CHAPTER 1

INTRODUCTION

Data mining is widely utilized in the medical diagnosis field to extract useful information from patient data. Today, individual data can be inferred by humans as well as data mining algorithms maintained by companies and hospitals, leading to major privacy threats to individuals. In recent years, numerous algorithms have been proposed for modifying or transforming data so as to preserve privacy. The data processed by privacy preserving algorithms becomes less efficient compared to the original data. In order to retain the fidelity of the data and also apply adequate privacy, an efficient balanced algorithm has to be developed.

1.1 DATA MINING

Data Mining (Divya et al 2013) is an important and inspiring area of research for extracting meaningful information from huge data sets. Data Mining is becoming popular today in healthcare because there is a need for efficient analytical methodology for detecting unknown and valuable information from health data. Data Mining provides several benefits in healthcare such as fraud detection in health insurance, providing medical solutions to patients at lower cost, detection of causes of diseases and identification of medical treatment methods. The data generated by health organizations (details regarding hospitals, patients, medical claims, treatment costs, etc.) is difficult to analyze (as it is very vast and complex) for making important decision regarding patient health. So, there is a need to generate a
powerful tool for analyzing and extracting important information from this complex data. The analysis of health data improves healthcare by enhancing the performance of patient management tasks. The outcome of Data Mining technologies is to provide benefits to healthcare organizations for grouping patients having similar diseases or health issues so that healthcare organizations provide effective treatments. It can also be useful for predicting the length of stay of patients in hospital, for medical diagnosis and planning for effective information system management.

Recent technologies are being used in medical field to enhance the medical services in cost effective manner. Data Mining techniques are also used to analyze various factors that are responsible for diseases, e.g. type of food, different working environments, education levels, living conditions, availability of potable water, health care services, cultural, environmental and agricultural factors. Data Mining Algorithms (Aura Conci et al 2002) have a huge potential to create a knowledge-rich environment to significantly improve the quality of clinical decisions.

1.2 PRIVACY PRESERVING ALGORITHMS

Various techniques available for the privacy preserving data mining include randomization, \(k\)-anonymization, and distributed privacy-preserving data mining.

The randomization method: In the randomization method for privacy-preserving data mining, noise is added to the data in order to mask the attribute values of records (Agarwal 2002). The noise added is sufficiently large so that individual record values cannot be recovered. Therefore techniques are designed to derive aggregate distributions from the perturbed records. Subsequently, data mining techniques can be developed to extract useful information from these aggregate distributions.
The k-anonymity model and l-diversity: These avoid the possibility of indirect identification of records from public databases. This is because combinations of record attributes can be used to exactly identify individual records. In the k-anonymity method, the granularity of data representation is reduced by using techniques such as generalization and suppression. This granularity is reduced sufficiently so that any given record maps on to at least k other records in the data. The l-diversity model (Gaoming Yang et al 2013) was designed to handle some weaknesses in the k-anonymity model since protecting identities to the level of k-individuals is not the same as protecting the corresponding sensitive values, especially when there is homogeneity of sensitive values within a group. To do so, the concept of intra-group diversity of sensitive values is promoted within the anonymization scheme (Machanavajjhala et al 2006).

1.2.1 Distributed Privacy Preservation

Often, users may wish to derive aggregate results from data sets which are partitioned across entities. Such partitioning may be horizontal (when the records are distributed across multiple entities) or vertical (when the attributes are distributed across multiple entities). While the data owners may prefer not to share their entire data sets, they may consent to limited information sharing after applying a variety of protocols. The overall desired result is to maintain privacy for each individual entity, while deriving satisfactory aggregate results over the entire data.

1.3 TEMPORAL DATA MINING

The main aim of temporal data mining (Claudia et al 2001) is to identify or extract relations between sequences and sub-sequences of events. A sequence composed by a series of nominal symbols from a particular alphabet is usually called a temporal sequence and a sequence of continuous,
real-valued elements is known as time series. The field of temporal data mining is concerned with such analyses in the case of ordered data streams with temporal interdependencies studied by Srivatsan et al (2006).

