# Hiding Sensitive Knowledge using Multidimensional Association Rule Mining for Privacy Preserving

*Developed by:*

**Mehvish Khanum**

**Registration #: 675-FBAS/MSCS/S12**

*Supervised by:*

**Ms. Zakia Jalil**

**Lecturer**

Department of Computer Science & Software Engineering

Faculty of Basic and Applied Sciences

International Islamic University Islamabad

[2015]

## Department of Computer Science & Software Engineering

## International Islamic University Islamabad.

**Date:**

## Final Approval

This is to certify that we have read the thesis submitted by Mehvish Khanum registration number 675-FBAS/MSCS/S12. It is our judgment that this thesis is of sufficient standard to warrant its acceptance by International Islamic University, Islamabad for the degree of MS in Computer Science.

**Committee**

**External Examiner:**
Dr. Shareefullah Kahn
Senior HOD and Associate Professor
Department of Computing (Doc)
School of Electrical Engineering and Computer Science (SEECS)
National University of Sciences and Technology (NUST)
H-12, Islamabad,Pakistan.

**Internal Examiner:**
*Ms Sadia Arshid*
*Lecturer*
*Department of Computer Science & Software Engineering,*
*International Islamic University,*
*H-10, Islamabad,Pakistan.*

Sadia

**Supervisor:**
Ms. Zakia Jalil
*Lecturer*
*Department of Computer Science & Software Engineering,*
*International Islamic University,*
*H-10, Islamabad,Pakistan.*

**A dissertation Submitted To**

**Department of Computer Science & Software Engineering,**

**Faculty of Basic and Applied Sciences,**

**International Islamic University, Islamabad**

**As a Partial Fulfillment of the Requirement for the Award of the**

**Degree of MS Computer Science.**

# Declaration

I hereby declare that this Thesis" **Hiding Sensitive Knowledge using Multidimensional Association Rule Mining for Privacy Preserving"** neither as a whole nor as a part has been copied out from any source. It is further declared that I have done this research with the accompanied report entirely on the basis of our personal efforts, under the proficient guidance of my supervisor Ms. Zakia Jalil. 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 of the training institute or educational institutions, I shall stand by the consequences.

**Mehvish Khanum**

**Registration#: 675-FBAS/MSCS/S12**

# Acknowledgement

My last remaining task is to acknowledge all those that have contributed to the work described in this thesis. This is an impossible task to give gratitude to many people that have helped me design, implement, criticize and evangelize the work.

First of all I would like to thank Allah Almighty whose blessings, help and guidance has been a real source of all my achievements in my life.

I would like to thank my supervisor **Ms. Zakia Jalil**. With her too much busy schedule she always managed to take her time out to guide me.

My family deserves special attention for their inseparable support and prayers. They have helped me in shaping my thoughts and being confident and source to understand different tedious aspects. It was their vision, care and persistence confidence in me. It is largely due to their efforts that I am as I am today. Thank you all.

I would like to thank my husband **Sajid** and my in-laws. This thesis would not have been possible without their kind support. Thank you all for the motivation and encouragement. Special thanks to Naheed, Naveed, Mobeen, Zainab, Nadeem, Bkhtawar, Saad, , Fatima, Kiran, Ibrahim, Shaista and Haad.

**Mehvish Khanum**
**(675-FBAS/MSCS/S12)**

# Project In Brief

| | |
|---|---|
| **Project Title:** | Hiding Sensitive Knowledge using Multidimensional Association Rule Mining for Privacy Preserving |
| **Undertaken By:** | Mehvish Khanum<br>675-FBAS/MSCS/S12 |
| **Supervised By:** | Ms. Zakia Jalil<br>Lecturer<br>Department of Computer Science & Software Engineering,<br>Faculty of Basic and Applied Sciences,<br>International Islamic University Islamabad. |
| **Start Date:** | December, 2013 |
| **Completion Date:** | March ,2015 |
| **Tools & Technologies:** | Weka 3.6.9 |
| **Documentation Tools:** | MS Office 2007 for Documentation |
| **Operating System:** | Windows 7 Ultimate (32-bit) |
| **Processor:** | Inter (R) Core(TM)2 Duo CPU T6600 @2.20GHz |
| **Installed Memory:** | 3.00 GB |

# Abstract

Data mining aims to find patterns from huge amounts of data. These patterns are expressed in the form of association rules. Data mining functionalities are widely used to specify the kind of patterns which are to be found in data mining tasks. However association rules are not directly used for prediction. They are a helpful starting point for further exploration which helps analyzers to understand enormous data.

Hiding sensitive knowledge is a new dimension in data mining. Privacy preserving deals with the protection of sensitive information against unauthorized usage. Algorithms are designed to transform original data so that private information will not be discovered after mining process. Exploratory data analysis has been performed using classification and clustering. Moreover, a data hiding approach which changes percentage of a given attribute values is presented. A given percentage of the instances are changed in the way that a set of instances are randomly selected. The attribute given by its index is changed from its current value to one of the other possible ones, also randomly. This is done with leaving a portion the same. By adding noise in the data or by adding false values, sensitive information can't be extracted.

# Table of Contents

# ACRONYMS

| | |
|---|---|
| PPDM | Privacy Preserving Data Mining |
| KKD | Knowledge Discovery in Database |
| KHD | Knowledge Hiding in Database |
| OLAP | On-line Analytical Processing |
| MCT | Minimum Confidence Threshold |
| MST | Minimum Support Threshold |
| SMO | Sequential Minimal Optimization |
| DM | Data Mining |
| QA | Quantitative Attribute |
| ARFF | Attribute Relation File Format |
| CID | Central Intelligent Data |
| DBM | Data Base Management |
| PCA | Principal Component Analysis |
| MCA | Multiple Correspondence Analysis |
| AHC | Agglomerative Hierarchical Clustering |
| (ITh) | Linkage Inconsistency Threshold |

# Chapter 1

# Introduction

# 1. Introduction

Data Mining provides techniques and tools that add intelligence to the data warehouse. It's a process of identifying patterns from a large amount of data. Once information is extracted it's really important to protect private information against unauthorized usage. A number of privacy preserving algorithms are developed to transform the original data so that private information and knowledge is not discovered after the process of data mining techniques. Hiding association rules or altering the data before mining process are famous techniques to prevent the extraction of sensitive association or information in data mining process.

## 1.1 Motivations and Challenges

High dimensional data can be in terms of data volume or the number of dimensions. In such huge data it is difficult task for designers to decide which dimension is most informative and should be included in analysis phase. Data can be sensitive containing individuals personal information so its disclosure can violate the privacy e.g. for the individuals who are recorded in data, if their identity is revealed to untrusted third party or if sensitive knowledge about them can be mined from data. This problem generates the need of hiding association rules containing sensitive information by altering the data before mining process begins.

## 1.2 Background

Data mining aims to find useful patterns from large amounts of data. These patterns represent knowledge and are expressed in the form of decision tress, clusters or association rules. Data mining functionalities are used to specify kind of patterns to be found in data mining tasks. Han and Kamber (2003) described that data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database while Predictive mining tasks perform inference on the current data in order to make predictions. Association rules are not directly used for prediction. They are, however, a helpful starting point for further exploration, which makes it a popular tool for understanding data.

Basically in Privacy Preserving Data Mining (PPDM), two problems are addressed: one is the protection of private data which deals that how to get normal mining results when private data can't be accessed accurately; another is the protection of sensitive rules which deals that how to protect sensitive rules contained in data from being discovered, while non-sensitive rules can still be mined normally, procedure called Knowledge Hiding in Database (KHD) which is opposite to Knowledge Discovery in Database (KDD). A study by Atallah et al. (1999) concluded the concept of data sanitization. Main idea was to select some transactions to modify (delete or add items) from original database through some heuristics. There are basically two groups of data sanitization

1- Data modification
2- Data reconstruction

Privacy preserving data mining can be categorized into data hiding and knowledge hiding. In data hiding main concept is how the privacy of data can be maintained before hiding process begins. In knowledge hiding instead of protecting raw data, sensitive information from data mining results is protected using distortion and blocking techniques.

## 1.3 Problem Domain

In high dimensional data, there can be some dimensions containing private or sensitive information. Assuming there is no prior knowledge of the underlying data, classification and clustering methods are used to find most informative dimensions. Knowledge is discovered in the form of association rules to predict the future behavior and data is modified after the mining process to hide sensitive information. Multidimensional association rules R are mined from the database D after the analysis process. Database D is then transformed into $D_M$, such that no sensitive information can be extracted.

## 1.4 Key Points

Some definitions and notations which are necessary to understand the work done in this research are given in this chapter.

### 1.4.1 Association Rule

Association rules are if/then statements that help to uncover relationships between seemingly unrelated data in a relational database or other information repository as concluded by Agrawal et al. (1996). Association rule contain antecedent (if) and consequent (then). An antecedent is an item found in the data and consequent is an item that is found in combination with the antecedent. Association rules are used for analyzing and predicting customer behavior and play an important part in catalog design, store layout, product clustering and shopping basket analysis.

### 1.4.2 Association Rule Mining

Association rule mining discovers interesting relations between variables in large databases. Association rules are also employed in web usage mining, intrusion detection and bioinformatics.

### 1.4.3 Classification of Association Rules

Classification rules can be classified in the following categories

#### 1.4.3.1 Boolean Association Rule

If a rule concerns associations between the presence and absence of items, it is Boolean association rule.

#### 1.4.3.2 Single Dimensional or Intra-dimension Association Rule

If the items or attributes in an association rule reference only one dimension, it is single dimension association rule. Such rules are commonly mined from transactional data.

#### 1.4.3.3 Multidimensional Association Rule

If a rule involves two or more dimensions/predicates, it is a multidimensional association rule. Such rules are commonly mined from relational database or data warehouse. Multidimensional association rules with no repeated predicates are called interdimensional association rules.

#### 1.4.3.4 Hybrid-dimensional Association Rules

Multidimensional association rules with repeated predicates, which contain multiple occurrences of some predicates are called hybrid-dimension association rules.

### 1.4.4 Techniques Used for Mining Multidimensional Association Rules

Techniques used to mine multidimensional association rules are

#### 1.4.4.1 Mining Multidimensional Association Rules using Static Discretization of Quantitative Attributes

Quantitative attributes are discretized prior to mining using predefined concept hierarch ies, numeric values are replaced by ranges. Categorical attributes may also be generalized to higher conceptual levels. If all the resulting task relevant data is stored in a relational table, apriori algorithm requires just a slight modification to find all frequent predicate sets rather than frequent item sets.

#### 1.4.4.2 Mining Quantitative Association Rules

Quantitative association rules are multidimensional association rules in which the numeric attributes are dynamically discretized during the mining process to satisfy some mining criteria such as maximizing the confidence or compactness of the rules mined. Association Rule Clustering System (ARCS) approach is used to find such kind of rules.

#### 1.4.4.3 Mining Distance Based Association Rules

Distance based partitioning group values that are close together within the same interval. Distance based partitioning considers the density or number of points in an interval. A cluster is a set of tuples defined on an attribute X, where the tuples satisfy a density threshold and frequency threshold (minimum number of tuples in a cluster). Diameter defines the closeness of tuples. These clusters are combined to form distance based association rules. Degree of association measure can be defined using average inter cluster distance or the centroid, where the centroid of a cluster represents the average tuple of the cluster. Sometimes it is difficult to find strong association rules among data items at low or primitive level of abstraction. Data mining systems provide capabilities to mine association rules at multiple levels of abstraction. Rules generated from association rule mining with concept hierarchies are called multiple-level or multilevel association rules, as they consider more than one concept level. Basically a top down approach is used, where counts are accumulated for the calculation of frequent item sets at each concept level.

#### 1.4.4.4 Boolean Matrix Based Approach

Liu & Wang (2007) suggested an algorithm based on Boolean matrix to generate the multidimensional rule which has no repetitive predicates from relational databases. A Boolean Matrix based approach has been used to find the frequent itemsets, the items forming a rule come from different dimensions. Algorithm adopts Boolean relational calculus to discover frequent predicate sets. When using this algorithm first time, it scans the database once and will generate the association rules. Apriori property is used in algorithm to prune the item sets. It is not necessary to scan the database again, it uses Boolean logical operations to generate the association rules. It stores all data in the form of bits, so it needs less memory space and can be applied to large relational databases.

### 1.4.5 Database Attributes

Database attributes can be categorical or quantitative. Categorical attributes have a finite number of possible values with no ordering among the values. Categorical attributes are also called nominal attributes since their values are names of things e.g. occupation, brand, colour etc. Quantitative attributes are numeric and have an implicit ordering among values e.g. age, income, price etc.

### 1.4.6 Pattern Interestingness Measures

A pattern is interesting if it is easily understood by humans, valid on a new or test data with some degree of certainty. A pattern is also interesting if it validates a hypothesis that the user sought to confirm. Several objective measures of pattern interestingness exist. An objective measure is support, representing the percentage of transactions from a transaction database that the given rule satisfies.

$$\text{support}(A \rightarrow B) = P(A \cup B) \qquad (1)$$

Another objective measure is confidence, which assess the degree of certainty of the detected association.

$$\text{confidence}(A \rightarrow B) = P(B|A) \qquad (2)$$

Han and Kamber (2003) explain that each interesting measure is associated with a threshold, which is controlled by user. Rules that cannot satisfy a threshold value can be considered uninteresting. Rules below the threshold likely reflect noise, exceptions or minority cases and are probably of less value. Objective measures are insufficient unless combined with subjective measures that reflect the needs and interests of a particular user. User provided constraints and interestingness measures should be used to focus the search. Association rule mining is an example where the use of constraints and interestingness measures can ensure the completeness of mining.

