial population is injected to the process of PPGA. The PPGA run repeatedly until the $SARs\ \{\} \neq \emptyset$.

3.3.3 Phase-3 of PPGA

In phase-3, the original database D is compared to modified database D`, to fined the number of ghost rules and lost rules.

3.4 Flow of the Architecture

Flow of the proposed architecture starts from CSV file format as shown in Figure 3.5. In first step, the data which are in the form of <T, F> or <Yes, No> is booleanized (convert to 0 or 1). The Boolean data is then imported to MySQL database. In next step, Apriory is applied with MST to generate k-frequent itemsets. Similarly, association rules are generated from these k-frequent itemsets. Assume that, some of the rules are selected as sensitive association rules which leak-out confidential information. Initially, three inputs original database, sensitive SAR, and MCT are passed to the condition. Then, MCT is compared with the confidence of the SARs. If the confidence of the SAR is greater than or equal to MCT, it means that the SAR set is not equal to empty then process of Privacy Preserving Genetic Algorithm PPGA will start. In this step, first the fitness of each and every transaction will be calculated based on fitness function as shown in Equation 3.3. Next different operators of PPGA are applied. Generally, the process of PPGA will repeat until the confidence of the SAR drops below the MCT or SAR {}=Ø. This repetition performs some modification in the sanitize database. Hence, if the condition becomes false, then the process of PPGA will stop. In the beginning, the original database and sanitize

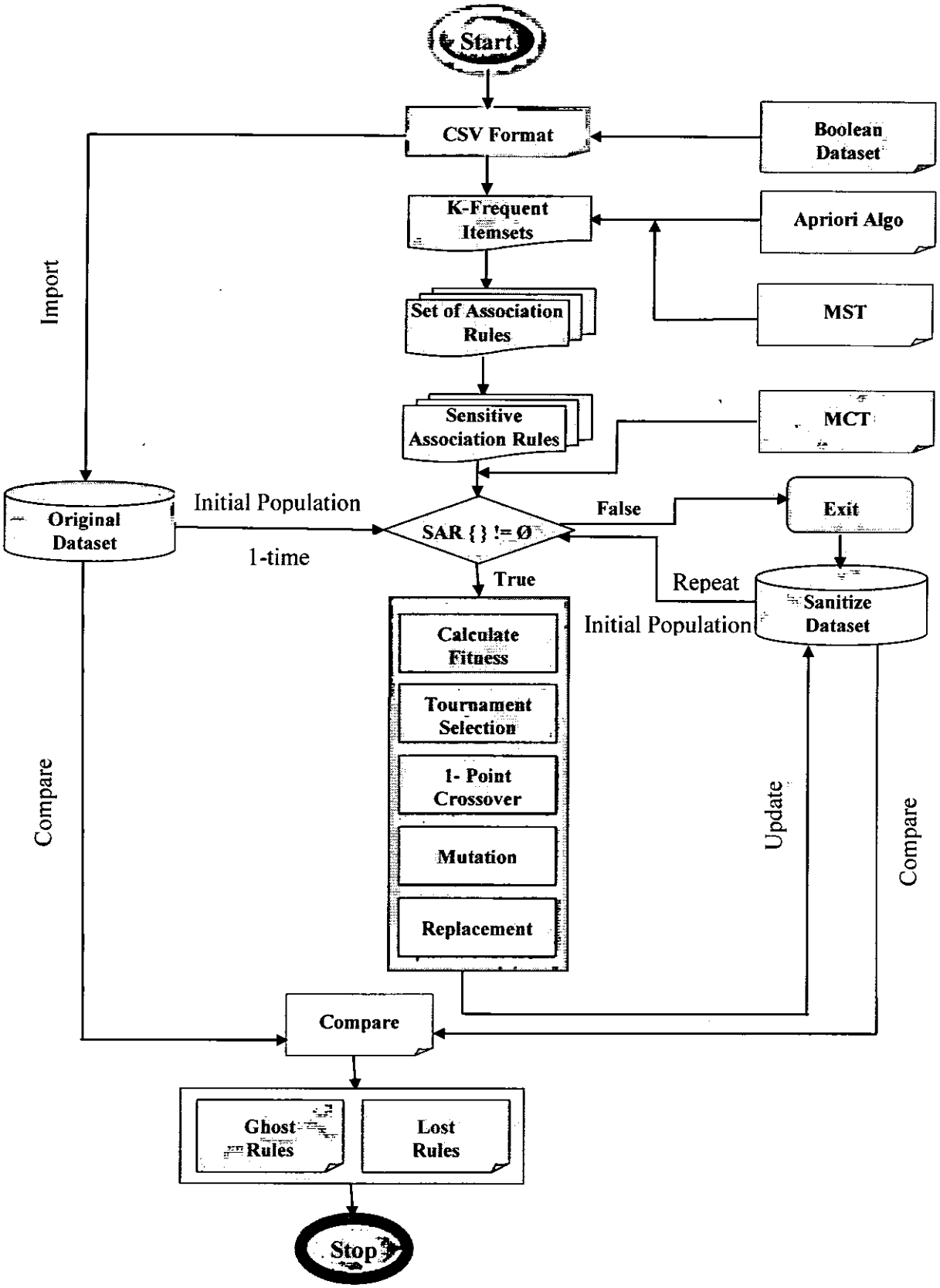

Figure 3.5: Flowchart of PPGA

database are same. At the end, the original database is compared to the sanitized database, to find number of ghost rules and lost rules.

Example: Table 3.6 and 3.7, shows the example dataset in CSV file format. The example dataset contains four transactions and five items in each transaction. If the MST is 50%, then *{Bread, Egg}* is the only 2-itemset that satisfies the minimum support as shown in Table 3.8. Thus, if the MST is increased from 50% then the information of *{Butter}* and *{Egg}* is lost which affect on the association of *{Bread, Egg}*. Hence, if the MST is decreased then the ratio of un-useful information is increased.

Table 3.8: Frequent itemset

Frequent itemset	Support
{Bread}	75%
{Butter}	50%
{Egg}	50%
{Bread, Egg}	50%

If the MCT is 50%, then the only two rules are generated from this 2-itemset, that have confidence greater than 50% as shown in Table 3.9. As we know, that association rules are generated from the frequent itemset. Therefore, the MCT must be greater than or equal to MST. In the current example, if the MCT is 66% then the same result will be shown. Similarly, if the MCT is 67% then the rule, *Bread→Egg*, is lost, which we want to hide.

Table 3.9: Association rules from frequent itemset

Antecedent	→	Consequent	Support	Confidence
Bread	→	Egg	50%	66%
Egg	→	Bread	50%	100%

Table 3.10: Fitness of example dataset

Population	Fitness
11100	3.5
10100	2.5
10010	3.0
01001	3.0

Assume that, rule, *Bread→Egg*, is sensitive or leak out confidential information and need to hide. At first, some input parameter, such as initial population, MST, MCT and SAR passed to PPGA. Later on, the PPGA calculate the fitness for each transaction. On the base fitness transactions are selected for new generation. Hence, the fitness generated by one rule is deferred

from other rule. However, the fitness depends on rule or rule oriented. Table 3.10, describes the fitness of the rule *Bread→Egg* generated from example dataset.