1.3.1 Mining Sequencing Patterns

Sequential Pattern Mining (Jian Pei et al 2001) plays a major role in discovering the patterns in the time series data. It extracts all possible sub-sequence in the pattern. The sequential pattern mining (Goryczka et al 2013) problem was first introduced by Agrawal & Srikant (1995). Given a set of sequences, where each sequence consists of a list of elements and each element consists of a set of items, and given a user-specified min support threshold, sequential pattern mining is to find all of the frequent subsequences, i.e. the sub-sequences whose occurrence frequency in the set of sequences is not less than minimum support.

1.3.2 Prefix - Projected Sequential Pattern Mining

In their study, Agrawal & Srikant (1993) developed a novel sequential pattern mining method called PrefixSpan (i.e. Prefix-projected Sequential pattern mining). Its general idea is to examine only the prefix sub-sequences and project only their corresponding postfix sub-sequences into projected databases. In each projected database, sequential patterns are grown by exploring only local frequent patterns. To further improve mining efficiency, two kinds of database projections are explored: level-by-level projection and bi-level projection. Moreover, a main-memory-based pseudo-projection technique is developed for saving projection cost and speeding up processing when the projected (sub)-database and its associated pseudo-projection processing structure can fit in main memory.
1.3.2.1 Sequential database with example

A sequence database S is a set of tuples \( \langle \text{sid}.s \rangle \), where \( \text{sid} \) is a **sequence_id** and \( s \) is a sequence. A tuple \( \langle \text{sid}.s \rangle \) is said to contain a sequence \( \alpha \), if \( \alpha \) is a sub-sequence of \( s \), i.e. \( \alpha \subseteq s \). The support of a sequence \( \alpha \) in a sequence database S is the number of tuples in the database containing \( \alpha \), i.e. \( \text{support}_\alpha = \left| \{ \langle \text{sid}.s \rangle \mid (\{ \langle \text{sid}.s \rangle \in S \}\land (\alpha \subseteq s)) \} \right| \). It can be denoted as \( \text{support} \alpha \) if the sequence database is clear from the context. Given a positive integer as the support threshold, a sequence is called a (frequent) sequential pattern in sequence database S if the sequence is contained by at least \( \xi \) tuples in the database, i.e. \( \text{support}_\alpha (\alpha) \geq \xi \). A sequential pattern with length \( l \) is called an \( l \)-pattern.

**Example (Running example)** Let the running database be the sequence database given in Table 1.1 and min support = 2. The set of items in the database is \{a, b, c, d, e, f, g\}

<table>
<thead>
<tr>
<th>Sequence_id</th>
<th>Sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>( \langle a(abc)(ac)d(cf) \rangle )</td>
</tr>
<tr>
<td>20</td>
<td>( \langle (ad)c(bc)(ae) \rangle )</td>
</tr>
<tr>
<td>30</td>
<td>( \langle (ef)(ac)(df)cb \rangle )</td>
</tr>
<tr>
<td>40</td>
<td>( \langle eg(af)cbc \rangle )</td>
</tr>
</tbody>
</table>

A sequence \( \langle a(abc)(ac)d(cf) \rangle \) has five elements: (a), (abc),(ac), (d) and (cf), where items \( \alpha \) and c appear more than once respectively in different elements. It is also a 9-sequence since there are 9 instances
appearing in it. Item \( \alpha \) happens 3 times in this sequence, so it contributes 3 to the \textit{length} of the sequence. However, the whole sequence \( \langle a(abc)(ac)d(cf) \rangle \) contributes only 1 to the \textit{support} of \( (a) \). Also, sequence \( (a(bc)df) \) is a \textit{sub-sequence} of \( \langle a(abc)(ac)d(cf) \rangle \). Since both sequences 10 and 30 contain sub-sequence \( s = \langle (ab)c \rangle \), \( s \) is a \textit{sequential pattern} of length 3 (i.e. 3-pattern).