## 1.5 Overview of the Manuscript

In this section overview of the remaining contents of this manuscript which is structured into 4 main chapters is given.

**Chapter 2** discusses previous work of researchers and highlights the critical points, theoretical and methodological contributions of already existing work.

**Chapter 3** highlights proposed technique in detail.

**Chapter 4** consist the dataset used for experimentation and evaluated results are also discussed in detail.

**Chapter 5** contains the review of the application and the possible future work.

# Chapter 2

# Literature Review

# 2. Literature Review

Chapter highlights the critical points, theoretical and methodological contributions of already existing work.

## 2.1 Multidimensional Association Rules

As cited in literature Messaoud et al. (2007), Kamber et al. (1997), Nestorov & Jukic (2003) and Kaya & Alhajj (2005) were the first to target the issue of multidimensional rules in multidimensional environment. In all these approaches rules are mined in the form of meta rules provided by the user, which ensures that the rules mined will be of interest of the user. But it has the drawback that the rules lying outside the template will not be discovered and many interesting rules can be unexplored due to the lack of the domain knowledge by the user.

## 2.2 Knowledge Discovery from Multidimensional Data

Usman et al. (2013a) proposed an approach for discovery of cubes of interest in large multidimensional datasets which does not rely on availability of domain knowledge because there are scenarios exist, where we have limited knowledge of domain. Reliance on domain knowledge tends to constrain only encapsulate known knowledge. Hierarchical clustering is used along with PCA (Principal Component Analysis) and MCA (Multiple Correspondence Analysis) at multiple levels to construct data cubes. PCA and MCA are data reduction techniques used to capture greatest degree of variation in data and to rank significant dimensions and facts. Two main contributions are made in the proposed method: Data cubes are generated at different levels of data abstraction and at each level of data abstraction most significant interrelationships between numeric and nominal variables are identified, which enables the identification of the cubes of interest. PCA is data reduction technique used for visualization of high dimensional data, identification of significant variables and dimensionality reduction. It ranks numeric variables in terms of degree of variance. Agglomerative Hierarchical Clustering (AHC) is used on numeric data to generate binary clustering tree or dendrogram. Distance between two clusters is represented by length of the link. Linkage inconsistency threshold (ITh) is used to determine cut-off point. Higher the value of this threshold indicates that less similar clusters are connected by links. Threshold value is used to define a number of clusters and the users are not required to set the number of clusters. Numeric values are ranked on the basis of calculated communalities. AHC requires nominal variables to be transformed into numeric, but here it is not done and MCA is used for nominal variables. MCA is specially designed for nominal variables and is an extension of simple correspondence analysis technique to account more than 2 variables. After applying AHC and ranking numeric and nominal variables multidimensional schema is generated. It can be done on the basis of top k and m thresholds where k is the number of highest ranked dimensions and m is the number of highly ranked facts. Operations like Drill-down, Roll-up, Slice and Dice can be applied by the user to construct informative data cubes. At the end informative data cubes are constructed on the basis of highly ranked facts and dimensions and by providing his/her own choice from multidimensional schema. As each user has specific analysis needs. Future work may include the use of entropy to identify information content of variables at different points of data hierarchy. Association rule mining can be integrated with proposed framework to get further insights at various levels of data abstraction.

## 2.3 Diverse Association Rules

Usman et al. (2013b) presented an approach that extends the capability of OLAP to detect hidden associations and to predict future trends based on historical data by deriving association rule mining from multidimensional schema. Three main contributions are made in this paper: a knowledge discovery methodology combining machine learning and statistical methods is presented to identify interesting regions of information and diverse association rules in large multidimensional data cubes. Information gain is used to identify and rank most informative dimensions. Through the application of Principal Component Analysis (PCA) and information gain, most informative data cubes are extracted at different level of abstraction and the effects of abstraction level on information content is studied. Methodology proposed works as follows.

First step is to perform Agglomerative Hierarchical Clustering (AHC) to group similar objects (automatic method is adopted for clustering which generates binary tree of clusters based on an automatically identified cut-off point and labels clusters with automatically determined labels that are based on their position in the data hierarchy), entropy and information gain is used to rank dimensions with high information. Multidimensional schema is constructed from which association rules are generated and different measures are used to test diversity of rules. One issue with any form of clustering is to determine the number of clusters and in hierarchical clustering it is to determine at what point to terminate the generation of the dendrogram (dendrogram is a tree diagram frequently used to illustrate the arrangement of the clusters produced by hierarchical clustering). A threshold value is used to determine at which point to stop the generation of dendrogram. Distance between 2 clusters is represented as length of links. A link whose height differs from the height of the links below it indicates that the clusters at this level in the dendrogram are much farther apart from their child clusters. This link is said to be inconsistent with the links below it and inconsistency threshold value at this level in the hierarchy is cut-off point. Using PC numeric variables are ranked and nominal variables are ranked using information gain.

Next step is to obtain natural grouping for each of the nominal variables. Multidimensional scaling is used to identify the semantic relationships among values in each nominal variable. After receiving ranked lists of numeric and nominal variables, a multidimensional STAR schema is created by treating nominal variables as dimensions and numeric variables as facts. SQL quries are used to create necessary tables and define table relationships that are needed to implement multidimensional schema. A number of interestingness measures Rae, CON and Hill have been proposed in literature. Methodology presented by Usman et al. (2013b ) permits interestingness evaluation only through diversity measure. Dimensions that were designed include two level hierarchies, first consisting groups and second of individual values. Future work may include exploring deeper hierarchies or/and to find automatic method for detecting certain fact variables to define dimensions. Such as numeric variable (used as fact variable in proposed algorithm) can be used as dimension variable by using discretization process to discretize its value into distinct ranges. Methodology can also be used with complex datasets from different domain.

## 2.4 Association Rule Hiding Approaches

Association rule hiding approach falls into 3 categories. First category is of heuristic approach. This approach generated efficient, scalable and quick response results. It is further divided into sensitive transaction identification method, sensitive association clustering method and sanitization matrix method. But this approach suffers from undesirable side effects. Second category is of boarder based approach, which utilizes the concept of borders to track the impact of altering transactions by greed selecting those modifications while minimizing the side effects. This approach focuses on preserving border of non sensitive frequent item sets rather than considering all non sensitive item sets during sanitization process. Third is non-heuristic or exact approach which conceives hiding process as constraint satisfaction problem. These problems are solved by inter programming. Compared to heuristic and boarder based, this guarantees quality for hiding sensitive information.

## 2.5 Privacy Preserving Data Mining Algorithms

Sensitive knowledge hiding problem was proven to be nondeterministic polynomial time hard problem. Chen et al. (2004) proposed a novel framework inspired by inverse frequent set mining problem. Algorithm first performed sanitization on an itemset lattice and then association rules were derived. In other words algorithm was used to conceal sensitive rules by sanitizing itemset lattice rather than sanitizing original dataset. Then a reconstruction procedure was performed to construct a new released dataset from sanitized itemset lattice. Algorithm lacks to give concrete guidance on how to sanitize itemset lattice also does not guarantee that itemset lattice with consistent support value configuration relationship can be found in polynomial time .

### 2.5.1 Association Rule Hiding using Support and Confidence

Patidar et al. (2013) suggested a new modified hybrid algorithm for privacy preserving. Algorithm is a modification of work presented in Belwal et al. (2008) and takes lesser number of passes to hide a specific association rule. Main contribution of the paper is that all the given sensitive rules are successfully hidden without any side effect in small as well as large databases. Previous algorithms also hide all sensitive rules without any side effect in small data but when the database is too large then existing algorithms require a large number of passes to produce the required results resulting in increased time complexity. Main idea proposed in this is as follows: to hide the rule X→Y: where X is the sensitive item on LHS, algorithm increases the special variable of the rule X→Y until confidence(X→Y) goes below a minimum specified confidence threshold(MCT). As the confidence goes below MCT rules X→ Y is hidden i.e., it will not be discovered through data mining algorithm.

### 2.5.2 Heuristic Approach

Lakshmi & Rani (2013) presented a heuristic approach to hide sensitive item sets efficiently by adopting two criterions. Methodology presented in this paper is based on the work proposed by Wang et al. and uses 2 criterions to obtain distorted database. Wang et al. suggested a procedure in which all the sensitive item sets whose length is greater than two are considered to find the pairs sub patterns. From this pair sub patterns only significant pair sub patterns are considered as sensitive to hide the sensitive patterns. This procedure is very important in a way that it avoids the problem of forward inference attack. In order to avoid forward inference attack problem, at least one sub pattern with length of two of the patterns should be hidden. This split pattern procedure helps to speed up the hiding process. This methodology protects private information by doing sanitization process but before participating in the sanitization process, the method analysis the side effects and select the most promising one to change so that side effects can be fully avoided or accepting few side effects which will not harm the informational accuracy. If an itemset <Ai,Aj> is to be hidden in first criteria that item will be selected as victim item whose frequency will be less in non sensitive frequent itemset. If the count of both the items in non sensitive frequent itemset is equal then any of the items Ai or Aj can be selected randomly. After identifying victim item, in the second criteria, minimum number of suitable transactions has to be selected from all supporting transactions for itemset <Ai,Aj> . A minimum number of transactions required to hide itemset is

$$<Ai,Aj> *sup-Mintrans+1 \qquad (3)$$

For each support transactions for item set <Ai,Aj> weight is computed by

A number of dependent items with the victim item-number of infrequent item sets associated with the victim item. Based on weight, supporting transactions are stored in the Mintrans table in ascending order. Mintrans is then used for sanitization process.

### 2.5.3 Blocking Technique

To hide sensitive association rules in binary datasets by blocking some data values an algorithm Blocking Algorithm (BA) is presented. Quantitative comparison of proposed algorithm with already published algorithms and qualitative comparison of the efficiency of proposed algorithm in hiding association rules is presented. Concept of border rules by putting weights in each rule is utilized. Effective data structure for the representation of the rules to minimize side effects of hiding process and to speed up the selection of victim transactions is presented. To hide sensitive rule R

$$I_L \Rightarrow I_R conf(\mathrm{R}) = \frac{\sup(I_L \cup I_R)}{\sup(I_L)} \tag{4}$$

$$\sup(\mathrm{R}) = \sup(I_L \cup I_R) \tag{5}$$

By blocking 1's either minconf(R) below MCT or minsup(R) below MST can be reduced. To reduce support, some items from $I_L$ or $I_R$ are selected and blocked (by replacing 1's with ?'s) from transactions that support sensitive rule. If these items are selected from $I_R$ both minsup(R) and minconf(R) will be decreased. If items are selected from $I_L$ to block, minconf(R) may not be decreased because both numerator and denominator are decreased. So preferably items from $I_R$ should be selected. Alternatively by blocking 0's we decrease minconf(R) by selecting transactions that partially supports R (transactions in which exactly one item of $I_L$is 0 and at the same time at least one item of $I_R$must be 0) and by replacing 0's with ?'s. If 0 item in $I_L$is blocked (replacing 0 with a ?) then minimum confidence of R will be reduced because the denominator of conf(R) will be increased while the numerator will remain the same.

Algorithm adds uncertainty in the database by adding question marks in a way that the database can be usable by a data miner that receives the database and at the same time an adversary cannot infer the sensitive rules that BA will hide. The algorithm aims to achieve the following goals

a) Reduce the minimum confidence of sensitive rules below (MCT-SM).

b) Do not reduce the minimum confidence of non-sensitive rules

If the adversary finds the maximum confidence of all the rules in the modified database, she will find many new ghost rules that did not exist in the initial database so the adversary cannot assume with certainty which of the rules that have maximum confidence above MCT were the sensitive rules. On the other hand, a data miner who wants to find useful information from the database can find the minimum confidence of all the rules, excluding in that way the sensitive rules from his information.

### 2.5.4 Data Distortion Technique

Shah et al. (2012) suggested that association rule hiding algorithms minimally modify original database such that no sensitive association rule is derived from it. Paper is a survey of association rule hiding methods for privacy preservation. All existing techniques for association rule hiding are summarized. Heuristic approach is based on 2 techniques: data distortion technique and blocking technique. Data distortion technique is performed by altering an attribute value from 1's to 0's or from 0's to 1's in selected transactions. Two approaches are used for rule hiding in data distortion based technique: by reducing the confidence of rules and by reducing the support of rules. A comparison of described techniques in term of new generated rules, lost rules and hiding failure is given in following table.

TABLE 2.1: COMPARISON OF EXISTING HEURISTIC TECHNIQUES

| ISL | DSR | DCIS | DCDS | DSC | DSRRC |
|---|---|---|---|---|---|
| -Increases support of LHS<br>-Requires more running time<br>-Hiding failure 12.9%<br>-New rules 33%<br>-Lost rules 0% | -Decrease support of RHS<br>- hiding failure 0%<br>-New rules 5%<br>-Lost rules 11% | -Increase support of LHS<br>-Hiding failure 0%<br>-New rules 75%<br>-Lost rules 0% | -Decrease support of RHS<br>-Hiding failure 0%<br>-New rules 1%<br>-Lost rules 4% | -uses pattern inversion tree to store related information<br>-So only one scan of database is required<br>-New rules 4%<br>-Lost rules 9% | -Provides privacy while ensuring data quality<br>-Clusters sensitive association rules based on RHS of rules and hides as many as possible rules at a time by modifying transactions<br>-cannot hide rules having multiple RHS items |

Table 2.1 is a comparison of Increase Support of LHS (ISL), Decrease Support of RHS (DSR), Decrease Confidence by Increase Support (DCIS), Decrease Confidence by Decrease Support (DCDS), Decrease Support and Confidence (DSC) and Decrease Support of R.H.S. item of Rule Clusters (DSRRC) in terms of hiding failure, generation of new rules and number of lost rules. ISL and DSR are two heuristic algorithms. Predicting items are given as input to both algorithms to automatically hide sensitive association rules without pre mining and selection of hidden rules. DCIS and DCDS hides collaborative recommendation association rules without pre mining and selection of hidden rules.