Table 3.11: First iteration of PPGA

T-Selection		Crossover	Offspring	Mutation	Fitness	Inversion
10100 10010	10010	10 010	10100	00100	5.0	11100
11100 10010	10010	11 100	11010	11110	1.5	00100
11100 10100	11100	11 100	11100	10100	3.5	10010
11100 10100	11100	11 100	11100	11000	4.0	01001

Table 3.11 demonstrates the first iteration or generation of PPGA. In the next generation, the sensitivity of the rule is checked by comparing the confidence and support of the rule to user specified threshold. If the sensitivity of the rule is below then specified threshold, it means that the rule is hidden. Subsequently, the modified dataset is compared to original dataset to achieve lost rule and ghost rule side effect. Table 3.12 illustrates that rule, *Bread→Egg*, hides successfully. The rule, *Egg→Bread*, is lost and no new rule is generated during hiding process.

Table 3.12: Performance measure of PPGA

Performance Measure	Association Rule	MCT	MST
SAR	Bread→Egg	50%	50%
Lost Rule	Egg→Bread		

3.5 Summary

In this chapter, we proposed a model to preserve the privacy of association rules. Generaly, we discussed in detail the funtionality, components and flow of the proposed architecture. Moreover, the components of the proposed architecture are divided into three phases. In phase 1, Apriori algorithm is applied to generate k-frequent itemsets. In phase 2, PPGA is applied to transform the original database into sanitize database, by performing some modification in original dataset, in order, to hide sensitive association rules. In phase 3, the original database is compare to sanitize database, to find the number of lost rules and ghost rules.

CHAPTER 4: VALIDATION & EVALUATION

Chapter 4

VALIDATION AND EVALUATION

In the previous chapter, we have proposed a privacy preserving genetic algorithm PPGA to preserve the privacy of confidential information. It also described the flow of the architecture and the number of steps involve in PPGA. In this chapter, we illustrate the implementation of the privacy preserving genetic algorithm PPGA. Hence, for its implementation NetBeans IDE 6.9.1 is used as a development tool and java as a programming language. The java is selected because of its elevated performance graphical user interface. Moreover, we validate the PPGA for sensitive association rules hiding with experimental results. Finally, we compare the results of the proposed framework with the existing techniques. Thus, on the base of experimental results the claim will be verified that the proposed model gives us better results as compare to previous work done.

4.1 Implementation

In this section, the screen shot of the Privacy Preserving Genetic Algorithm PPGA is demonstrated. The development of PPGA is coded in java using NetBeans IDE 6.9.1 as a development tool. Java is selected as a programming language because of its prominent features. Moreover, it provides a high performance graphical user interface. In addition, we have performed series of experiments on a PC with ~2.0 GHz CPU and 2046 MB memory, under the Windows Vista. At first, the data which is in the CSV file format is imported to MYSQL database. After this, Apriory algorithm is used to mine frequent k-itemset [28]. Figure 4.1 describes the association rules generated from frequent k-itemsets before and after sanitization from synthetic dataset well be discussed later on. It indicates the sensitive association rule, 4→7, with support 41% and confidence 81%. The rule is hidden by decreasing the confidence to 74% as shown in the column of Association Rules from Sanitize Dataset. Moreover, the figure also represents the lost rule that is, 7→4. During this hiding process no ghost rules are generated and the number of transaction modified 766 is also shown. Initially, Sensitive Association Rule SAR, Minimum Supporting Threshold MST, Minimum Confidence Threshold MCT and original dataset pass to Privacy Preserving Genetic Algorithm PPGA. The PPGA transform the database

repeatedly until the confidence or support of the sensitive association rule drop below the user specified threshold.

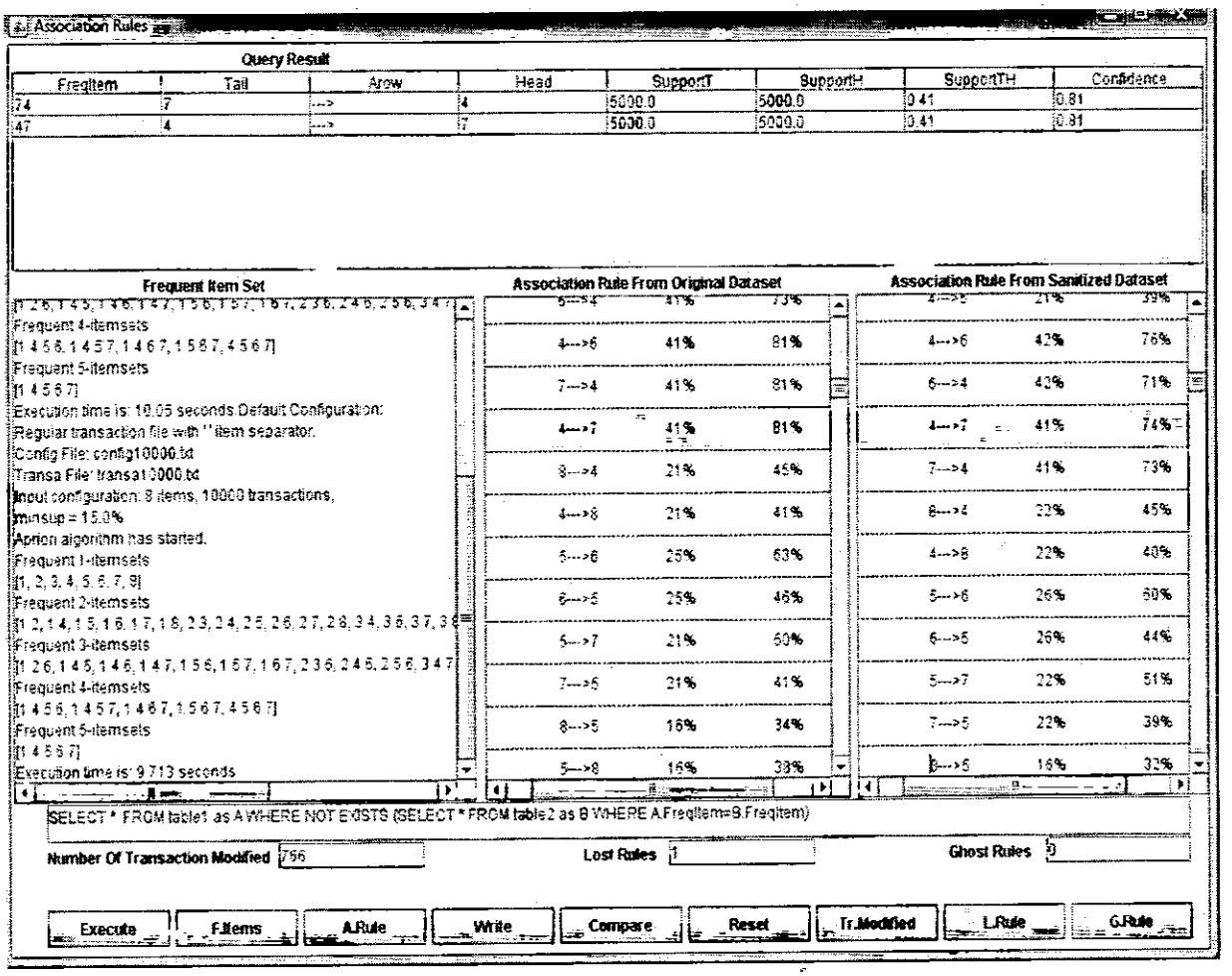

Figure 4.1: Mining Frequent Itemsets and Association Rules from Synthetic Dataset

Figure 4.2 represents the numbers of step involve in Privacy Preserving Genetic Algorithm PPGA. It also depicts the number of repetition or iteration of PPGA. In addition, the figure also describes that what will be the support and confidence value for next generation. The process will stop when the support or confidence drop below the user specified threshold. It indicates that confidence of the rule is decreased after each iteration of PPGA.