1.3.2.2 Mining sequential patterns by prefix projections: An example

Since items within an element of a sequence can be listed in any order without loss of generality, it is assumed they are listed in alphabetical order. For example, the sequence with Sequence_id 10 in the running example is listed as \( \langle a(abc)(ac)d(cf) \rangle \) instead of \( \langle a(abc)(ac)d(fc) \rangle \). With such a convention, the expression of a sequence is unique.

\textbf{Definition: (Prefix, projection)} Suppose all the items in an element are listed alphabetically. Given a sequence \( \alpha = \langle e_1 e_2 \ldots e_n \rangle \), a sequence \( \beta = \langle e'_1 e'_2 \ldots e'_m \rangle (m \leq n) \) is called a \textbf{prefix} of \( \alpha \) if and only if (1) \( e'_i = e_i \) for \( (I \leq m - 1) \); (2) \( e'_m \subseteq e_m \); and (3) all the items in \( e_m - e'_m \) are alphabetically after those in \( e'_m \).

Given sequences \( \alpha \) and \( \beta \) such that \( \beta \) is a sub-sequence of \( \alpha \), i.e. \( \beta \subseteq \alpha \). A sub-sequence \( \alpha' \) of sequence \( \alpha \) (i.e. \( \alpha' \subseteq \alpha \) ) is called a \textbf{projection} of \( \alpha \) w.r.t. prefix \( \beta \) if and only if (1) \( \alpha' \) has prefix \( \beta \) and (2) there exists no proper super-sequence \( \alpha'' \) of \( \alpha' \) (i.e. \( \alpha' \subseteq \alpha'' \) but \( \alpha' \neq \alpha'' \)) such that \( \alpha'' \) is a sub-sequence of \( \alpha \) and also has prefix \( \beta \). For example, \( \langle a \rangle , \langle aa \rangle , \langle a(ab) \rangle \) and \( \langle a(abc) \rangle \) are prefixes of sequence \( \langle a(abc)(ac)d(cf) \rangle \), but neither \( \langle ab \rangle \) nor \( \langle a(bc) \rangle \) is considered as a prefix.
Prefix span Algorithm steps

**Step 1**: **Find length-1 sequential patterns.** Scan S once to find all frequent items in sequences. Each of these frequent items is a length-1 sequential pattern.

**Step 2**: **Divide search space.** The complete set of sequential patterns can be partitioned into n subsets according to n (number of prefixes).

**Step 3**: **Find subsets of sequential patterns.** The sub-sets of sequential patterns can be mined by constructing corresponding projected databases and mining each recursively.

### 1.4 ASSOCIATION RULE HIDING

Hiding both data and knowledge is an important issue in privacy of data (Vassilios et al 2008), privacy can be maintained before and after mining. Association rule hiding performs the process of hiding or changing the original database into the modified database, by which certain association rules are hidden.

Association Rule Mining (Agarwal et al 1993) is the process of extracting sets of items that frequently co-occur in a transactional database so as to produce association rules that hold for the data. Each association rule is defined as an implication of the form \( A \Rightarrow B \), where \( A, B \) are frequent item sets in the transactional database, such that \( A \cap B = \emptyset \). The item set \( A \cup B \) that leads to the generation of an association rule is known as the generating item set and consists of two parts: the Left Hand Side (LHS), which is the part on the left of the arrow of the rule, and the Right Hand Side (RHS), which is the part on the right of the arrow of the rule. Two metrics, support and
confidence, are incorporated in the task of association rule mining to drive the generation of association rules and expose only those that are expected to be interesting to the owner (Hospital) of the data.

Knowledge hiding, in the context of association rule mining (Guoling et al 2009), aims at sanitizing the original dataset in a way that at least one of the following goals is accomplished:

i) No rule that is considered as sensitive from the owner’s perspective, can be revealed from the sanitized dataset when this is mined at pre-specified thresholds of confidence and support (or at any value higher than these thresholds),

ii) All the non-sensitive rules can be successfully mined in the sanitized database at pre-specified thresholds of confidence and support (or higher), and

iii) No rule that was not found in the original dataset can be found at the sanitized database when mining this database at pre-specified thresholds of confidence and support (or higher).