In Fast Hiding Sensitive Association Rules (FHSAR) correlation between sensitive association rules and each transaction in original database is analyzed, which effectively selects proper item to modify. Naïve, Minimum Frequency Item Algorithm (MinFIA), Maximum Frequency Item Algorithm (MaxFIA) and Item Grouping Algorithm (IGA) selects the sensitive transactions to sanitize based on degree of conflict. Naive removes all items of selected transaction except the item with the highest frequency in the database. MinFIA selects item with the smallest support in the pattern as a victim item and it removes the victim item from the sensitive transactions. MaxFIA selects the item with the maximum support in the restrictive pattern as a victim item. IGA groups restricted patterns in groups of patterns sharing the same itemsets so that all sensitive patterns in the group will be hidden in one step.

Heuristic approach uses two strategies to hide sensitive rules: by inserting a new itemset in selected transactions e-g for rule X→Y, inserting itemset X in transactions not supporting Y will decrease confidence of rule X→Y, and by removing itemset from selected transactions either by remove itemset Y to decrease confidence of rule or by reducing support count of large itemset XY by removing items in X or Y from selected transactions.1.a, ISL and DCIS uses first strategy while 1.b, 2a,DSR,DCDS,FHSAR,DSRRC,DSC,Naïve, MinFIA, MaxFIA and IGA uses second strategy. Another way is by hiding approach. Item hiding algorithms hide all the association rules having particular itemset. Rule hiding algorithms hide all the sensitive rules given. Algorithm 1.a, 1.b, FHSAR and DSRRC are based on rule hiding approach while others are based on item hiding approach. From the analysis on heuristic algorithms it is concluded that whenever item insertion approach is used, there are more artifactual patterns created and it also have more hiding failure than deletion approach. So item insertion affects two parameters more, which are hiding failure and artifactual patterns. Sometimes it also shows missing cost. Whenever deletion approach is used at that time it affects misses cost parameter more. All other parameters are affected comparatively less than insertion approach.

In blocking existing value is replaced with a "?" it inserts unknown values in the data to blur the rules. When unknown values are inserted, support and confidence values would fall into a range instead of a fixed value. In this technique maximum confidence of a sensitive rule cannot be reduced. If it does not add much uncertainty in the database, adversary can infer the hidden values if he applies a smart inference technique. In database both 0's and 1's must be hidden during blocking, because if only 1's were hidden the adversary would simply replace all the?'s with 1's and would restore easily the initial database and many ?'s must be inserted, if we don't want an adversary to infer hidden data. Heuristics algorithms suffer from undesirable side effects due to the fact that it always aims at taking locally best decisions with respect to the hiding of the sensitive knowledge which, however, are not necessarily also globally best. Table 2.2 describes the comparative analysis of heuristic algorithms.

**Table 2.2: Comparative analysis of Heuristic Algorithms**

| Rule Hiding Strategy | Algorithm | Item Hiding Algorithm | | Rule Hiding Algorithm |
|---|---|---|---|---|
| | | LHS | RHS | |
| Insertion | ISL | Y | | |
| | DCIS | | Y | |
| | 1.a | | | Y |
| Deletion | DSR | Y | | |
| | DCDS | | Y | |
| | DSC | Y | Y | |
| | Naïve | Y | Y | |
| | MinFIA | Y | Y | |
| | MaxFIA | Y | Y | |
| | IGA | Y | Y | |
| | FHSAR | | | Y |
| | 1.b | | | Y |
| | 2.a | | | Y |
| | DSRRC | | | Y |

Boarder based approach hides sensitive association rule by modifying the borders in the lattice of the frequent and the infrequent itemsets of the original database. The itemsets which are at the position of the borderline separating the frequent and infrequent itemsets forms the borders. The algorithms in this approach differ in the methodology they follow to enforce the new, revised borders, in the modified database.

### 2.5.5 Exact Approach

Non-Heuristic or Exact approach formulates the hiding process as a constraints satisfaction problem or an optimization problem and solved by integer programming. These algorithms can provide optimal hiding solution with ideally no side effects. Algorithms falling in this category have very high time complexity due to the time that is taken by the integer programming to solve the optimization problem.

### 2.5.6 Boolean Matrix Based Approach

Khare & Pardasni (2010) suggested a Boolean Matrix based approach to discover frequent itemsets, items forming a rule come from different dimensions. Algorithm adopts Boolean relational calculus to discover frequent predicate sets. Database is scanned once to generate association rules. Apriori property is used to prune itemsets. It's not necessary to scan database again, it uses Boolean logical operations to generate association rules. Main idea proposed is as follows: Relational database is transformed in Boolean matrix. Frequent 1-itemset $L_1$ is generated. Apriori property is used to prune Boolean matrix. AND operation is performed on frequent 1-itemset of different dimensions to generate frequent 2-itemsets. Process is repeated to generate (K+1) itemsets from $L_K$. Table 2.3, 2.4 and 2.5 is matrix explanation of the literature review given in this section.

TABLE 2.3: CONCEPT MATRIX OF PRIVACY PRESERVING TECHNIQUES

| | Purpose | Technique | Results/Outcome | Advantage | Drawback | Future Work |
|---|---|---|---|---|---|---|
| Pontikakis et al. (2004) | -To hide sensitive association rules in binary datasets by blocking some data values | Blocking Algorithm(BA)<br>-Reduces the minimum confidence of sensitive rules below (MCT-SM).<br>-Does not reduce the minimum confidence of non-sensitive rules | -There is a trade-off between privacy and data loss<br>-At 60% safety margin BA blocks many 0's resulting in many ghost rules and number of side effects is increased | -Main advantage to other techniques is that database is not distorted, only some values become unknown.<br>-No false information is included.<br>-Data miner can assume that all the remaining values of database are true | -Privacy breaches of modified database<br>-Adversary who wants to infer can use placement of question marks and remaining data to find hidden values | -Ghost association rules using BA are suggested.<br>-So that adversary will not be able to infer which of the association rules having confidence above MCT are sensitive and which are the ghost one. |
| Khare & Pardasni (2010) | -To mine multidimensional association rules | - Boolean Matrix approach is used to find frequent itemsets<br>-Algorithm adopts Boolean relational calculus to discover frequent predicate sets | -Stores all data in the form of bits<br>-Less memory space<br>-Can be used on large relational databases | -Scans database only once<br>-Does not generate candidate item sets<br>-Uses Boolean vector(relational calculus ) to generate frequent item sets | -Additional computation is performed to transform relational database into Boolean Matrix | -Is not discussed in the paper |
| Lakshmi & Rani (2013) | -To hide sensitive association rules using a methodology based on | -Before sanitization process , methodology analysis side effects and selects most promising one to | -Proposed methodology utilizes two criterions(discussed above) to hide sensitive knowledge<br>-These criterions are | -Functionality of the proposed solution is given with sample databases<br>-Two cases are | | -Not given in the paper yet. |

| | | | | | | |
|---|---|---|---|---|---|---|
| | heuristic approach | change<br>-Resulting avoidance of side effects<br>-Or partially accepting few side effects, which does not harm information accuracy. | used to speed up the process of sensitive item hiding.<br>-Distorted database is obtained which hides all sensitive item sets. | taken under consideration: non overlapping and overlapping<br>-Criterions used in methodology are well described. | | |
| Patidar et al. (2013) | -To extract non sensitive knowledge from collaborative database while protecting sensitive information. | -Algorithm works on the basis of statistical measures (support and confidence). | -Algorithm claims to hide sensitive information in small as well as large databases without any side effects. | -Algorithm takes fewer numbers of passes to hide specific association rules.<br>-Algorithm hides association rules in which sensitive items occur in either left side or right side of the rule. | -There is no evidence of algorithm's success on large database, as it claims. | -Future work is not discussed in the paper. |

**TABLE 2.4: CONCEPT MATRIX OF A SURVEY PAPER OF PRIVACY PRESERVING TECHNIQUES**

| | Association Rule Hiding Approaches | Algorithms discussed in paper | Limitations of association rule hiding approaches | Analysis of algorithms based on heuristic approach |
|---|---|---|---|---|
| Shah et al. (2012) | -Heursitic<br>-Boarder based<br>-Non-Heuristic | -ISL<br>-DCID<br>-1.a<br>-DSR<br>-DCDS<br>-DSC<br>-Naïve<br>-MinFIA<br>-MaxFIA<br>-IGA | -Heuristics algorithms suffer from undesirable side effects it always aims at taking locally best decisions with respect to the hiding of the sensitive knowledge which, however, are not necessarily also globally best<br>-Boarder based algorithms rely on to decide upon the item modifications that they apply on the original database resulting in to identify optimal hiding | -Sensitive rules can either be hidden by inserting new itemset in selected transactions (1.a,ISL and DSL uses this strategy)<br>-Creates more artifactual patterns by increasing support of some itemsets Sometimes fails to hide sensitive rules due to new patterns are created as side effects.<br>-Or by removing itemset from selected |

| | | | | |
|---|---|---|---|---|
| | | -FHSAR<br>-1.b<br>-2.a<br>-DSRRC | solutions, although such solutions may exist for the problem at hand.<br>- Non-heuristic algorithms have very high time complexity due to the time that is taken by the integer programming to solve the optimization problem. | transactions(1.b,2a,DSR,DCDS,FHSAR, DSRRC,DSC,Naïve,MinFIA MaxFIA and IGA uses this strategy).<br>-Some frequent itemsets become infrequent<br>-Effects non sensitive rules which are hidden as side effects |

**TABLE 2.5: CONCEPT MATRIX OF LITERATURE REVIEW**

| | Purpose | Main Contribution | Technique | Results/Outcome | Drawback | Future Work |
|---|---|---|---|---|---|---|
| Usman et al. (2013a) | -To discover cubes of interest in large multidimensional datasets with minimal or no domain knowledge. | -Agglomerative Hierarchical Clustering is performed at multiple levels to construct data cubes.<br>-Principle Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) are used as data reduction techniques.<br>-PCA operates with numeric data while MCA operates with nominal data. | -- Linkage inconsistency threshold (ITh) is used to determine cut-off point<br>- Threshold value is used to define number of clusters and the users are not required to set the number of clusters. | -Proposed methodology is implemented on two real data sets also the results are verified. | | -Future work may include the use of entropy to identify information content of variables at different points of data hierarchy. -Association rule mining can be integrated with proposed framework to get further insights at various levels of data abstraction. |
| Usman et al. (2013b) | -To design multidimensional schema.<br>-To predict future trends based on historical data by deriving association rule mining from multidimensional | - Past work focuses on automatic derivation of database schema using conceptual models<br>-Methodology for designing multidimensional schema is adopted.<br>-Targets cases where | - Agglomerative Hierarchical Clustering (AHC) is performed to group similar objects.<br>-Automatic method is adopted for clustering, which is when to select cut off points<br>-Entropy and | -Three real world data sets are selected for experiments.<br>-Results show that association rules generated using proposed methodology are more diverse and have better predictive | - Permits interestingness evaluation only through diversity measure.<br>- A rule that is diverse does not mean that it is novel.<br>- Other rule | . Future work may include exploring deeper hierarchies or/and to find automatic method for detecting certain fact variables to define dimensions. Such as numeric variable |

| | | | | | | |
|---|---|---|---|---|---|---|
| | schema. | limited domain knowledge exists and no operational system is in place. | information gain is used to rank dimensions with high information.<br>-Multidimensional schema is constructed using most informative dimensions.<br>-Association rules are generated using Apriori.<br>-Diversity measure is used to evaluate rule interestingness. | accuracy than rules generated from same data without using multidimensional structure. | interestingness measures could also be used. | (which is used as fact variable) can be used as dimension variable by using discretization process to discretize its value into distinct ranges.<br>- Methodology can also be used with complex datasets from different domain. |

# Chapter 3
# Proposed Solution

# 3. Proposed Solution

All the discussion in the literature review emphasis that in high dimensional data, to predict that which dimension will be more informative is almost difficult for the analyst. We have suggested a way to analyze data and to discover knowledge in the form of association rules. More over data modification method is used to add noise to the original data base so that individual's private information cannot be disclosed.