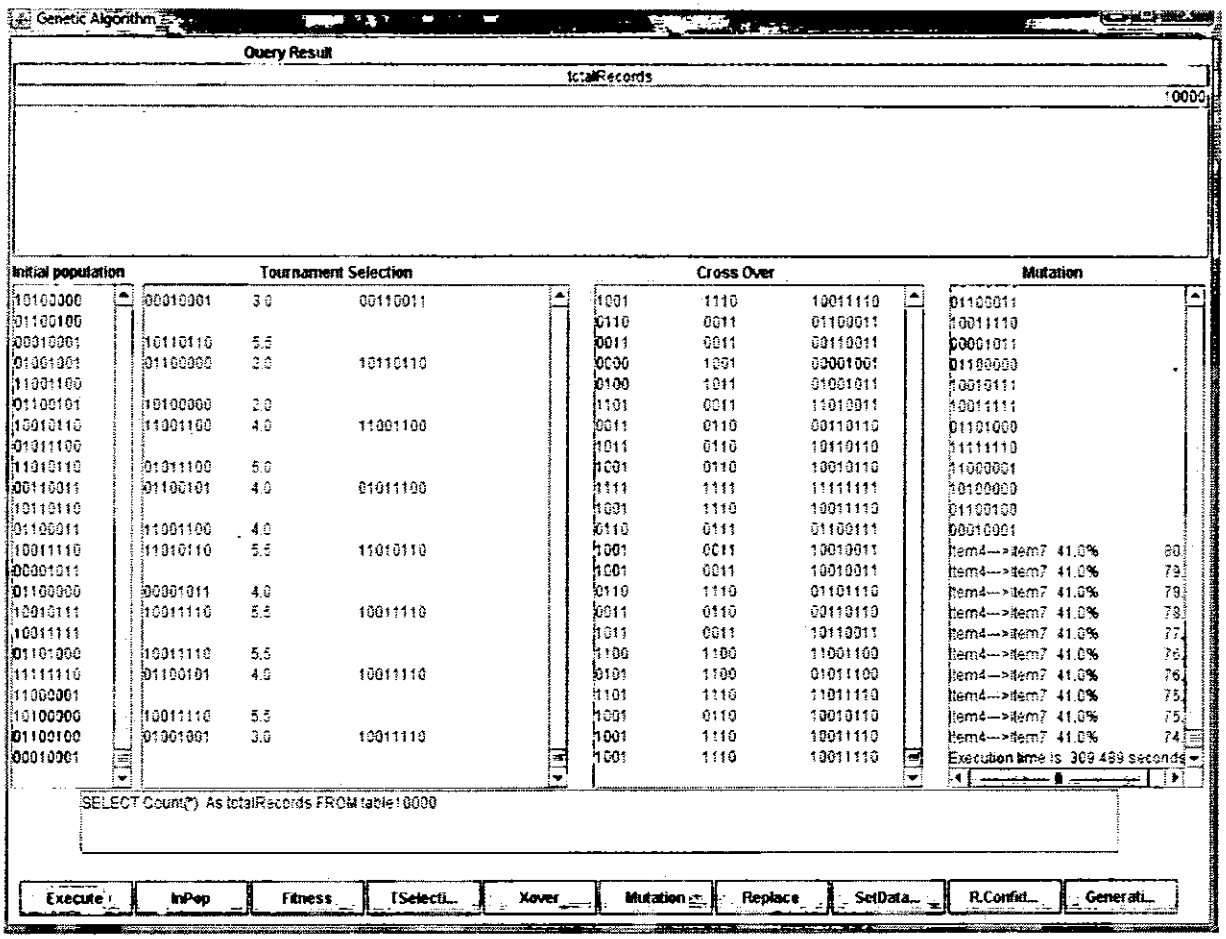

Figure 4.2: Hiding process of PPGA

4.2 Datasets

To test and validate the Privacy Preserving Genetic Algorithm PPGA, experiments were conducted on Zoo dataset [75], Synthetic dataset [76] and Extended Bakery dataset [77]. Moreover, the experiments were performed on those data items which are in Boolean format or convertible to Boolean format. The data items which we cannot convert to Boolean data will be removed as shown in Table 4.1.

Table 4.1: Dataset

Dataset	Total Records	Total Attributes	Ordinal Attributes
Zoo	101	17	15
Synthetic	10000	8	8
Extended Bakery	20000	50	50

4.2.1 Zoo Dataset

This dataset consist of 101 transactions. Each transaction consist 17 attributes. These can be divided into one categorical attribute and 16 quantitative attributes. As we have described, that we are only concern the data items which are convertible into Boolean format. Therefore, we neglect one categorical attribute (Type) and one quantitative attribute (Legs). Hence, the remaining 15 quantitative attributes which are in the form of 'yes' or 'no' are used. Furthermore, the attributes description of this dataset is shown in Table 4.2.

Table 4.2: Zoo Dataset Attributes Description

Attribute	Description	Attribute	Decription
Item-1	Hair	Item-9	Backbone
Item-2	Feather	Item-10	Breathes
Item-3	Egg	Item-11	Venomous
Item-4	Milk	Item-12	Fins
Item-5	Airborne	Item-13	Tail
Item-6	aQuatic	Item-14	Domestic
Item-7	Predator	Item-15	Catsize
Item-8	Toothed		

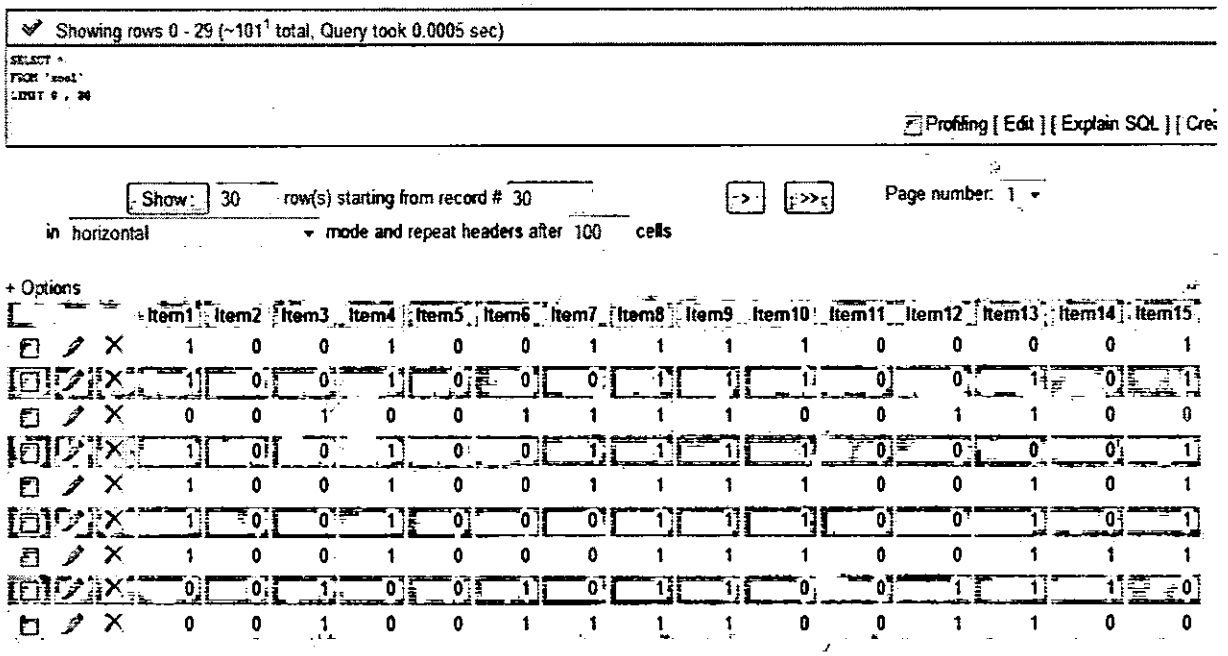

Figure 4.3: Zoo Dataset

Figure 4.3 depicts the sample of Zoo dataset. The attributes values are in the form of '0' and '1' format. Here '0' represents the non-availability of data items while '1' shows the availability of data items.