The first goal requires sensitive rules to disappear. The second goal simply states that there should be no lost rules in the sanitized dataset. The third goal says that no false rules should be produced as a side-effect of the sanitization process. Generally in the typical case-hiding scenario, the sanitization process has to be accomplished in a way that minimally affects the original dataset, preserves the general patterns and trends, and achieves to conceal the sensitive knowledge. To be able to classify the various algorithms, propose a set of orthogonal dimensions based on which, the algorithm will present the following approaches:
i) The first dimension considers whether the hiding algorithm uses the support or the confidence of the rule to drive the hiding process. By this, the hiding algorithms are separated into support-based and confidence-based.

ii) The second dimension in the classification is related to the modification in the raw data caused by the hiding algorithm. The two forms of the modification are the distortion and the blocking of the original values. Distortion is the process of replacing 1’s by 0’s and 0’s by 1’s, while blocking refers to replacing original values by question marks.

iii) The third dimension refers to whether a single rule or set of rules can be hidden during an iteration of the hiding algorithms. This criterion differentiates hiding algorithms into single rule and multiple rule schemes.

iv) The fourth dimension has to do with the nature of the hiding algorithm, which can be either heuristic or exact. Heuristic techniques rely on optimizing certain sub-goals in the hiding process, while they do not guarantee optimality. The formulation of the association rule hiding problem presented implies that there are two specific sub-goals that need to be attained by every association rule hiding algorithm.

The first and most important sub-goal is to try to hide as many sensitive rules as possible. The second sub-goal is to manage to hide the sensitive rules while minimizing the possible side-effects.
As side-effects in the hiding process, consider:

a) The number of data items affected by the hiding process,

b) The number of non-sensitive rules which are accidentally hidden during the hiding process, and

c) The number of rules which recreated by the hiding process. Different hiding algorithms give different priorities to the satisfaction of the sub-goals presented, thus producing a list with primitive hiding of hiding primitively. Exact techniques, on the other hand, rely on formulating the association rule hiding problem in such a way that a solution can be found to satisfy all the sub-goals.

1.4.1 Approaches in Association Rule Hiding

Association rule hiding algorithms can be divided into three distinct classes: heuristic approaches, border-based approaches and exact approaches.

1.4.1.1 Heuristic approach

The first approach involves efficient fast algorithms that selectively sanitize a set of transactions from the database to hide the sensitive knowledge. Due to their efficiency and scalability, heuristic approaches have been the focus of attention for the vast majority of researchers in the knowledge hiding field.

1.4.1.2 Border-Based Approach

The second approach considers the task of sensitive rule hiding through modification of the original borders in the lattice of frequent and infrequent patterns in the dataset. In these schemes, the sensitive knowledge is
hidden by enforcing the revised borders in the sanitized database. The algorithms in this class differ both in (a) the borders they track and use for the hiding strategy and (b) the methodology they follow to enforce the revised borders in the sanitized dataset.

1.4.1.3 Exact approach

The third approach contains non-heuristic algorithms which conceive the hiding process as a constraint satisfaction problem that is solved by using integer or linear programming. The main difference of this approach compared to the previous ones is that the sanitization process guarantees optimality in the hiding solution, provided that an optimal solution exists.

1.5 MOTIVATION OF THE THESIS

The output of the typical data mining algorithms does not result in privacy preservation. In the medical data set, the privacy of an individual is more important than the output of the extracted knowledge, so the need of an efficient privacy preserving is needed either before mining the data or privacy preservation in the extraction of the rules. Various privacy preserving techniques are available to protect or reveal the data of an individual. In the pre-mining method, the privacy preservation is done in data bases itself. When the sensitive information is hidden or removed in the data bases (Arumugam et al 2013), the sanitized file (extracted information) usually loses authenticity because the sensitive knowledge is also removed or hidden.