## 3.1 Design Requirements

Parameters required for the construction of the model are as

### 3.1.1 Task Relevant Data

- Database
- Database tables
- Conditions for data selection
- Relevant attributes/ Dimensions
- Data grouping criteria

### 3.1.2 Knowledge Type to be Mined

- Association Rules

### 3.1.3 Pattern Interestingness Measures

- Certainty (confidence)
- Utility (support)

### 3.1.4 Visualization of Discovered Patterns

- Rules and graphs

## 3.2 Proposed System Framework

The proposed architecture of our methodology consists of the following steps

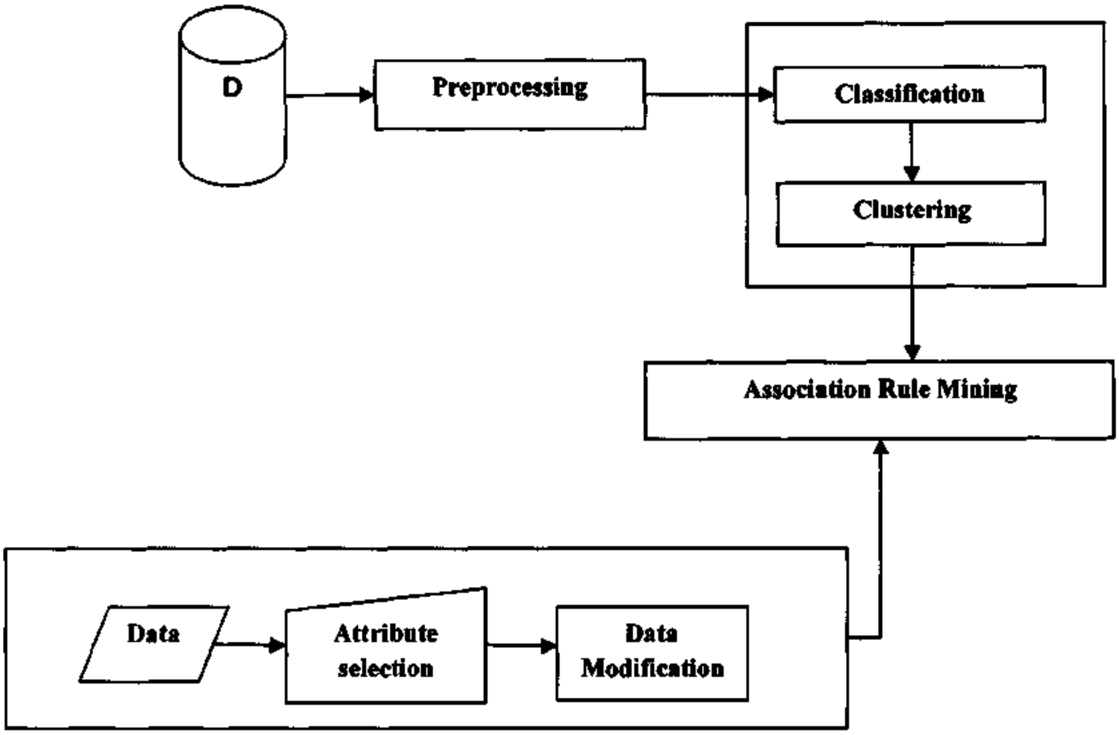


**Fig 3.1: Proposed Framework**

As the user does not have any prior knowledge of the underlying data. So classification and clustering is performed for exploratory data analysis. Preprocessing is performed to remove unknown values. For SMO data set is divided into training set and test set. Model is built by using training data set and then is evaluated by using test set to avoid over fitting. Clustering is performed to make groups of data to recognize patterns from data. This analysis phase gives a clear picture of most informative dimensions in our data. After the mining process, in our scenario, we are interested to publish these associations for which we have to hide private/sensitive information of individuals. So data modified method is used to add random values in sensitive attributes so that when this information is publically available, private information cannot be extracted. Following section describes each step in detail.

## 3.3 Preprocessing

As the dataset contains unknown values so preprocessing is applied to replace the missing values. Next discretization filter is used to discretize a range of numeric attributes of dataset into nominal attributes by using simple binning. Binning is a smoothing technique which smoothes sorted data values by consulting its neighborhood values. Sorted values are distributed into a number of buckets called bins. As binning method consults its neighborhood values so it performs local smoothing. For SMO an additional preprocessing step to reduce the sample size is performed.

## 3.4 Sequential Minimal Optimization Classifier

Platt (1998) proposed Sequential Minimal Optimization (SMO) algorithm for training support vector machines which requires solution of a very large quadratic programming (QP). SMO breaks this problem in smallest possible QP problems. These smaller QP problems are then solved analytically avoiding time consuming numeric QP optimization. Memory required for SMO is linear in the training set size resulting to handle a very large dataset. As matrix computation is not done so SMO is between linear and quadratic in training set size of various test problems. Classification algorithm SMO is applied for training support vector classifier to know which class a new object belongs to. There is a training set at the beginning which contains labeled items. SMO is built by using training dataset. To validate our classification tree, test data set is run through the model to ensure the accuracy of the model that test set is not too much different from the training set.

## 3.5 Cluster Formation using K-Means Clustering

K-means clustering is a partitioning method. The function kmeans partitions data into k mutually exclusive clusters, and returns the index of the cluster to which it has assigned each observation. Unlike hierarchical clustering, k-means clustering operates on actual observations (rather than the larger set of dissimilarity measures), and creates a single level of clusters. K-means clustering is often more suitable than hierarchical clustering for large amounts of data. kmeans treats each observation in data as an object having a location in space. It finds a partition in which objects within each cluster are as close to each other as possible, and as far from objects in other clusters as possible. Each cluster in the partition is defined by its member objects and by its centroid, or center. The centroid for each cluster is the point to which the sum of distances from all objects in that cluster is minimized. kmeans computes cluster centroids differently for each distance measure, to minimize the sum with respect to the user specified measure.

kmeans uses an iterative algorithm that minimizes the sum of distances from each object to its cluster centroid, over all clusters. This algorithm moves objects between clusters until the sum cannot be decreased further. The result is a set of clusters that are as compact and well-separated as possible. Minimization details can be controlled by using several optional input parameters to kmeans, including ones for the initial values of the cluster centroids, and for the maximum number of iterations.

K-mean clustering algorithm clusters observation into k groups and k is provided as input which determines a number of clusters. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. Clusters mean is computed using following steps

1. The algorithm arbitrarily selects k points as the initial cluster centers called centorid (mean).
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.

Step 2 and 3 are repeated until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

## 3.6 Knowledge Mining in the Form of Multidimensional Association Rules

After the analysis phase, Apriori algorithm is applied to mine multidimensional association rules. Algorithm iteratively reduces the minimum support until it finds the required number of rules with the given minimum confidence. The algorithm also has an option to mine class association rules. Apriori algorithm was proposed by Agrawal et al. (1994). The algorithm works by finding frequent itemsets in the database. The algorithm is a bottom up search in which an item set lattice is formed. In the lattice frequent item set are marked while moving upward in each level. It prunes many of the sets which are unlikely to be frequent sets. This property avoids extra passes over the database. All other item sets which have infrequent sub sets are ignored and not considered for calculation support. The algorithm terminates when no further successful extensions are found.

## 3.7 Data Modification for Privacy Preserving

An instance filter that changes a percentage of a given attributes values. The attribute must be nominal and missing value can be treated as value itself. A given percentage of the instances are changed in the way, that a set of instances are randomly selected using seed. The attribute given by its index is changed from its current value to one of the other possibly ones, also randomly. This is done with leaving a portion the same. By adding noise in the data or by adding false values, sensitive information can't be extracted.

# Chapter 4

# Experimentation & Results

# 4. Experimentation and Results

In order to understand data set, we need to find answers of the following questions.

1. Which attributes are most decisive to determine the income of a person?
2. What are the values of the attributes that would determine high income or low income of a person?
3. What conclusion can be made by answering above questions?

To find answer of the first question classification is performed. Classification results on test dataset will be verified to determine most decisive attributes.

To find answer of the second question, it is required to find which values of attributes association with >50k or <=50k. For this purpose association rule mining is performed.

## 4.1 Dataset

The proposed methodology is validated using an adult dataset that was extracted from UCI machine learning repository*. Dataset is multivariate with 48,842 instances. The total numbers of attributes are 14 with 8 nominal and 6 numeric. The dataset contains unknown values for some attributes. The last attribute, annual-pay, is the default target or class variable used for prediction. In this case it is a numeric attribute with two values <=50k or >50k which makes it a binary classification problem. This dataset is to be used to predict whether the described attributes effects if an adult's annual pay is below or above 50k per year. Dataset was divided into 2 sets adult.data and adult.test. Adult.data was used to train the model and adult.test was used to test it. There were 32561 instances in adult.data and 16281 in adult.test. Table 4.1 describes the attribute name and its type in the dataset.

**Table 4.1: Adult Dataset**

| Attribute Name | Attribute Type |
|---|---|
| age | numeric |
| workclass | nominal |
| fnlwgt | numeric |

| | |
|---|---|
| education | nominal |
| education-num | numeric |
| marital-status | nominal |
| occupation | nominal |
| relationship | nominal |
| race | nominal |
| sex | nominal |
| capital-gain | numeric |
| capital-loss | numeric |
| hours-per-week | numeric |
| native-country | nominal |
| annual-pay | numeric |

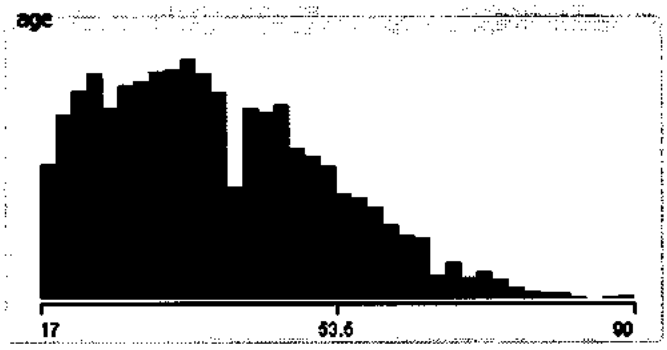


**Figure 4.1: Graphical representation of Attribute age**

*https://archive.ics.uci.edu/ml/datasets.html

Figure 4.1 is graphical representation of attribute age when is loaded into weka. One can view all the attributes in detail by exploring. Here for age attribute minimum value is 17, 1st quartile is 28, median is 37.

## 4.2 Preprocessing

Dataset contains unknown values. So a filter was applied to replace missing values of all the attributes with means and modes from training data set. This filter can handle string, nominal, numeric, binary, unary and relational classes. Attributes can be binary, numeric, unary, nominal, string and relational. Filter was selected by

Weka→filters→unsupervised→attribute→ReplaceMissingValues

Filter contains an option of ignoreClass which is used to unset class index temporarily before the filter is applied.

Fayyad &Irani (1993) MDL method was used for discretization. Discretization is a process of transferring continuous values into discrete. This process was carried out as a first step for numeric evaluation and implementation on digital computers. Attribute which is to be discretized was selected by its index and was transformed into nominal attribute.. Discretization was selected by

Weka→filters→unsupervised→attribute→Discretize

Hera age, fnlwgt,education-num,capital-gain,capital-loss,hours-per-week and annual-pay were numeric attributes. Attribute indices option was selected to specify the range of attributes to act on. Discretization was performed with a bin parameter of 10.Following values were given as input

AttributeIndices→first-3,5,11,12,13,last.

Bins→10

Table 4.2 describes the age attribute before discretization.

**Table 4.2: Attribute Age before Discretization**

| Statistic | Value |
|---|---|
| Minumum | 17 |
| Maximum | 90 |
| Mean | 38.58 |
| StdDev | 13.64 |

**Table 4.3: Attribute Age after Discretization**

| No | Label | Count |
|---|---|---|
| 1 | -inf-24.3 | 1119 |
| 2 | 24.3-31.6 | 1196 |
| 3 | 31.6-38.9 | 1217 |
| 4 | 38.9-46.2 | 1225 |
| 5 | 46.2-53.5 | 785 |
| 6 | 53.5-60.8 | 535 |
| 7 | 60.8-68.1 | 299 |
| 8 | 68.1-75.4 | 85 |
| 9 | 75.4-82.7 | 40 |
| 10 | 82.7-inf | 11 |

Table 4.3 describes the age attribute after discretization. Attribute's limit and its count is given.

## 4.3 Sequential Minimal Optimization

SMO classifier was applied on the dataset. SMO produced most accurate results with the lowest misclassification percentage. Several measures like TP Rate, FP Rate, Precision, Recall and F-Measure were used to evaluate the accuracy of the algorithm. . A sample of few values is given here. While the rest of the results can be viewed in appendix A.

```
=== Run information ===

Scheme:weka.classifiers.functions.SMO -C 1.0 -L 0.001 -P 1.0E-12 -N 0 -V -1 -W 1 -K
"weka.classifiers.functions.supportVector.PolyKernel -C 250007 -E 1.0"

Relation:     adult.data-weka.filters.unsupervised.attribute.ReplaceMissingValues-
weka.filters.unsupervised.attribute.Discretize-B10-M-1.0-Rfirst-3,5,11,12,13,last-
weka.filters.unsupervised.instance.Resample-S1-Z20.0

Instances:    6512

Attributes:   15

              age

               workclass

               fnlwgt

               education

               education-num
```

```
            marital-status
            occupation
            relationship
            race
            sex
            capital-gain
            capital-loss
            hours-per-week
            native-country
           annual-pay
Test mode:10-fold cross-validation
=== Classifier model (full training set) ===
SMO
Kernel used:
 Linear Kernel: K(x,y) = <x,y>
Classifier for classes:  <=50K,  >50K
BinarySMO
Machine linear: showing attribute weights, not support vectors
       -0.738  * (normalized) age='(-inf-24.3]'
 +      -0.222  * (normalized) age='(24.3-31.6]'
 +       0.2771 * (normalized) age='(31.6-38.9]'
 +       0.3655 * (normalized) age='(38.9-46.2]'
 +       0.4892 * (normalized) age='(46.2-53.5]'
 +       0.5186 * (normalized) age='(53.5-60.8]'
 +       0.1893 * (normalized) age='(60.8-68.1]'
+    -0.233  * (normalized) age='(68.1-75.4]'
+    -0.6467 * (normalized) age='(75.4-82.7]'
+     0      * (normalized) age='(82.7-inf)'
+    -0.2646 * (normalized)  workclass= State-gov
+    -0.312  * (normalized)  workclass= Self-emp-not-inc
+     0.0394 * (normalized)  workclass= Private
+     0.1489 * (normalized)  workclass= Federal-gov
+    -0.0884 * (normalized)  workclass= Local-gov
+    -0.0847 * (normalized)  workclass= State-gov
+     0.5784 * (normalized)  workclass= Self-emp-inc
+    -0.0171 * (normalized)  workclass= Never-worked
```

Weight of each attribute in the model is described in this section. Attributes with highest absolute numbers are most useful for classification

Below accuracy and confusion matrix measures are given for the learnt classifier. Confusion matrix describes how many samples are correctly classified. Here 691 samples from class >50k were wrongly classified as being from class <=50k.