4.2.2 Synthetic Dataset

It is compose of 10,000 transactions. Each transaction contain 8 items (attributes). Moreover, the attributes description of this dataset describe as the number of items available in a hotel for breakfast as shown in Table 4.3. Furthermore, the arrangement of these items is like this that some peoples take bread as well as tea, Bread→Tea, while others take egg with rice, Egg→Rice, and so on. Figure 4.4 describes the sample of synthetic dataset.

Table 4.3: Synthetic Dataset Attributes Description

Attribute	Description	Attribute	Description
Item-1	Bread	Item-5	Milk
Item-2	Egg	Item-6	Cake
Item-3	Tea	Item-7	Rice
Item-4	Butter	Item-8	Chicken Roost

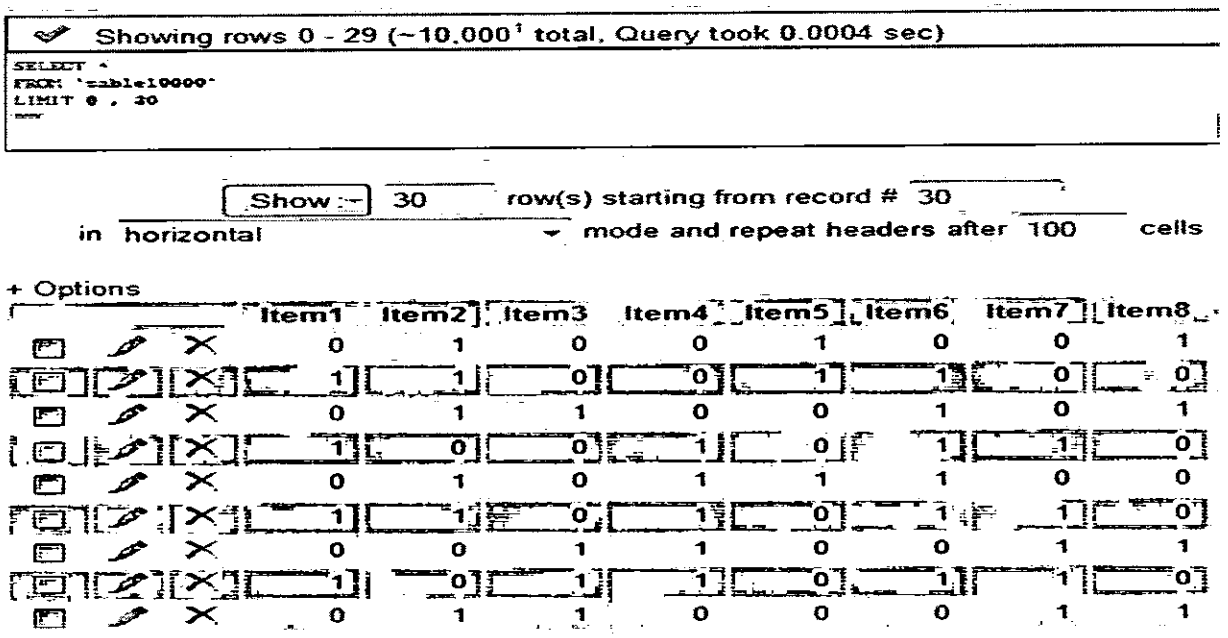

Figure 4.4: Synthetic Dataset

4.2.3 Extended Bakery Dataset

This dataset contains 20,000 transactions. The database store information about food or drink. The number of data items or attributes involve in this dataset are 50. Out of these 40 are pastry items and 10 are coffee drinks. Furthermore, the database is distributed in multiple locations in West Cost State. The numbers of location are California, Oregon, Arizona and Nevada. Table 4.4 shows the sample of the extended bakery dataset.

Table 4.4: Extended Bakery Dataset

Tr-ID	Item-1	Item-2	Item-3	Item-4	----------	Item-48	Item-49	Item-50
1	0	1	0	0	----------	0	0	0
2	0	0	0	0	----------	0	0	0
3	0	0	0	0	----------	0	0	0
.								
.								
.								
19998	0	0	0	0	----------	1	0	0
19999	1	0	0	0	----------	0	0	0
20000	0	1	0	0	----------	0	0	0

4.3 Performance Measures

The performance measures of the proposed model are:

- When some non-sensitive pattern falsely hidden during hiding process, we call this Lost Rules (LRs). In current research work we will minimize the lost rules side effect. It can be measure by formula as shown in Equation 1.4.
- When some artificial pattern discover during the hiding process which may not support by the original database, we call this Ghost Rules (GRs). In this work we will try to reduce the ghost rule side effect to zero. It can be measure by formula as shown in Equation 1.5.
- In current research work we are trying to minimize the number of Transaction Modification (TM). It is the number of transaction modify during the rules hiding process. It can be calculate by comparing the original database to modified database.
- The technique use in this research work will hide the sensitive association rules successfully. Sensitive association rules are rules which contain confidential information and whose support and confidence is greater than user specified threshold.

4.4 Results and Discussion

In this section we performed some experiments on each dataset described in the previous section. Initially, minimum supporting threshold is set for each experiment. At first step, the frequent k-item set is mined from each dataset. After this, association rules are generated from frequent k-itemset.