Another procedure of privacy preservation can be deployed in the extraction of the rules. Association rule hiding is one of the techniques to hide the sensitive rules in the database after extracting rules. This technique is applied in the post mining process. Comparing the privacy preservation either in pre- or post-mining, the result of the extracted knowledge is less and the
information loss is higher. Hence an effective algorithm needs to developed to increase the scope of the knowledge discovery and reduce the information loss.

The above limitations have motivated researchers to seek an efficient algorithm for pre- and post-mining with balance constraint measure on temporal datasets (privacy preserving in medical data sets).

1.6 CONTRIBUTION OF THE THESIS

The proposed work contributes:

- An efficient algorithm for privacy which also preserves temporal pattern mining
- Balanced Constraint Measure Algorithm to preserve privacy from sequential rule discovery
- Balanced Constraint Measure Based Algorithm on pre-mining for privacy preserved data publishing

1.6.1 An Efficient Algorithm for Privacy Preserving Temporal Pattern Mining

Numerous research works show that data mining generally deals with the extraction of potentially useful information from large collections of data for a variety of functional areas, including customer relationship management, market basket analysis and bio-informatics. Data mining is also used for predicting and analyzing the medical records of hospitals, e.g. potential outbreaks of infectious diseases, analysis of customer transactions for market research applications, etc.
The proposed work protects the privacy of the medical database stored by hospitals. Hospitals may share their medical database with other organizations, including drug manufacturers, medical insurance companies and other hospitals. But privacy needs to be preserved on the patient medical database because patients submit personal data only for proper and correct medical treatment. To ensure privacy of the original data, it is stipulated that the privacy level should be specified by the data owner (hospital) to adequately protect the privacy of patients’ medical data while sharing it with other organizations.

Using the proposed approach, the data owner (Hospital) can select which part(s) of the patient database are to be anonymized (made private, masked or hidden). Here, it is assumed that patient name, disease name and patient location are to be anonymized by using ASCII code. Diseases are classified into four types: sensitive, frequent, seasonal and geographical.

Sensitive diseases take longer to be treated and are classified by user-defined sensitive threshold values. The threshold value is the user-defined minimum time duration of the disease. Sensitive diseases are extracted from the database using the user-defined count threshold value $C_s t$. After identifying a sensitive disease, it is anonymized by modifying its count to equal that of normal diseases by repeatedly decreasing the sensitive disease count value by $C_s t$ until its count value is less than or equal to $C_s t$.

Frequent diseases are repetitive, i.e. they appear more frequently in the database in all seasons. The threshold value is user-defined minimum frequency of the disease in all seasons. Frequent diseases are extracted from the database using the user-defined count threshold value $C_f t$. After identifying a frequent disease, it is anonymized by modifying its count to
equal that of normal diseases by repeatedly decreasing the frequent disease count value by $C_n t$ until its count value is less than or equal to $C_n t$.

Seasonal diseases occur more frequently (have larger number of counts) in particular time periods (seasons) in the database. The threshold value is user-defined minimum occurrence of the disease during different seasons. Seasonal diseases are extracted from the database using the sliding window method. Normally the window size is 30 days. Seasonal diseases are extracted from the database using the user-defined count threshold value $C_n t$. After identifying a seasonal disease, it is anonymized by modifying its count to equal that of normal diseases by repeatedly decreasing the seasonal disease count value by $C_n t$ until its count value is less than or equal to $C_n t$.

Geographical diseases occur more frequently (have larger number of counts) in particular areas. The threshold value is user-defined geographical areas where the disease occurs. To identify geographical diseases, the frequent diseases in total area (whole database) are tabulated, from which the frequent disease in a particular area is identified. Taking the occurrence of two or more frequent diseases, one can identify the geographical disease. Geographical diseases are extracted from the database using the user-defined count threshold value $C_n t$. After identifying a geographical disease, it is anonymized by modifying its count to equal that of normal diseases by repeatedly decreasing the geographical disease count value by $C_n t$ until its count value is less than or equal to $C_n t$.