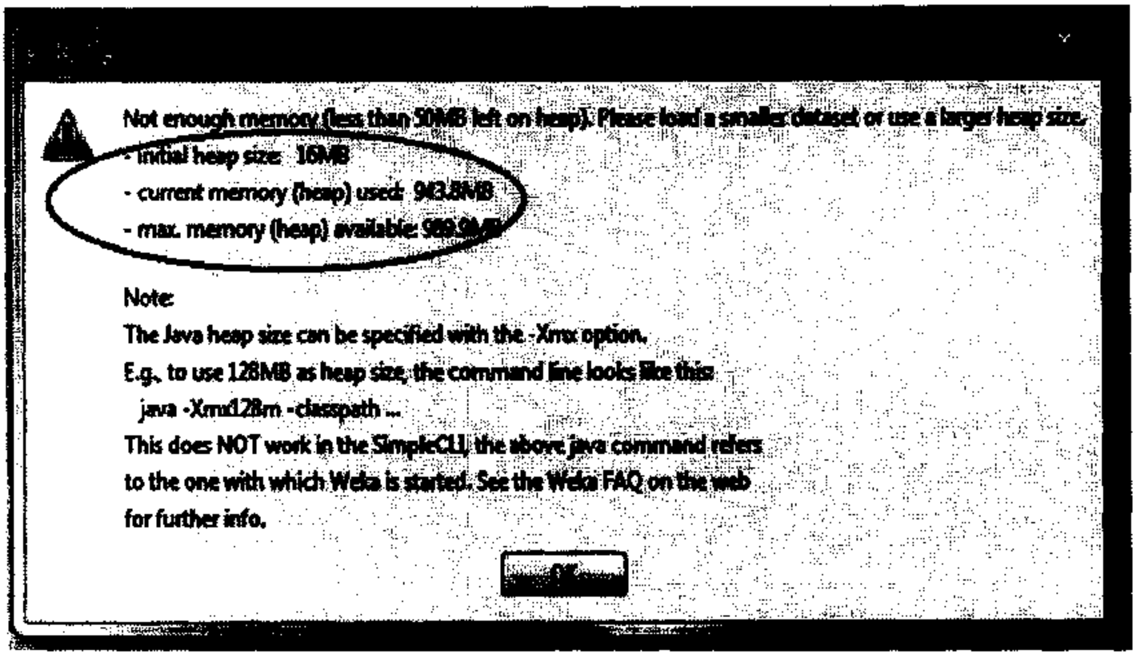
```
Number of kernel evaluations: 56973710 (50.222% cached)

Time taken to build model: 965.68 seconds

=== Stratified cross-validation ===

=== Summary ===

Correctly Classified Instances       5471          84.0141 %

Incorrectly Classified Instances     1041          15.9859 %

Kappa statistic                  0.5222

Mean absolute error                0.1599

Root mean squared error              0.3998
```

```
Relative absolute error          44.1431 %

Root relative squared error        93.9668 %

Total Number of Instances          6512

=== Detailed Accuracy By Class ===

           TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class

            0.93     0.447     0.87     0.93     0.899     0.741    <=50K

            0.553    0.07      0.71     0.553    0.622     0.741    >50K

Weighted Avg.  0.84    0.358    0.832   0.84    0.833    0.741

=== Confusion Matrix ===

  a   b   <-- classified as

 4616  350 |   a = <=50K

  691  855 |   b = >50K
```

**Table 4.4: SMO Classifier Summary**

| Total Instances | Misclassified Instances | Error Rate | Accuracy | Training Time |
|---|---|---|---|---|
| 6512 | 1041 | 15.98% | 84.01% | 965.68 seconds |

Model was built by using adult.dataset. This training model was then tested by using adult.test. It took so long to test this data set and ended up with the following message. Figure 4.2 is system generated message which shows memory status.



```
Not enough memory (less than 50MB left on heap). Please load a smaller dataset or use a larger heap size.
- initial heap size: 16MB
- current memory (heap) used: 943.8MB
- max. memory (heap) available: 989.9MB

Note:
The Java heap size can be specified with the -Xmx option.
E.g., to use 128MB as heap size, the command line looks like this:
  java -Xmx128m -classpath ...
This does NOT work in the SimpleCLI, the above java command refers
to the one with which Weka is started. See the Weka FAQ on the web
for further info.
```

**Figure 4.2**

For which dataset was further reduced to 40%, which also resulted with the above mentioned error. Again dataset was reduced to 30%. This time it worked well and output was generated after approximately 9 hours.

## 4.4 K- Mean Clustering

```
Time taken to build model (full training data) : 856 seconds

=== Model and evaluation on training set ===

Clustered Instances

0     4166 ( 64%)

1     2346 ( 36%)

Class attribute:  annual-pay

Classes to Clusters:

   0   1  <-- assigned to cluster

 2828 2138 |  <=50K

 1338  208 |  >50K
```

Cluster 0 <-- >50K

Cluster 1 <-- <=50K

Incorrectly clustered instances : 3036.0 46.6216 %

Euclidian distance was used as distance function and k was set to two to find that which attributes are strongly correlated. First column gives overall population centroid while second and third columns gives centroids for cluster 0 and 1. Each row gives the centroid coordinate for the specific dimension.

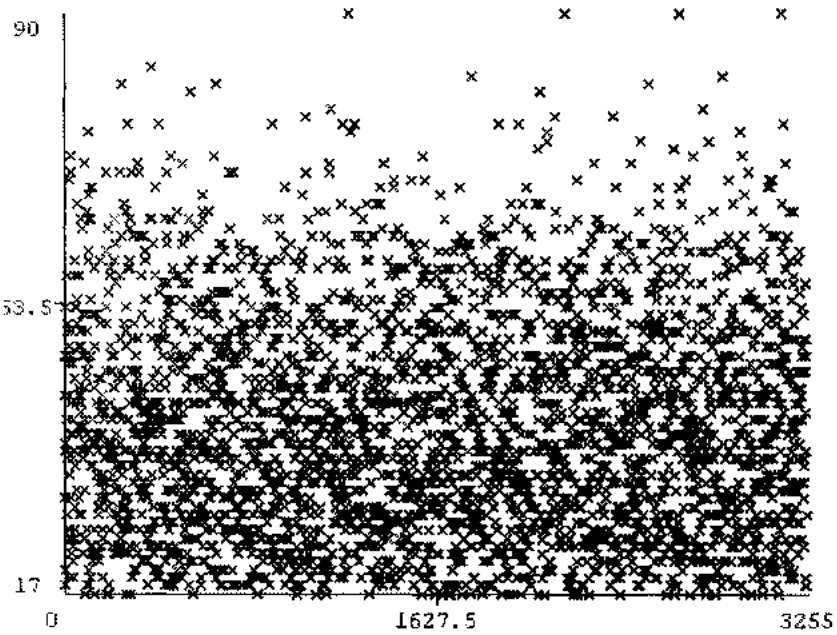


**Figure 4.3**

Figure 4.3 shows the distribution of clusters. X-axis contains number of instances and y-axis contains attribute age. In this way by exploring different clusters data correlation can be analyzed.

## 4.5 Association Rule Mining

Apriori algorithm was used to mine association rules. Apriori works only with categorical attributes. In preprocessing step discretization was performed to convert numeric attributes into nominal. After the analysis of data, association rules are mined. By entering user interestingness measures support=50% and confidence=70%. More than

1500 rules were selected to find the correlation between the attributes. A sample of few rules can be viewed in the following section while the rest of rules are in appendix B.

1. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3772 ==> class= <=50K 3767 conf:(1)

2. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3425 ==> class= <=50K 3420 conf:(1)

3. age='(-inf-24.3]' workclass= Private marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3270 ==> class= <=50K 3265 conf:(1)

4. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 3496 ==> class= <=50K 3490 conf:(1)

5. age='(-inf-24.3]' workclass= Private marital-status= Never-married race= White capital-gain='(-inf-9999.9]' 3331 ==> class= <=50K 3325 conf:(1)

6. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' 3844 ==> class= <=50K 3837 conf:(1)

7. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3827 ==> class= <=50K 3820 conf:(1)

8. age='(-inf-24.3]' workclass= Private marital-status= Never-married race= White capital-loss='(-inf-435.6]' 3273 ==> class= <=50K 3267 conf:(1)

9. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 4356 ==> class= <=50K 4348 conf:(1)

10. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4779 ==> class= <=50K 4770 conf:(1)

From the generated rules it can be assumed that marital status, education, occupation and age are most decisive attributes to predict income. It can be concluded that if a person is married and has many years of education earns more than 50k. And if a person studied less than 8 years, is younger than 25 years and is not married makes <50k. It can be concluded that relationship is the most important attribute to define a person's annual pay. Second most important attribute is marital status. Also sex, age, occupation, hours-per –week and capital gain are important attributes to define individual's annual pay more than 50k. Decisions can be made that if a person is more experienced (age), works more hours every week, has studied longer also has high occupation is expected to get >50k.

## 4.6 Data Modification

In our scenario attribute age and native country are assumed to be more sensitive as these two attributes were the most decisive. So when noise was added to these attributes data was modified and its extraction was not possible. An instance filter to change the percentage of given attributes was used. Attribute should be nominal before applying the filter. Following parameters were selected to apply the filter

attributeIndex →attribute index which is to be changed

percent→noise percentage which is to be added in data

randomSeed →Random number seed

useMissing → Flag to set if missing values are used.

Table 4.4 describes age attribute, which was changed from its current values to one of other possible ones. Noise percent parameter was set to 20%.

**Table 4.4: Attribute Age after Adding Noise**

| No | Label | Count |
|---|---|---|
| 1 | -inf-21.5 | 3198 |
| 2 | 21.5-23.5 | 1865 |
| 3 | 23.5-27.5 | 3285 |
| 4 | 27.5-29.5 | 1872 |
| 5 | 29.5-35.5 | 5041 |
| 6 | 35.5-43.5 | 6199 |
| 7 | 43.5-54.5 | 6251 |
| 8 | 54.5-61.5 | 2655 |
| 9 | 61.5-71.5 | 2195 |
| 10 | 61.5-inf | 1190 |

A sample of 10 rules are listed below while the rest of the results can be viewed in Appendix C.

1. age='(35.5-43.5]' marital-status= Never-married capital-loss='(-inf-368.3]' 444 ==> class= >50K 444 conf:(1)

2. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 444 ==> class= <=50K 444 conf:(1)

3. age='(-inf-24.3]' marital-status= Never-married capital-loss='(-inf-368.3]' native-country= United-States 412 ==> class= <=50K 412 conf:(1)

4. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 412 ==> class= <=50K 412 conf:(1)

5. age='(-inf-24.3]' marital-status= Never-married race= White capital-loss='(-inf-368.3]' 385 ==> class= <=50K 385 conf:(1)

6. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 385 ==> class= <=50K 385 conf:(1)

7. age=('35.5-43.5]' marital-status= Never-married race= White capital-loss='(-inf-368.3]' native-country= United-States 361 ==> class= >50K 361 conf:(1)

8. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 361 ==> class= <=50K 361 conf:(1)

9. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-loss='(-inf-368.3]' 352 ==> class= <=50K 352 conf:(1)

10. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 352 ==> class= <=50K 352 conf:(1)

Adding noise added new values in selected attribute. Association rules generated after adding noise gives a new meaning to make decisions about data. It was concluded that a person whose age is less than 25, who studied less than 10 years and is never married earns <50k (based on 500 generated rules). By adding noise attribute age contains new values and by analyzing the rules it can be concluded that a person whose age is less than 25 can earn more than 50k.

# Chapter 5

# Conclusion & Future Work

# 5. Conclusion and Future Work

In this work exploratory data analysis is performed to get insight of the dataset. Association rules are extracted to find out correlation between the attributes. A data hiding approach is adopted to change a percentage of sensitive attributes so that information cannot be extracted later on. To understand data, and to determine which attribute is most decisive to determine a person's pay classification and clustering has been performed while association mining has been performed to extract most correlated attributes.

Proposed methodology has been implemented on adult data set. From the generated rules marital states,age, education and occupation were found to be most decisive attributes .It can be concluded that relationship is the most important attribute to define a person's annual pay. The second most important attribute is marital status. Also sex, age, occupation, hours-per–week and capital gain are important attributes to define individual's annual pay more than 50k. Decisions can be made that if a person is more experienced (age), works more hours every week, has studied longer also has high occupation is expected to get greater than 50k.