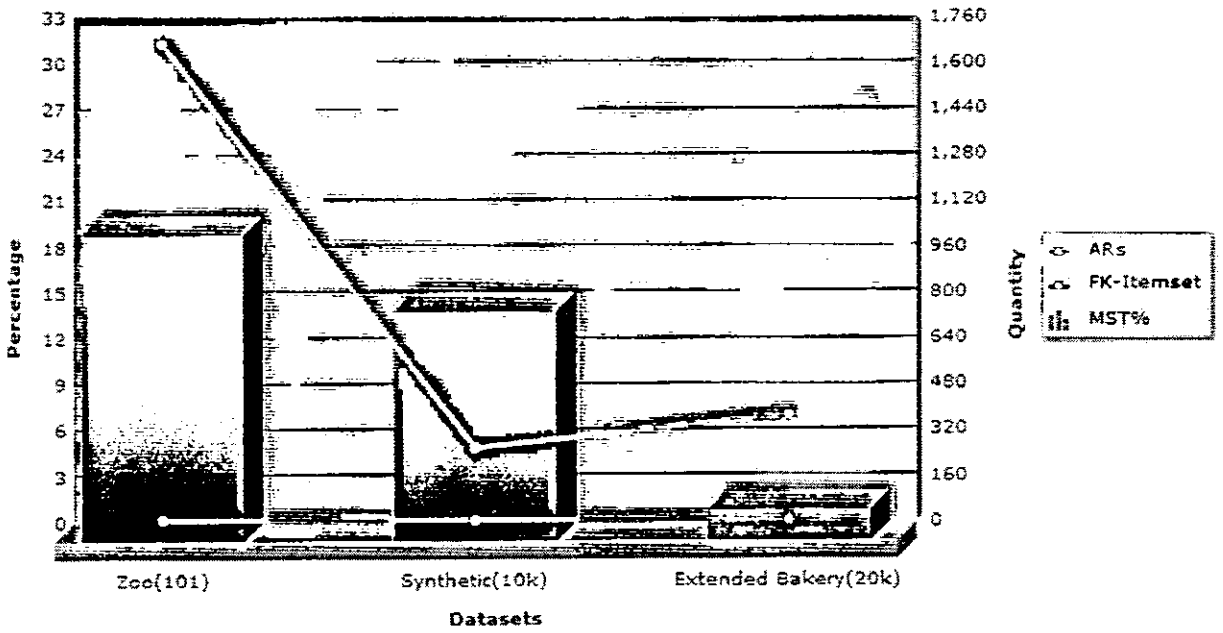

Figure 4.5: Frequent k-Itemset and their corresponding ARs

Figure 4.5 describes frequent k-Itemset (Fk-Itemset) and their corresponding Association Rules (ARs) with some Minimum Supporting Threshold (MST). The X-axis represents the size of different datasets and Y-axis in left hand side describes minimum supporting threshold for each dataset. While the Y-axis in right hand side indicates the number of frequent k-itemset and their corresponding association rules generated for each dataset. As the MST decreases the number of FK-Itemset and ARs will increases and vice versa. Consider that, some of these ARs leak out confidential information. We call it Sensitive Association Rules SARs or sensitive pattern. The SARs are randomly selected on the base of their support and confidence. Table 4.5 describes sensitive association rules for each dataset and their support and confidence.

Table 4.5: SARs their Support and Confidence

Dataset	|D|	SAR	Support %	Confidence %
Zoo	101	9→13	75	90
		9→8	61	74
		13,9→10	60	82
		9→10,13	60	73
Synthetic	10k	4→7	41	81
		6→4	41	73
		1→6	41	81
		7→4,6	35	70
Extended Bakery	20k	19→36	6	58
		34→43	5	54
		4,19→36	5	94

Three parameters play an important role in rule hiding process such as MST, MCT and the number of Transactions Modified in Each Generation (TMG) of PPGA. Therefore, if the values of these parameters are changed then the result will be changed. Moreover, we conducted several experiments on each dataset. The parameters were set for each experiment. Additionally, the hiding process loss some non sensitive patterns, called lost rules, and also new patterns are generated, called ghost rules. Thus, optimal sanitization is an NP-hard problem [6]. The PPGA preserve the privacy of restrictive patterns by decreasing the ghost rules to zero and lost rules to one in most of the cases.

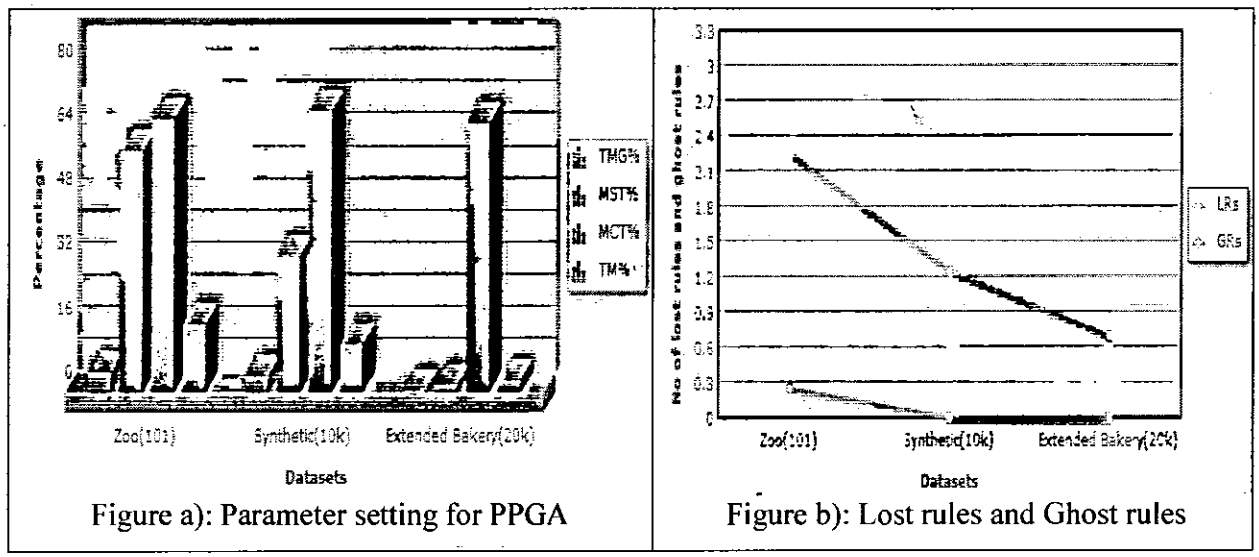

Figure a): Parameter setting for PPGA

Figure b): Lost rules and Ghost rules

Figure 4.6: Figure a and b depicts the experimental results of PPGA

Figure 4.6 illustrates the experimental results of PPGA. The X-axis describes the size of the different datasets and Y-axis of Figure a, indicates different parameter setting and the number of transactions modified during the hiding process. Similarly, Y-axis of Figure b, represents lost rules and ghost rules side effects. It describes that the proposed technique generates the lost rules between 0-3 and minimized the ghost rules side effects to zero. It also shows the number of transaction modified during the hiding process. The flow of the graph represents that as the size of database is increased the side effects in term of lost rules, ghost rules and transaction modification is decreased.

In the start of this chapter, the author presented four performance measures. On the base of experimental results, we claim that the proposed architecture hides sensitive association rules successfully with no hiding failure. Moreover, the approach used in this work minimizes the side effects in term of lost rules and ghost rules. Furthermore, minimizing the number of Transaction Modification (TM) is still remains. Additionally, one accidental measure which we have found is the CPU time. It is the amount of time taken by PPGA to preserve the privacy of confidential information. For small dataset it is no mater. But for large dataset the PPGA required huge amount of CPU time to preserve the privacy of association rules.

4.5 Comparison

The idea of using genetic algorithm for preserving the privacy of sensitive association rules was first introduced by Dehkordi et al. [30]. They performed experiments on example dataset which contain 5 transactions and 6 items in each record. They did not perform experiment on large

database. Therefore, we did not compare the proposed technique to their work. The proposed technique is compared to the technique presented by Naeem et al. [31], Varykios et al. [22] and Chih-Chia et al. [26]. Naeem et al. [31], proposed a novel architecture in the domain of Privacy Preserving Data Mining PPDM. They performed experiments on Zoo, Lymphography, Thyroid0387 and Hypothyroid datasets taken from UCI machine learning repository. The author claims that the techniques generate zero ghost rules. In addition, the technique causes high side effect in term of lost rules. Thus the experimental result of PPGA is compared to their work. This comparison is based on zoo dataset. Furthermore, the PPGA is compared to the technique proposed by Varykios et al. [22]. They introduced five algorithms namely algorithm 1.a, 1.b, 2.a, 2.b and 2.c in order to preserve the privacy of confidential information. They carried out experiments on databases of size 10k, 50k and 90k. These techniques generate high side effects in term of lost rules and ghost rules. Similarly, the proposed technique is judged against to these techniques. The judgment is based on synthetic dataset of size 10k. In the same direction, the proposed technique is compared to the technique presented by Chih-Chia et al. [26]. They proposed a novel architecture in the domain of PPDM, called FHSAR, for Fast Hiding Sensitive Association Rules. They conducted experiments on databases of size 10k, 20k, 30k, 50k and 100k. Each of them contains 50 data items, $|I|=50$. The experimental results of FHSAR describes that the technique hide sensitive association rules successfully. The experimental results of FHSAR shows that this is outperform in term of lost rules and ghost rules side effect then previous work done. The experimental results of PPGA are compared to the FHSAR. The comparison is made on the base of Extended Bakery dataset of size 20k. Moreover, the purpose of selecting zoo dataset of size 101 for Naeem et al. [31], synthetic dataset of size 10k for Varykios et al. [22] and extended bakery dataset of size 20k for Chih-Chia et al. [26] is to standardize the results of PPGA.