The sensitive, frequent, seasonal and geographical diseases are plotted into the window table. The count of each disease in the window table is divided by the size of the window table.
1.6.2 Balanced Constraint Measure Algorithm to Preserve Privacy from Sequential Rule Discovery

The methodology of this proposed system has three main phases:

1. In the first (pre-processing) phase, the significant diseases are mined from the medical database and converted into a sequential database by classifying the diseases into four categories: sensitive, frequent, seasonal and geographical.

2. In the second phase, the prefix span algorithm is applied to the generated sequential database to make sequential patterns by sequential rule generation. These sequential patterns are converted into sequential rules, which are evaluated through support and confidence measures.

3. Next, the unwanted sequential rules are filtered out using minimum support and minimum confidence rules. The filtered rules are sent to the proposed privacy algorithm, which is designed to create the output which balances the knowledge discovery versus information loss.

The proposed privacy preservation algorithm initially generates a random value for every rule. Based on the random value, it makes modifications on the sequential rules and evaluates if the modified sequential rules meet the appropriate values of knowledge discovery and information loss.

If the random value of the modified sequential rule is between 0 and 0.2, the algorithm checks the corresponding item with a set of significant diseases. If the rule has any disease from the significant database, the algorithm removes that disease from that sequential rule. If the rule has more
than one disease from significant database, the algorithm removes one disease at a time from that rule. If the rule has no disease from significant database, the algorithm selects and removes a disease randomly.

If the random value of the modified sequential rule is between 0.2 and 0.4, the algorithm changes the position of the diseases from the rule. By changing the order of the disease in the sequential rule, the sensitive information is changed (hidden). If the random value of the modified sequential rule is between 0.4 and 0.6, the algorithm reduces the support value of the rule to 10%. If the random value of the modified sequential rule is greater than 0.6, the algorithm reduces the confidence value to 10%.

If the evaluated result satisfies the user-defined thresholds, the proposed algorithm releases the modified sequential rules, else further iteration is carried out until the sequential rules satisfy the user-defined threshold value.

1.6.3 Balanced Constraint Measure Based Algorithm on Pre-Mining for Privacy Preserved Data Publishing

This proposed work presents an algorithm that generates a sanitized database which maintains a proper balance between information privacy and knowledge discovery. It first generates the significant diseases from the original database, then converts the original database into a sanitized database through the proposed random sanitization process.

Next, the generated sanitized database is evaluated on the basis of Knowledge Discovery (KD) and Privacy Factor (PF). This evaluation process calculates the KD and PF by comparing the significant diseases generated from the original database vs. the sanitized database. If the values of KD and PF satisfy the user-specified threshold value, the sanitized database is ready.
for publishing, else the proposed algorithm repeats the sanitization process until the KD and PF values are satisfied.

Finally, the proposed algorithm is evaluated by comparing it with similar popular algorithms on the criteria of KD and PF. The experimental results show that the proposed algorithm performed well with 77% running time compared to similar popular algorithms. The memory requirement for the proposed algorithm is 25%, compared to 31% memory by previous algorithms. The proposed algorithm also performed 53% better than previous algorithms in terms of number of significant diseases affected by the proposed sanitization process.

1.7 ORGANIZATION OF THE THESIS

The remainder of the thesis is organized as follows:

Chapter 2 reviews related work in data mining, privacy preserving on databases, sequential mining, rule hiding and classification, balancing the knowledge discovery and information loss.

Chapter 3 discusses an efficient algorithm developed in this work for the privacy preserving mining which clearly demonstrates its superiority over the privacy preservation on databases and the mining performed by prefix span algorithm.

Chapter 4 deals with Balanced Constraint Measure-Based Algorithm for Privacy Preserved Sequential Rule Discovery which performs the privacy preservation after the rule discovery with minimal side effects information loss.
Chapter 5 discusses the Balanced Constraint Measure Based Algorithm On Pre-Mining for Privacy Preserved Data Publishing. It deals with privacy preservation of databases in which the scope knowledge discovery increases compared to other mining processes.

Chapter 6 discusses the comparative analyses of various processes of privacy preservation algorithm with running time, disease changes and memory requirement.

Chapter 7 deals the conclusion and discusses future directions of this thesis.