A number of other interestingness measures can be used for patterns discovery including diversity, novelty, conciseness, peculiarity, generality etc. These measures can be used in different scenarios either independently or can be correlated with each other accordingly.

In this work, data is modified after the mining process. Another work can be to hide association rules after mining process which results in $R_H \subseteq R$ such that a subset $R_H$ is considered as sensitive if a certain rule in this subset can't be made public.

# References

# References

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H., & Verkamo, A. I. (1996). Fast Discovery of Association Rules. *Advances in knowledge discovery and data mining, 12*(1), 307-328.

Atallah, M., Bertino, E., Elmagarmid, A., Ibrahim, M., & Verykios, V. (1999).Disclosure limitation of sensitive rules. In *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on* (pp. 45-52). IEEE.

Belwal, R. C., Varshney, J., Khan, S. A., Sharma, A., & Bhattacharya, M. (2008, October). Hiding sensitive association rules efficiently by introducing new variable hiding counter. In *Service Operations and Logistics, and Informatics, 2008. IEEE/SOLI 2008. IEEE International Conference on* (Vol. 1, pp. 130-134). IEEE.

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

Han,J., & Kamber, M.(2003), Mining association rules in large databases, ," in *Data Mining Concepts and Techniques*. (3rd ed.).(pp.21). India: Elsevier.

Han,J., & Kamber, M.(2003), Mining association rules in large databases, ," in *Data Mining Concepts and Techniques*. (3rd ed.).(pp.28). India: Elsevier.

Kamber, M., Han, J., & Chiang, J. (1997, August). Metarule-Guided Mining of Multi-Dimensional Association Rules Using Data Cubes. In *KDD* (Vol. 97, p. 207).

Kaya, M., & Alhajj, R. (2005). Fuzzy OLAP association rules mining-based modular reinforcement learning approach for multiagent systems. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on, 35*(2), 326-338.

Khare, N., Adlakha, N., & Pardasani, K. R. (2010, March). An Algorithm for Mining Multidimensional Association Rules using Boolean Matrix. In *Recent Trends in*

*Information, Telecommunication and Computing (ITC), 2010 International Conference on* (pp. 95-99). IEEE.

Lakshmi. N.V.M. & Rani. K.S. (2012, Feb.-Mar.). A novel method for finding privacy preserving association rule mining . *Indian Journal of Computer Science and Engineering (IJCSE)* .3(1), pp.104-113.

Liu, H., & Wang, B. (2007). An association rule mining algorithm based on a Boolean matrix. *Data Science Journal, 6*, pp.559-565.

Messaoud, R. B., Rabaséda, S. L., Boussaid, O., & Missaoui, R. (2006, November). Enhanced mining of association rules from data cubes. In*Proceedings of the 9th ACM international workshop on Data warehousing and OLAP* (pp. 11-18). ACM.

Nestorov, S., & Jukic, N. (2003, January). Ad-hoc association-rule mining within the data warehouse. In *System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on* (pp. 10-pp). IEEE.

Patidar. V., Shrivastava. V., & Shrivastava. V. (2013).A generalized association rule based method for Privacy Preserving in Data Mining. *International Journal of Advanced Research in Computer Science and Software Engineerin.* 3(9), pp.703-707.

Platt, J. (1998). Sequential minimal optimization: A fast algorithm for training support vector machines.

Pontikakis, E. D., Theodoridis, Y., Tsitsonis, A. A., Chang, L., & Verykios, V. S. (2004, October). A quantitative and qualitative analysis of blocking in association rule hiding. In *Proceedings of the 2004 ACM workshop on Privacy in the electronic society* (pp. 29-30). ACM.

Shah. K., Thakkar. A. & Gantra. A. (2012, Feb.).A study on association rule hiding approaches . *International Journal of Engineering and Advanced Technology.*1(3), pp. 72-76.

Usama M. F., Keki B. I (1993). Multi-interval discretization of continuous valued attributes for classification learning. In: Thirteenth International Joint Conference on Artificial Intelligence, pp.1022-1027.

Usman, M., Pears, R., & Fong, A. C. M. (2013a). A data mining approach to knowledge discovery from multidimensional cube structures. *Knowledge-Based Systems, 40,* 36-49.

Usman, M., Pears, R., & Fong, A. C. M. (2013b). Discovering diverse association rules from multidimensional schema. *Expert Systems with Applications, 40*(15), 5975-5996.

Wang. E. T., Lee. G. & Lin.Y.T., A novel method for protecting sensitive knowledge in association rules mining". Ph.D. thesis , National Dong Hwa University Hualien, Taiwan.

# Appendix A

# Appendix A. Normalized Attribute List

```
+    0.0858 * (normalized)  fnlwg='(-inf-159527]'

+    0.1475 * (normalized)  fnlwg='(159527-306769]'

+    0.2109 * (normalized)  fnlwg='(306769-454011]'

+   -0.1779 * (normalized)  fnlwg='(454011-601253]'

+    1      * (normalized)  fnlwg='(601253-748495]'

+   -1.2663 * (normalized)  fnlwg='(748495-895737]'

+    0      * (normalized)  fnlwg='(1337463-inf)'

+    0.3972 * (normalized)  education= Bachelors

+   -0.1335 * (normalized)  education= HS-grad

+   -0.136  * (normalized)  education= 11th

+    0.3724 * (normalized)  education= Masters

+   -0.1567 * (normalized)  education= 9th

+    0.1041 * (normalized)  education= Some-college

+   -0.0486 * (normalized)  education= Assoc-acdm

+    0.1264 * (normalized)  education= Assoc-voc

+   -0.2463 * (normalized)  education= 7th-8th

+    0.3644 * (normalized)  education= Doctorate

+    0.3831 * (normalized)  education= Prof-school

+   -0.1982 * (normalized)  education= 5th-6th

+   -0.0695 * (normalized)  education= 10th

+   -0.3    * (normalized)  education= 1st-4th

+    0      * (normalized)  education= Preschool

+   -0.4589 * (normalized)  education= 12th
+   -0.3    * (normalized)  education-num='(-inf-2.5]'
+   -0.4445 * (normalized)  education-num='(2.5-4]'
+   -0.1567 * (normalized)  education-num='(4-5.5]'
+   -0.2055 * (normalized)  education-num='(5.5-7]'
+   -0.4589 * (normalized)  education-num='(7-8.5]'
+   -0.0294 * (normalized)  education-num='(8.5-10]'
+    0.1264 * (normalized)  education-num='(10-11.5]'
+    0.3486 * (normalized)  education-num='(11.5-13]'
+    0.3724 * (normalized)  education-num='(13-14.5]'
+    0.7475 * (normalized)  education-num='(14.5-inf)'
+   -0.4027 * (normalized)  marital-status= Never-married
+    0.6342 * (normalized)  marital-status= Married-civ-spouse
+   -0.353  * (normalized)  marital-status= Divorced
+   -0.6427 * (normalized)  marital-status= Married-spouse-absent
+    0.1216 * (normalized)  marital-status= Separated
+    0.6451 * (normalized)  marital-status= Married-AF-spouse
+   -0.0026 * (normalized)  marital-status= Widowed
+    0.2024 * (normalized)  occupation= Adm-clerical
+    0.6759 * (normalized)  occupation= Exec-managerial
+   -0.2511 * (normalized)  occupation= Handlers-cleaners
+    0.6017 * (normalized)  occupation= Prof-specialty
+   -0.5681 * (normalized)  occupation= Other-service
+    0.1248 * (normalized)  occupation= Sales
+   -0.1378 * (normalized)  occupation= Craft-repair
+   -0.2502 * (normalized)  occupation= Transport-moving
+   -0.6134 * (normalized)  occupation= Farming-fishing
+   -0.2887 * (normalized)  occupation= Machine-op-inspct
+    0.6627 * (normalized)  occupation= Tech-support
+   -0.1018 * (normalized)  occupation= Armed-Forces
+    0.7017 * (normalized)  occupation= Protective-serv
+   -0.7581 * (normalized)  occupation= Priv-house-serv
+   -0.2606 * (normalized)  relationship= Not-in-family
+    0.1652 * (normalized)  relationship= Husband
+    0.8302 * (normalized)  relationship= Wife
+   -0.2961 * (normalized)  relationship= Own-child
+   -0.4483 * (normalized)  relationship= Unmarried
+    0.0097 * (normalized)  relationship= Other-relative
+    0.0868 * (normalized)  race= White
+    0.2765 * (normalized)  race= Black
+    0.1038 * (normalized)  race= Asian-Pac-Islander
+   -0.6126 * (normalized)  race= Amer-Indian-Eskimo
+    0.1456 * (normalized)  race= Other
+   -0.3348 * (normalized)  sex
+   -2.6577 * (normalized)  capital-gain='(-inf-9999.9]'
+    1.4281 * (normalized)  capital-gain='(9999.9-19999.8]'
+    0.2297 * (normalized)  capital-gain='(19999.8-29999.7]'
+    1     * (normalized)  capital-gain='(89999.1-inf)'
+   -0.6001 * (normalized)  capital-loss='(-inf-435.6]'
+    0     * (normalized)  capital-loss='(435.6-871.2]'
+   -1.963  * (normalized)  capital-loss='(871.2-1306.8]'
+   -1     * (normalized)  capital-loss='(1306.8-1742.4]'
+    0.5724 * (normalized)  capital-loss='(1742.4-2178]'
+    0.6678 * (normalized)  capital-loss='(2178-2613.6]'
+    2      * (normalized)  capital-loss='(2613.6-3049.2]'
+    0.3229 * (normalized)  capital-loss='(3484.8-3920.4]'
+   -0.768  * (normalized)  hours-per-week='(-inf-10.8]'
+   -0.5934 * (normalized)  hours-per-week='(10.8-20.6]'
+   -0.415  * (normalized)  hours-per-week='(20.6-30.4]'
+    0.0147 * (normalized)  hours-per-week='(30.4-40.2]'
+    0.279  * (normalized)  hours-per-week='(40.2-50]'
+    0.3023 * (normalized)  hours-per-week='(50-59.8]'
+    0.399  * (normalized)  hours-per-week='(59.8-69.6]'
+    0.1838 * (normalized)  hours-per-week='(69.6-79.4]'
+    0.8139 * (normalized)  hours-per-week='(79.4-89.2]'
+   -0.2162 * (normalized)  hours-per-week='(89.2-inf)'
+    0.1283 * (normalized)  native-country= United-States
+    0      * (normalized)  native-country= Cuba
+   -0.1278 * (normalized)  native-country= Jamaica
+    0.1772 * (normalized)  native-country= India
+    0.1273 * (normalized)  native-country= Mexico
+   -1.1561 * (normalized)  native-country= South
+   -0.0911 * (normalized)  native-country= Puerto-Rico
+    1.0822 * (normalized)  native-country= England
+    0.7095 * (normalized)  native-country= Canada
+    0.9046 * (normalized)  native-country= Germany
+   -0.567  * (normalized)  native-country= Iran
+    0.7603 * (normalized)  native-country= Philippines
+    0.5001 * (normalized)  native-country= Italy
+    0      * (normalized)  native-country= Poland
+   -0.1855 * (normalized)  native-country= Columbia
+    0      * (normalized)  native-country= Cambodia
+   -0.1877 * (normalized)  native-country= Thailand
+    0      * (normalized)  native-country= Ecuador
+    0      * (normalized)  native-country= Laos
+   -0.9661 * (normalized)  native-country= Taiwan
+   -0.2176 * (normalized)  native-country= Haiti
+    0.3544 * (normalized)  native-country= Portugal
+   -0.8676 * (normalized)  native-country= Dominican-Republic
+    0      * (normalized)  native-country= El-Salvador
+    0      * (normalized)  native-country= Guatemala
+    0.1123 * (normalized)  native-country= China
+    0.1254 * (normalized)  native-country= Japan
+   -0.2755 * (normalized)  native-country= Peru
+    0      * (normalized)  native-country= Outlying-US(Guam-USVI-etc)
+    0      * (normalized)  native-country= Scotland
+    1      * (normalized)  native-country= Trinadad&Tobago
+   -0.397  * (normalized)  native-country= Greece
+   -0.2867 * (normalized)  native-country= Nicaragua
+   -1      * (normalized)  native-country= Vietnam
+    0.4935 * (normalized)  native-country= Ireland
```

# Appendix B

# Appendix B. Association Rules

11. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4112 ==> class= <=50K 4104 conf:(1)

12. age='(-inf-24.3]' workclass= Private marital-status= Never-married race= White 3334 ==> class= <=50K 3327 conf:(1)

13. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 4453 ==> class= <=50K 4443 conf:(1)

14. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' native-country= United-States 3906 ==> class= <=50K 3897 conf:(1)

15. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-loss='(-inf-435.6]' native-country= United-States 3430 ==> class= <=50K 3422 conf:(1)

16. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-loss='(-inf-435.6]' 3778 ==> class= <=50K 3769 conf:(1)

17. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' 4192 ==> class= <=50K 4182 conf:(1)

18. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' 4880 ==> class= <=50K 4868 conf:(1)

19. age='(-inf-24.3]' workclass= Private marital-status= Never-married native-country= United-States 3501 ==> class= <=50K 3492 conf:(1)

20. age='(-inf-24.3]' marital-status= Never-married race= White capital-loss='(-inf-435.6]' native-country= United-States 3833 ==> class= <=50K 3823 conf:(1)