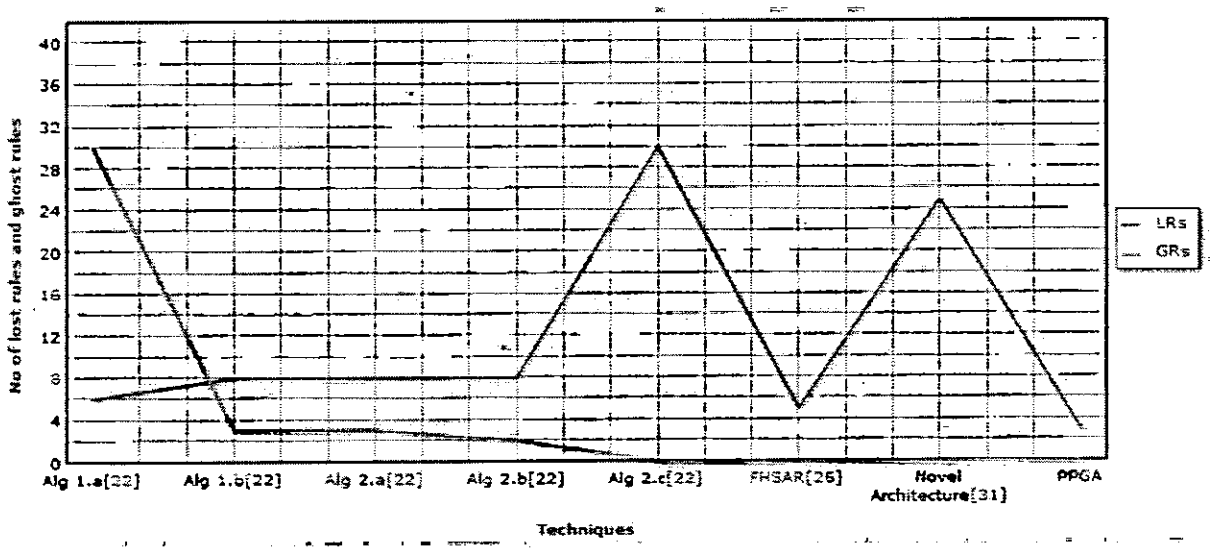

Figure 4.7: Comparison of PPGA with existing techniques average over three dataset zoo, synthetic and extended bakery dataset.

Figure 4.7 depicts the comparison of PPGA to other techniques in the literature as described. It shows that these techniques minimized the side effects in one direction such as minimized the ghost rules side effect and remain or ignore the lost rules side effect as well. It also describes that the proposed technique hide sensitive association rules by decreasing the ghost rules side effect to zero. The figure also represents that PPGA generate lost rules between 0-3. Similarly, the proposed technique hides sensitive association rules successfully with no hiding failure. On the base of such comparison we claim that PPGA is out perform than other techniques presented in the literature.

4.6 Summary

In this chapter, we used three datasets zoo [75], synthetic [76] and extended bakery dataset [77]. The experiments were conducted on these datasets. Moreover, the author claimed that the proposed technique minimizes lost rules and ghost rules side effects. Finally, the claim is validated by comparing the experimental results of PPGA to other techniques in the literature.

CHAPTER 5: CONCLUSION & FUTURE WORK

Chapter 5

CONCLUSION AND FUTURE WORK

5.1 Conclusion

Organizations such as Customer Relationship Management (CRM), telecommunication industry, financial sector investment trends, web technologies, demand and supply analysis, direct marketing, health industry, e-commerce, stocks & real estates, understanding consumer research marketing and product analysis often share data in order to achieve mutual benefits. However, sharing of data disclose confidential data. Therefore, data modification or data sanitization techniques are applied to preserve the confidentiality of their confidential data or restrictive pattern in the form of sensitive association rules. Moreover, it preserves the privacy of restrictive patterns by concealing the frequent itemsets subsequent to those patterns. This process overcomes the leak out of confidential information while sharing data. It causes impact on data effectiveness in the form of non-restrictive patterns lost and also new patterns are generated. The problem of optimal sanitization is very complex or NP-hard [6]. In current research work, we minimized the side effects caused by hiding sensitive association rules or frequent itemset. Furthermore, we presented a fitness function. It calculates fitness value of each transaction. Moreover, this approach hides sensitive patterns or sensitive association rules successfully. Additionally, the hiding process modifies some transaction in original dataset. Here binary dataset is passed as initial population to Privacy Preserving Genetic Algorithm PPGA. Similarly, the PPGA modifies the database recursively until the support or confidence of the restrictive patterns drop below the user specified threshold. This process takes CPU time to complete. It is the amount of time taken by PPGA to hide sensitive association rules. It depends on dataset. For small dataset the process complete in short time. If the dataset is large then the prototype runs in huge amount of CPU time. Additionally, to test and validate the PPGA experiments were performed on Zoo dataset [75], Synthetic dataset [76] and Extended Bakery dataset [77]. Similarly, the experimental results of PPGA compared to the technique presented by Naeem et al. [31], Varykios et al. [22] and Chih-Chia et al. [26]. Thus, the claim is verified that PPGA outperform then other techniques available in the literature. Furthermore, the technique presented in this approach generates the lost rules 0-3 and minimized ghost rule to zero.

5.2 Future Work

In the future, we will design a confidence base privacy preserving genetic algorithm PPGA. It will improve the existing fitness function of PPGA. Moreover, it will modify those items in a sensitive transaction that will reduce the confidence of the rule. Hence, this will minimize the number of transaction modification and also ensure to minimize lost rule and ghost rule side effects. The PPGA takes huge amount of CPU time to preserve the privacy of confidential information. Therefore, we will also try to improve the CPU time of privacy preserving genetic algorithm PPGA. Additionally, we will apply other evolutionary approaches, to preserve the privacy of sensitive association rules.

References:

[1] R. Agrawal and R. Srikant, "Privacy preserving data mining", In ACM SIGMOD Conference on Management of Data, pages 439-450, Dallas, Texas, May 2000

[2] L. Brankovic and V. Estivill-Castro, "Privacy Issues in Knowledge Discovery and Data Mining", AustralianInstitute of Computer Ethics Conference, July, 1999, Lilydale.

[3] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining", in SIGMOD Workshop on Research Issues on Data Mining and knowledge Discovery, 1996

[4] D. Agrawal and C. C. Aggarwal, "On the design and quantification of privacy preserving data mining algorithms", In Proceedings of the 20th Symposium on Principles of Database Systems, Santa Barbara, California, USA, May 2001.

[5] C. Clifton, "Protecting Against Data Mining Through Samples", in Proceedings of the Thirteenth Annual IFIP WG 11.3Working Conference on Database Security, 1999

[6] M. Atallah, E. Bertino, A. Elmagarmid, M. Ibrahim, and Verykios V. S., "Disclosure Limitation of Sensitive Rules," in Proc. 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), pp 45–52, 1999

[7] Y. Lindell and B. Pinkas, "Privacy preserving data mining", In CRYPTO, pages 36–54, 2000

[8] C. Clifton, "Using Sample Size to Limit Exposure to Data Mining", Journal of Computer Security, 8(4), 2000.

[9] C. Clifton, M. Kantarcioglu, X. Lin and M.Y. Zhu, "Tools for Privacy Preserving Distributed DataMining", SIGKDD Explorations, 4(2), 1-7, Dec. 2002.

[10] A. Evfimievski, R. Srikant, R. Agrawal, and J. Gehrke,"Privacy preserving mining of association rules", In Proc, Of the 8th ACM SIGKDD Int'l Conference on Knowledge Discoveryand Data Mining, Edmonton, Canada, July 2002

[11] D.E.O. Leary, "Knowledge Discovery as a Threat to Database Security", In G. Piatetsky-Shapiro and W. J. Frawley,editors, Knowledge Discovery in Databases, 507-516, AAAIPress/ MIT Press, Menlo Park, CA, 1991

[12] V. Verykios, E. Bertino, I.G. Fovino, L.P. Provenza, Y.Saygin, and Y. Theodoridis, "State-of-the-art in Privacy Preserving Data Mining", SIGMOD Record, Vol. 33, No. 1, 50-57,March 2004.

[13] S.L. Wang, and A. Jafari, "Using Unknowns for Hiding Sensitive Predictive Association Rrules, " In Proceedings of the 2005 IEEE International Conference on Information Reuse and Integration (IRI 2005), pp.223–228, 2005

[14] E. Dasseni, V. Verykios, A. Elmagarmid and E. Bertino, "Hiding Association Rules by Using Confidence and Support" in Proceedings of 4th Information Hiding Workshop, 369-383,Pittsburgh, PA, April 2001.

[15] A. Evfimievski, "Randomization in Privacy Preserving Data Mining", SIGKDD Explorations, 4(2), Issue 2,43-48, Dec. 2002.

[16] A. Evfimievski, J. Gehrke and R. Srikant, "Limiting Privacy Breaches in Privacy Preserving DataMining", PODS 2003, June 9-12, 2003, San Diego, CA.

[17] S.R. Oliveira and O.R. Zaiane, "Privacy Preserving Frequent Itemset Mining", Proceedings of IEEE International Conference on Data Mining, December 2002, pp. 43-54.

[18] S.R. Oliveira and O.R. Zaiane, "Algorithms for Balancing Priacy and Knowledge Discovery in Association Rule Mining",Proceedings of 7th International Database Engineering and Applications Symposium (IDEAS03), Hong Kong, July 2003, pp.54-63.

[19] M. Kantarcioglu and C. Clifton, "Privacy-preserving distributed mining of association rules on horizontally partitioned data", In ACM SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery, June 2002

[20] S.R. Oliveira and O.R. Zaiane, "Protecting Sensitive Knowledge by Data Sanitization", Proceedings of IEEE International Conference on Data Mining, November 2003, pp. 613-616.

[21] J. Vaidya and C.W. Clifton. "Privacy preserving association rule mining in vertically partitioned data", In Proc. of the 8th ACM SIGKDD Int'l Conference on Knowledge Discovery and Data Mining, Edmonton, Canada, July 2002

[22] V. Verykios, A. Elmagarmid, E. Bertino, Y. Saygin, and E.Dasseni, "Association Rules Hiding", IEEE Transactions on Knowledge and Data Engineering, Vol. 16, No. 4, 434-447, April 2004

[23] Y. Saygin, V. Verykios, and C. Clifton, "Using Unknowns to Prevent Discovery of Association Rules", SIGMOD Record 30(4): 45-54, December 2001

[24] K. Duraiswamy, D. Manjula, and N. Maheswari, "A New approach to Sensitive Rule Hiding", Journal of Computer and Information science, 2008

[25] M. Gupta et al., "Privacy Preserving Fuzzy Association Rules Hiding in Quantitative Data", International Journal of Computer Theory & Engineering, Vol. No.4 Oct 2009

[26] Chih-Chia Weng, C. Shan-Tai, and L. Hung-Che, "A Novel Algorithm for Completely Hiding Sensitive Association Rules," in Proc. Eighth International Conference on Intelligent Systems Design and Applications Taiwan, 2008

[27] S.L. Wang, and A.Jafari, "Hiding Sensitive Predictive Association Rules", Systems, Man and Cybernetics, 2005 IEEE International Conference on , vol.1, no., pp. 164- 169 Vol. 1, 10-12 Oct. 2005

[28] R. Agarwal, T. Imielinski, and A. Swami, "Mining Associations Between Sets of Items in Large Databases", SIGMOD93, pages 207-216, Washington, D.C, USA, May 1993

[29] M. Sulaiman Khan, M. Muyeba and F. Coenen, "Weighted Association Rule Mining from Binary and Fuzzy Data", Proceedings of the 8th industrial conference on Advances in Data Mining, pp.200-212, 2008

[30] M. Naderi Dehkordi, K. Badie, and A. K. Zadeh, "A Novel Method for Privacy Preserving in Association Rule Mining Bassed on Genetic Algorithms", Journal of Software, Vol. 4, No. 6, August 2009

[31] M. Naeem, S. Asghar and S. Fong, "Hiding sensitive association rules using central tendency," Advanced Information Management and Service (IMS), 6th International Conference on , vol., no., pp.478-484, Dec. 2010

[32] C.B. Ramesh, V. Jitendra, A.K. Sohel, S. Anand, "Hiding Sensitive Association Rules Efficiently By Introducing New Variables Hiding Counter", IEEE International Conference on Service Operations and Logistics, and Informatics, Beijing , 2008.

[33] G. Krishna and V. P. R. Krishna, "A Novel Approach for Statistical and Fuzzy Association Rule Mining on Quantitative Data," Journal of Scientific and Industrial Research, Vol. 67 July 2008, pp. 512-517

[34] C.N. Modi, U.P. Rao, D.R. Patel, "Maintaining privacy and data quality in privacy preserving association rule mining," Computing Communication and Networking Technologies (ICCCNT), 2010 International Conference on , vol., no., pp.1-6, 29-31 July 2010

[35] S.T. Chen, S.M. Lin, C.Y. Tang, and G.Y. Lin, "An Improved Algorithm for Completely Hiding Sensitive Association Rule Sets," Computer Science and its Applications, CSA '09. 2nd International Conference on, vol., no., pp.1-6, 10-12 Dec. 2009

[36] Y.C. Chang, and S.T. Chen, "Fast Algorithm for Completely Hiding Sensitive Association Rule Sets," Proceedings of the Cryptology and Information Security Conference (CISC2008), Hualien, Taiwan, R.O.C., pp. 547-560, 2008.

[37] S.R.M. Oliveira, O.R. Zaïane, "Privacy Preserving Frequent Itemset Mining," Proceedings of the IEEE international conference on Privacy, security and data mining - Volume 14, Austrlia, 2002

[38] X. Zhang, X. Qiao, "New Approach for Sensitive Association Rule Hiding," International Workshop on Education Technology and Training & International Workshop on Geosciences and Remote Sensing, vol. 2, pp.710-714, , 2008

[39] R. Agarwal, T. Imielinski and A. Swami, "Mining associations between sets of items in large databases," SIGMOD93, pages 207-216, Washington, D.C, USA, May 1993

[40] C. Clifton and D. Marks, "Security and Privacy Implications of Data Mining," In Proc. ACM Workshop Research Issues in Data Mining and Knowledge Discovery, 1996.

[41] H.H. Malik, J.R. Kender, "Clustering Web Images using Association Rules, Interestingness Measures, and Hypergraph Partitions," ICWE'06, July 11–14, 2006, Palo Alto, California, USA. ACM 1-59593-352-2/06/0007

[42] 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, 19(1), 2007.

[43] W. Du, M.J. Atallah, "Secure Multi-Party Computation Problems and their Applications: A Review and Open Problems," In Proc. of 10th ACM/SIGSAC 2001 New Security Paradigms Workshop, pages 13-22, Cloudcroft, New Mexico, September 2001

[44] O. Goldreich, S. Micali, A. Wigderson, "How to Play Any Mental Game - A Completeness Theorem for Protocols with Honest Majority," In Proc. of the 19th Annual ACM Symposium on Theory of Computing, pages 218-229, New York City, USA, May 1987

[45] B. Pinkas, "Cryptographic Techniques for Privacy-Preserving Data Mining," SIGKDD Explorations, 4(2):12-19, December 2002.

[46] L. Chang and I. Moskowitz, "An integrated framwork for database inference and privacy protection," Data and Applications Security. Kluwer, 2000

[47] L. Yongcheng et al., "A Survey on the Privacy Preserving Algorithm of Association Rule Mining", 2009 Second International Symposium on Electronic Commerce and Security, 2009

[48] Y. Saygin, Vassilios S. Verykios, and Ahmed K. Elmagarmid, "Privacy preserving association rule mining," In Proceedings of the 12th International Workshop on Research Issues in Data Engineering , pp.151–158, February 2002.

[49] G. Yuhong, "Reconstruction-Based Association Rule Hiding", in Proc. SIGMOD2007 Ph.D. Workshop on Innovative Database Research 2007(IDAR2007), June 10, 2007, Beijing, China.

[50] A. Gkoulalas-Divanis, V.S. Verykios, "An Integer Programming Approach for Frequent Itemset Hiding," In Proc. ACM Conf. Information and Knowledge Management (CIKM 06), Nov. 2006.

[51] Ali Amiri, "Dare to share: Protecting sensitive knowledge with data sanitization," Decision Support Systems, 43(1): pp.181–191, 2007.

[52] Estivill-Castro, Brankovic. "Data Swapping: Balancing Privacy against Precision in Mining for Logic Rules," Data Warehousing and Knowledge Discovery (D*a*Wak'99*)*, M. Mukesh and A. M. Tjoa (Eds.), Springer-Verlag, Berlin, 1999, pp. 389-398.

[53] X. Sun and P.S. Yu, "A border-based approach for hiding sensitive frequent itemsets" In: Proc. of the 5th IEEE Int'l Conf. on Data Mining (ICDM'05). IEEE Computer Society, 2005. 426-433.

[54] S.R.M. Oliveira, O.R. Zaiane, "Toward standardization in privacy preserving data mining", CiteSeer, In Proc. Of the 3rd workshop of Data Mining Standards (DM-SSP 2004), in conjunction with KDD 2004.

[55] X. Sun, P.S. Yu, "Hiding Sensitive Frequent Itemsets by a Border-Based Approach," *Computing Science and Eng.,* vol. 1, no. 1, pp. 74-94, 2007.

[56] H. Mannila, H. Toivonen, "Level wise search and border of theories in knowledge discovery," Data Mining and Knowledge Discovery, 1(3):241-258, 1997.

[57] S. Menon, S. Sarkar, S. Mukherjee, "Maximizing accuracy of shared databases when concealing sensitive patterns," Information Systems Research, 16(3):256–270, 2005.

[58] A. A. Veloso, W. Meira Jr., S. Parthasarathy, and M. B. Carvalho, "Efficient, Accurate and Privacy-Preserving Data Mining for Frequent Itemsets in Distributed Databases," In Proc. of the 18th Brazilian Symposium on Databases, pages 281-292, Manaus, Brazil, October 2003.

[59] S. Meregu and J. Ghosh, "Privacy-Preserving Distributed Clustering Using Generative Models," In Proc. of the 3rd IEEE International Conference on Data Mining (ICDM'03), pages 211-218, Melbourne, Florida, USA, November 2003.

[60] C.C. Aggarwal, Yu, "Privacy-Preserving Data Mining: Models and Algorithms," Springer, 2008.

[61] S.L. Wang, B. Parikh, and A. Jafari, "Hiding Informative Association Rule Sets," Expert Systems with Application, vol. 33, pp. 316-323, Aug. 2007.

[62] S. R. M. Oliveira and O. R. Zaiane, "Achieving Privacy Preservation When Sharing Data For Clustering," In Proc. of the Workshop on Secure Data Management in a Connected World (SDM'04) in conjunction with VLDB'2004, pages 67-82, Toronto, Ontario, Canada, Aug. 2004

[63] S. R. M. Oliveira and O. R. Zaiane, "Privacy-Preserving Clustering by Object Similarity-Based Representation and Dimensionality Reduction Transformation," In Proc. of the Workshop on Privacy and Security Aspects of Data Mining (PSADM'04) in conjunction with the Fourth IEEE International Conference on Data Mining (ICDM'04), pages 21-30, Brighton, UK, November 2004.

[64] J. Han and M. Kamber, "Data Mining: Concepts and Techniques," Morgan Kaufmann Publishers, San Francisco, CA, 2001.

[65] T. Johnsten and V. V. Raghavan, "Impact of Decision-Region Based Classification Mining Algorithms on Database Security," In Proc. of 13th Annual IFIP WG 11.3 Working Conference on Database Security, pages 177-191, Seattle, USA, July 1999.

[66] T. Johnsten and V. V. Raghavan, "Security Procedures for Classification Mining Algorithms," In Proc. of 15th Annual IFIP WG 11.3 Working Conference on Database and Applications Security, pages 293-309, Niagara on the Lake, Ontario, Canada, July 2001.

[67] T. Johnsten and V. V. Raghavan, "A Methodology for Hiding Knowledge in Databases," In Proc. of the IEEE ICDM Workshop on Privacy, Security, and Data Mining, pages 9-17, Maebashi City, Japan, December 2002.

[68] S. R. M. Oliveira, O. R. Zaiane, and Y. Saygin, "Secure Association Rule Sharing," In Proc. of the 8th Paci_c-Asia Conference on Knowledge Discovery and Data Mining (PAKDD'04), pages 74-85, Sydney, Australia, May 2004.

[69] J. Kennedy and R. Eberhart, "Particle swarm optimization," In Proc. IEEE International Conference on Neural Networks, Nov/Dec 1995

[70] M. Dorigo, V. Maniezzo, and A. Colorni, "Ant System: Optimization by a colony of cooperating agents," IEEE Transactions on Systems, Man, and Cybernetics—Part B, vol. 26, no. 1, pp. 29–41, 1996.

[71] S. Kirkpatrick, C. D. Gelatt, Jr., M. P. Vecchi, "Optimization by Simulated Annealing," 13 May 1983.

[72] F. Glover, "Tabu Search — Part I," ORSA Journal on Computing, Vol. 1, No. 3, pp. 190 206, 1989.

[73] J. Holland, " Genetic Algorithm," Scientific American, July 1992

[74] A. Afshar, O. Bozorg Haddad, M.A. Mariño, B.J. Adams, "Honey-bee mating optimization (HBMO) algorithm for optimal reservoir operation," Journal of the Franklin Institute, Volume 344, Issue 5, August 2007

[75] A. Frank, A. Asuncion,2010 "{UCI} Machine Learning Repository", University of California, Irvine, School of Information and Computer Sciences, Available: http://archive.ics.uci.edu/ml 2010. [Accessed 02-03-2012].

[76] H. Hamilton, "DBD: Data Mining Projects", University of Regina Available: http://www2.cs.uregina.ca/~dbd/cs831/index.html, 2000-9. [Accessed 15-03-2012].

[77] Track Open Source Project, "Extended Bakery Dataset", Integrated SCM & Project Management, Available: https://wiki.csc.calpoly.edu/datasets/wiki/ExtendedBakery20k, 2003. [Accessed 02-03-2012].