21. age='(-inf-24.3]' workclass= Private marital-status= Never-married 3850 ==> class= <=50K 3839 conf:(1)

22. age='(-inf-24.3]' marital-status= Never-married race= White capital-loss='(-inf-435.6]' 4119 ==> class= <=50K 4107 conf:(1)

23. age='(-inf-24.3]' marital-status= Never-married race= White native-country= United-States 3912 ==> class= <=50K 3900 conf:(1)

24. age='(-inf-24.3]' marital-status= Never-married race= White 4199 ==> class= <=50K 4185 conf:(1)

25. age='(-inf-24.3]' marital-status= Never-married capital-loss='(-inf-435.6]' native-country= United-States 4367 ==> class= <=50K 4352 conf:(1)

26. age='(-inf-24.3]' marital-status= Never-married native-country= United-States 4464 ==> class= <=50K 4447 conf:(1)

27. workclass= Private marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3346 ==> class= <=50K 3333 conf:(1)

28. age='(-inf-24.3]' marital-status= Never-married capital-loss='(-inf-435.6]' 4793 ==> class= <=50K 4774 conf:(1)

29. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 4061 ==> class= <=50K 4044 conf:(1)

30. age='(-inf-24.3]' marital-status= Never-married 4894 ==> class= <=50K 4872 conf:(1)

31. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4375 ==> class= >=50K 4355 conf:(1)

32. workclass= Private marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' 3420 ==> class= >=50K 3404 conf:(1)

33. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3500 ==> class= <=50K 3483 conf:(1)

34. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3696 ==> class= <=50K 3678 conf:(1)

35. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 4152 ==> class= <=50K 4131 conf:(0.99)

36. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' 4472 ==> class= <=50K 4448 conf:(0.99)

37. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' 3771 ==> class= >=50K 3749 conf:(0.99)

38. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' native-country= United-States 3573 ==> class= <=50K 3552 conf:(0.99)

39. marital-status= Never-married relationship= Own-child capital-loss='(-inf-435.6]' native-country= United-States 4072 ==> class= <=50K 4047 conf:(0.99)

40. workclass= Private marital-status= Never-married relationship= Own-child capital-loss='(-inf-435.6]' 3357 ==> class= <=50K 3335 conf:(0.99)

41. marital-status= Never-married relationship= Own-child race= White capital-loss='(-inf-435.6]' 3706 ==> class= <=50K 3681 conf:(0.99)

42. marital-status= Never-married relationship= Own-child capital-loss='(-inf-435.6]' 4388 ==> class= <=50K 4358 conf:(0.99)

43. marital-status= Never-married relationship= Own-child race= White capital-loss='(-inf-435.6]' native-country= United-States 3510 ==> class= <=50K 3486 conf:(0.99)

44. age='(-inf-24.3]' education-num='(8.5-10]' capital-gain='(-inf-9999.9]' native-country= United-States 3347 ==> class= <=50K 3324 conf:(0.99)

45. marital-status= Never-married relationship= Own-child native-country= United-States 4163 ==> class= <=50K 4134 conf:(0.99)

46. age='(-inf-24.3]' education-num='(8.5-10]' capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3551 ==> class= <=50K 3526 conf:(0.99)

47. age='(-inf-24.3]' education-num='(8.5-10]' capital-gain='(-inf-9999.9]' 3628 ==> class= <=50K 3602 conf:(0.99)

48. workclass= Private marital-status= Never-married relationship= Own-child 3431 ==> class= <=50K 3406 conf:(0.99)

49. marital-status= Never-married relationship= Own-child 4485 ==> class= <=50K 4451 conf:(0.99)

50. marital-status= Never-married relationship= Own-child race= White 3781 ==> class= >=50K 3752 conf:(0.99)

51. marital-status= Never-married relationship= Own-child race= White native-country= United-States 3583 ==> class= <=50K 3555 conf:(0.99)

52. age='(-inf-24.3]' capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 4932 ==> class= >=50K 4893 conf:(0.99)

53. age='(-inf-24.3]' education-num='(8.5-10]' native-country= United-States 3354 ==> class= >=50K 3327 conf:(0.99)

54. age='(-inf-24.3]' capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 5436 ==> class= >=50K 5392 conf:(0.99)

55. age='(-inf-24.3]' workclass= Private capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3896 ==> class= <=50K 3864 conf:(0.99)

56. age='(-inf-24.3]' race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 4333 ==> class= <=50K 4297 conf:(0.99)

57. age='(-inf-24.3]' workclass= Private capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4312 ==> class= <=50K 4276 conf:(0.99)

58. age='(-inf-24.3]' capital-gain='(-inf-9999.9]' native-country= United-States 5044 ==> class= <=50K 5001 conf:(0.99)

59. age='(-inf-24.3]' workclass= Private capital-gain='(-inf-9999.9]' native-country= United-States 3979 ==> class= <=50K 3945 conf:(0.99)

60. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3478 ==> class= <=50K 3448 conf:(0.99)

61. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3449 ==> class= <=50K 3419 conf:(0.99)

62. age='(-inf-24.3]' education-num='(8.5-10]' capital-loss='(-inf-435.6]' 3560 ==> class= <=50K 3529 conf:(0.99)

63. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3766 ==> class= <=50K 3733 conf:(0.99)

64. age='(-inf-24.3]' race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4677 ==> class= <=50K 4636 conf:(0.99)

65. age='(-inf-24.3]' education-num='(8.5-10]' 3637 ==> class= <=50K 3605 conf:(0.99)

66. age='(-inf-24.3]' workclass= Private capital-loss='(-inf-435.6]' native-country= United-States 3901 ==> class= >50K 3866 conf:(0.99)

67. age='(-inf-24.3]' workclass= Private race= White capital-loss='(-inf-435.6]' native-country= United-States 3452 ==> class= <=50K 3421 conf:(0.99)

68. age='(-inf-24.3]' capital-gain='(-inf-9999.9]' 5555 ==> class= <=50K 5505 conf:(0.99)

69. age='(-inf-24.3]' race= White capital-gain='(-inf-9999.9]' native-country= United-States 4427 ==> class= >=50K 4387 conf:(0.99)

70. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3743 ==> class= >=50K 3709 conf:(0.99)

71. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' native-country= United-States 3522 ==> class= <=50K 3490 conf:(0.99)

72. age='(-inf-24.3]' workclass= Private capital-gain='(-inf-9999.9]' 4398 ==> class= >=50K 4358 conf:(0.99)

73. age='(-inf-24.3]' race= White capital-loss='(-inf-435.6]' native-country= United-States 4340 ==> class= <=50K 4300 conf:(0.99)

74. age='(-inf-24.3]' workclass= Private capital-loss='(-inf-435.6]' 4318 ==> class= >=50K 4278 conf:(0.99)

75. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 3561 ==> class= <=50K 3528 conf:(0.99)

76. age='(-inf-24.3]' workclass= Private native-country= United-States 3984 ==> class= <=50K 3947 conf:(0.99)

77. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' 3854 ==> class= <=50K 3818 conf:(0.99)

78. age='(-inf-24.3]' workclass= Private race= White capital-loss='(-inf-435.6]' 3746 ==> class= <=50K 3711 conf:(0.99)

79. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4600 ==> class= >=50K 4557 conf:(0.99)

80. age='(-inf-24.3]' workclass= Private race= White native-country= United-States 3525 ==> class= <=50K 3492 conf:(0.99)

81. age='(-inf-24.3]' race= White capital-gain='(-inf-9999.9]' 4773 ==> class= <=50K 4728 conf:(0.99)

82. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' 3817 ==> class= >=50K 3781 conf:(0.99)

83. age='(-inf-24.3]' capital-loss='(-inf-435.6]' native-country= United-States 4944 ==> class= <=50K 4897 conf:(0.99)

84. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 4192 ==> class= <=50K 4152 conf:(0.99)

85. age='(-inf-24.3]' workclass= Private race= White 3820 ==> class= <=50K 3783 conf:(0.99)

86. age='(-inf-24.3]' race= White capital-loss='(-inf-435.6]' 4685 ==> class= >=50K 4639 conf:(0.99)

87. relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 4570 ==> class= <=50K 4525 conf:(0.99)

88. age='(-inf-24.3]' race= White native-country= United-States 4434 ==> class= <=50K 4390 conf:(0.99)

89. age='(-inf-24.3]' workclass= Private 4404 ==> class= <=50K 4360 conf:(0.99)

90. age='(-inf-24.3]' native-country= United-States 5056 ==> class= <=50K 5005 conf:(0.99)

91. age='(-inf-24.3]' capital-loss='(-inf-435.6]' 5451 ==> class= <=50K 5396 conf:(0.99)

92. education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' hours-per-week='(30.4-40.2]' 3344 ==> class= <=50K 3310 conf:(0.99)

93. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 4294 ==> class= <=50K 4250 conf:(0.99)

94. relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 4937 ==> class= <=50K 4886 conf:(0.99)

95. workclass= Private education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' 3765 ==> class= <=50K 3726 conf:(0.99)

96. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' 4713 ==> class= <=50K 4664 conf:(0.99)

97. age='(-inf-24.3]' race= White 4781 ==> class= <=50K 4731 conf:(0.99)

98. relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 4678 ==> class= <=50K 4629 conf:(0.99)

99. workclass= Private education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-435.6]' native-country= United-States 3520 ==> class= <=50K 3483 conf:(0.99)

100. relationship= Own-child capital-gain='(-inf-9999.9]' 5053 ==> class= <=50K 4998 conf:(0.99)

# Appendix C

# Appendix C. Association Rules after Data Modification

11. age='(-inf-24.3]' workclass= Private marital-status= Never-married native-country= United-States 331 ==> class= <=50K 331 conf:(1)

12. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 331 ==> class= <=50K 331 conf:(1)

13. age='(-inf-24.3]' marital-status= Never-married native-country= United-States 421 ==> class= <=50K 420 conf:(1)

14. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 421 ==> class= <=50K 420 conf:(1)

15. age='(-inf-24.3]' marital-status= Never-married race= White 390 ==> class= <=50K 389 conf:(1)

16. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' 390 ==> class= >=50K 389 conf:(1)

17. age='(-inf-24.3]' marital-status= Never-married race= White native-country= United-States 366 ==> class= >=50K 365 conf:(1)

18. age='(-inf-24.3]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' native-country= United-States 366 ==> class= <=50K 365 conf:(1)

19. age='(-inf-24.3]' workclass= Private marital-status= Never-married 360 ==> class= <=50K 358 conf:(0.99)

20. age='(-inf-24.3]' workclass= Private marital-status= Never-married capital-gain='(-inf-9999.9]' 360 ==> class= >=50K 358 conf:(0.99)

21. age='(-inf-24.3]' marital-status= Never-married 456 ==> class= <=50K 453 conf:(0.99)

22. age='(-inf-24.3]' marital-status= Never-married capital-gain='(-inf-9999.9]' 456 ==> class= >50K 453 conf:(0.99)

23. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 431 ==> class= <=50K 428 conf:(0.99)

24. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 388 ==> class= <=50K 385 conf:(0.99)

25. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 372 ==> class= <=50K 369 conf:(0.99)

26. relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 468 ==> class= >=50K 464 conf:(0.99)

27. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 458 ==> class= <=50K 454 conf:(0.99)

28. age='(-inf-24.3]' workclass= Private race= White native-country= United-States 343 ==> class= <=50K 340 conf:(0.99)

29. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' native-country= United-States 343 ==> class= >=50K 340 conf:(0.99)

30. workclass= Private marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 342 ==> class= <=50K 339 conf:(0.99)

31. age='(-inf-24.3]' workclass= Private race= White capital-loss='(-inf-368.3]' native-country= United-States 336 ==> class= >=50K 333 conf:(0.99)

32. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 336 ==> class= <=50K 333 conf:(0.99)

33. workclass= Private marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 336 ==> class= <=50K 333 conf:(0.99)

34. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 439 ==> class= <=50K 435 conf:(0.99)

35. marital-status= Never-married relationship= Own-child capital-loss='(-inf-368.3]' native-country= United-States 432 ==> class= >50K 428 conf:(0.99)

36. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 427 ==> class= >=50K 423 conf:(0.99)

37. relationship= Own-child race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 419 ==> class= <=50K 415 conf:(0.99)

38. relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 499 ==> class= >=50K 494 conf:(0.99)

39. relationship= Own-child race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 399 ==> class= <=50K 395 conf:(0.99)

40. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' 394 ==> class= <=50K 390 conf:(0.99)

41. marital-status= Never-married relationship= Own-child race= White capital-loss='(-inf-368.3]' 389 ==> class= >=50K 385 conf:(0.99)

42. relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 478 ==> class= <=50K 473 conf:(0.99)

43. marital-status= Never-married relationship= Own-child race= White capital-gain='(-inf-9999.9]' native-country= United-States 377 ==> class= <=50K 373 conf:(0.99)

44. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' native-country= United-States 376 ==> class= <=50K 372 conf:(0.99)

45. marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' 468 ==> class= <=50K 463 conf:(0.99)

46. marital-status= Never-married relationship= Own-child race= White capital-loss='(-inf-368.3]' native-country= United-States 373 ==> class= <=50K 369 conf:(0.99)

47. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 368 ==> class= <=50K 364 conf:(0.99)

48. marital-status= Never-married relationship= Own-child capital-loss='(-inf-368.3]' 459 ==> class= <=50K 454 conf:(0.99)

49. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 459 ==> class= >=50K 454 conf:(0.99)

50. age='(-inf-24.3]' workclass= Private race= White 367 ==> class= <=50K 363 conf:(0.99)

51. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' 367 ==> class= <=50K 363 conf:(0.99)

52. workclass= Private marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' 362 ==> class= >=50K 358 conf:(0.99)

53. education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 542 ==> class= <=50K 536 conf:(0.99)

54. age='(-inf-24.3]' workclass= Private race= White capital-loss='(-inf-368.3]' 360 ==> class= >=50K 356 conf:(0.99)

55. age='(-inf-24.3]' workclass= Private race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 360 ==> class= <=50K 356 conf:(0.99)

56. workclass= Private marital-status= Never-married relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 355 ==> class= <=50K 351 conf:(0.99)

57. marital-status= Never-married relationship= Own-child native-country= United-States 440 ==> class= <=50K 435 conf:(0.99)

58. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 440 ==> class= <=50K 435 conf:(0.99)

59. workclass= Private education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 346 ==> class= <=50K 342 conf:(0.99)

60. workclass= Private marital-status= Never-married relationship= Own-child native-country= United-States 343 ==> class= >=50K 339 conf:(0.99)

61. relationship= Own-child race= White capital-gain='(-inf-9999.9]' 427 ==> class= <=50K 422 conf:(0.99)

62. relationship= Own-child capital-gain='(-inf-9999.9]' 511 ==> class= <=50K 505 conf:(0.99)

63. age='(-inf-24.3]' education-num='(8.5-10]' 339 ==> class= <=50K 335 conf:(0.99)

64. age='(-inf-24.3]' education-num='(8.5-10]' capital-gain='(-inf-9999.9]' 339 ==> class= >=50K 335 conf:(0.99)

65. workclass= Private marital-status= Never-married relationship= Own-child capital-loss='(-inf-368.3]' native-country= United-States 337 ==> class= >=50K 333 conf:(0.99)

66. education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 581 ==> class= >=50K 574 conf:(0.99)

67. age='(-inf-24.3]' education-num='(8.5-10]' capital-loss='(-inf-368.3]' 332 ==> class= >=50K 328 conf:(0.99)

68. age='(-inf-24.3]' education-num='(8.5-10]' capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 332 ==> class= >=50K 328 conf:(0.99)

69. workclass= Private relationship= Own-child race= White capital-gain='(-inf-9999.9]' 331 ==> class= <=50K 327 conf:(0.99)

70. relationship= Own-child race= White capital-gain='(-inf-9999.9]' native-country= United-States 406 ==> class= <=50K 401 conf:(0.99)

71. education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' native-country= United-States 557 ==> class= <=50K 550 conf:(0.99)

72. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' 397 ==> class= <=50K 392 conf:(0.99)

73. marital-status= Never-married relationship= Own-child race= White 395 ==> class= <=50K 390 conf:(0.99)

74. workclass= Private education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' 472 ==> class= <=50K 466 conf:(0.99)

75. relationship= Own-child capital-loss='(-inf-368.3]' native-country= United-States 470 ==> class= <=50K 464 conf:(0.99)

76. marital-status= Never-married relationship= Own-child 469 ==> class= <=50K 463 conf:(0.99)

77. workclass= Private relationship= Own-child capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 388 ==> class= <=50K 383 conf:(0.99)

78. marital-status= Never-married relationship= Own-child race= White native-country= United-States 378 ==> class= <=50K 373 conf:(0.99)

79. education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' 597 ==> class= <=50K 589 conf:(0.99)

80. education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 445 ==> class= <=50K 439 conf:(0.99)

81. workclass= Private education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 364 ==> class= <=50K 359 conf:(0.99)

82. workclass= Private marital-status= Never-married relationship= Own-child 363 ==> class= <=50K 358 conf:(0.99)

83. relationship= Own-child capital-loss='(-inf-368.3]' 501 ==> class= <=50K 494 conf:(0.99)

84. workclass= Private marital-status= Never-married relationship= Own-child capital-loss='(-inf-368.3]' 356 ==> class= <=50K 351 conf:(0.99)

85. workclass= Private education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' native-country= United-States 355 ==> class= <=50K 350 conf:(0.99)

86. age='(-inf-24.3]' race= White capital-loss='(-inf-368.3]' native-country= United-States 422 ==> class= <=50K 416 conf:(0.99)

87. age='(-inf-24.3]' race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 422 ==> class= <=50K 416 conf:(0.99)

88. relationship= Own-child race= White capital-loss='(-inf-368.3]' 421 ==> class= <=50K 415 conf:(0.99)

89. relationship= Own-child native-country= United-States 480 ==> class= <=50K 473 conf:(0.99)

90. relationship= Own-child race= White capital-loss='(-inf-368.3]' native-country= United-States 401 ==> class= >=50K 395 conf:(0.99)

91. education-num='(8.5-10]' marital-status= Never-married capital-gain='(-inf-9999.9]' hours-per-week='(30.4-40.2]' 333 ==> class= <=50K 328 conf:(0.98)

92. education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 465 ==> class= <=50K 458 conf:(0.98)

93. age='(-inf-24.3]' workclass= Private native-country= United-States 391 ==> class= <=50K 385 conf:(0.98)

94. age='(-inf-24.3]' workclass= Private capital-gain='(-inf-9999.9]' native-country= United-States 391 ==> class= >=50K 385 conf:(0.98)

95. education-num='(8.5-10]' marital-status= Never-married race= White capital-gain='(-inf-9999.9]' native-country= United-States 455 ==> class= <=50K 448 conf:(0.98)

96. age='(-inf-24.3]' race= White capital-loss='(-inf-368.3]' 451 ==> class= >=50K 444 conf:(0.98)

97. age='(-inf-24.3]' race= White capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' 451 ==> class= <=50K 444 conf:(0.98)

98. relationship= Own-child 513 ==> class= <=50K 505 conf:(0.98)

99. age='(-inf-24.3]' workclass= Private capital-loss='(-inf-368.3]' native-country= United-States 382 ==> class= >=50K 376 conf:(0.98)

100. age='(-inf-24.3]' workclass= Private capital-gain='(-inf-9999.9]' capital-loss='(-inf-368.3]' native-country= United-States 382 ==> class= >=50K 376 conf:(0.98)

# Appendix D

# Appendix D. Resources

1. The book Data Mining Concepts and Techniques by Jiawei Han and Micheline Kamber has website containing many supplemental materials including slide presentations per chapter, instructor manual, supplemental reading lists etc for readers.
   http://www.cs.uiuc.edu/~hanj/bk2/
   http://www.cs.uiuc.edu/~hanj/bk3/
2. Data mining software package is available on the following link. http://illimine.cs.uiuc.edu/software/
3. Additional resources on WEKA, including sample data sets can be found from the official WEKA Web site.
   http://www.cs.waikato.ac.nz/ml/weka/
4. The UCI Machine Learning Repository is a collection of databases, domain theories, and data generators that are used by the machine learning community for the empirical analysis of machine learning algorithms https://archive.ics.uci.edu/ml/datasets.html
5. For information about citing data sets in publications citation policy is given on the following address
   http://archive.ics.uci.edu/ml/citation_policy.html
6. Data mining glossary
   www.twocrows.com/glossary.html
7. Adult dataset can be accessed from the following link
   http://archive.ics.uci.edu/ml/datasets/Adult

# Appendix E

# Appendix E. Glossary

**Relational Database** is a collection of tables, each of which is assigned a unique name. Each table consist a set of attributes (columns or fields) and stores a large set of tuples (records or rows). Each tuple in a relational table represents an object which is identified by a unique key. Relational data can be accessed by database quries written in relational query language, such as SQL, or with the assistance of graphical user interface.

**ER (entity-relationship)** model is a semantic data model and represents the database as a set of entities and their relationships.

A **data warehouse** is a subject-oriented, integrated, time-variant, and nonvolatile collection of data organized in support of management decision making.

A data warehouse is a repository of information collected from multiple sources, stored under a unified schema and usually resides at a single site. Data warehouses are constructed via the process of data cleaning, integration, transformation, loading and periodic data refreshing.

Data warehouse are usually modeled by using multidimensional database structure where each dimension corresponds to an attribute or a set of attributes in the schema. Actual physical structure of a data warehouse may be a relational data store or multidimensional data cube.

Data warehouse collects information about subjects that span an entire organization, and thus its scope is enterprise-wide where as **data mart** is a department subset of a data warehouse and focuses on selected subjects, and this its scope is department-wide.

By providing multidimensional data views and the pre computation of summarized data, data warehouse systems are well suited for on-line analytical processing (OLAP).

**Transactional database** consists of a file where each record represents a transaction. A transaction typically includes a unique transaction identity number (trans_ID) and a list of the items making up the transaction.

**Object-Relational Database** is constructed on an object-relational data model. This model extends the relational model by providing a rich data type for handling complex objects and object orientation.

**Temporal Database** stores relational data that include time-related attributes.

**Sequence Database** stores sequences of ordered events, with or without a concrete notion of time.

**Time-Series Database** stores sequences of values or events obtained over repeated measurements of time.

**Spatial Databases** contain spatial-related information like geographic (map) databases, medical or satellite image databases.

**Spatiotemporal Database** A spatial database that stores spatial objects that change with time is called a spatiotemporal database.

**Text databases** are databases that contain word descriptions for objects containing long sentences or paragraphs, such as product specifications, error reports, warning messages, summary reports etc.

**Multimedia database** stores image, audio and video data.

**Heterogeneous database** consist a set of interconnected, autonomous component databases.

**Legacy database** is a group of heterogeneous databases that combines different kinds of data systems, such as relational or object-oriented databases, hierarchical databases, network databases, spreadsheets, multimedia databases or file systems.

Data mining tasks can be classified into two categories: **descriptive and predictive**. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make predictions.

**Frequent patterns** are the patterns that occur frequently in data; kind of frequent patterns include itemsets, subsequences and substructures.

An itemset with a support greater than a given minimum support is **called frequent itemset**.

Integration of data mining components, as a whole, with a database or data warehouse system can involve either no coupling, loose coupling, semi tight coupling or tight coupling.

**Concept hierarchies** for numerical attributes can be constructed automatically based on data discretization. Methods used include binning, histogram analysis, entropy based discretization, cluster analysis and discretization by intuitive partitioning. Each method assumes that the values to be discretized are sorted in ascending order.

A rule containing more than one distinct predicates is a **multidimensional association rule.**

A rule in which all the predicates have distinct predicate names is called **non repetitive** predicate multi-dimensional association rule.

**Support** determines how often a rule is applicable to a given dataset, while confidence determines how frequently items in B appear in transactions that contain A.

Reporting and online analytical processing (OLAP) are important tools for understanding what happened in the past. Data mining is a process for understanding what will happen in the future.

Data mining doesn't replace statistics. In fact, **statistics** are a good complement to data mining. Traditional statistical techniques, such as regression, are used alongside data mining technologies, such as neural networks. Statistics are also used to validate data mining results.

**Statistical models** are usually used in early stages of data mining project to gain an overview of the structure of the data.

**Propensity models** are usually used for predicting customer behavior.

**Association** is the process of discovering which events occur together, which is contrast to sequence detection, used to discover the order in which the event happened.

**Attitudinal data** relates to personal attitude or opinions and is usually gathered through survey research.

**Attribute** is a property or characteristic of an entity also known as variable or field.

**Field** (also known as variable or attribute) is a space for an individual piece of data or information. For example, one data field may contain a customer's first name. The next data field may contain the customer's last name.

To analyze two categories of data, each category should have equal amount of data to simplify the modeling process called **balanced data**.

**Variable:** Any measured characteristic or **attribute** that differs for different subjects.

**Statistics** is the mathematics of the collection, organization, and interpretation of numerical data.

**Structured data** is the data in traditional numerical format, such as transactional data.

**Test set:** A dataset independent of the **training set**, used to fine-tune the estimates of the model parameters.

**Training set:** A dataset used to estimate or train a model.

**Behavioral data** relates to behavior or actions. This data type is most extensively used in data mining.

**Churn** is the process of customer attrition and is primary source of aggravation of many industries.

**Classification** is the process of identifying group to which an object belongs by examining characteristics of the object. In classification groups are defined by external criterion.

**Clustering** is good for finding natural groupings of cases that have the same characteristics, i-e detect fraud by using clustering to group similar cases of unusual credit card transactions. Clustering divides a dataset so that records with similar content are in the same group, and groups are as different as possible from each other (contrast with **classification**).

**Regression** is the process of discovering and predicting relationships between two or more variables.

**Cross-selling** is the practice of offering and selling additional products or services to existing customers.

**Data warehouse** is the database in which data is collected and stored for analysis.

**Decision trees** are graphical (tree-like) displays that show segments, patterns and hierarchies in data.

**Deployment** is the process of distribution and use of results obtained from data mining.

**Derived attributes** are constructed from one or more existing attributes in the same record.

**Gain tables** measure model effectiveness by showing the difference between results obtained by the model and results obtained without using the model.

**Lift charts** enable users to measure model effectiveness by showing the ratio between results obtained using the model and results obtained without using the model. The farther the lift line from the baseline, the more effective the model.

**Machine learning techniques** are used to enables a computer to learn a specific task, such as, decision making, estimation, classification, or prediction without manual programming.

**Model** is a set of representative rules, behaviors, or characteristics against which data are analyzed to find similarities. **Descriptive** models are used to analyze past events. **Predictive** models are used to discover what will happen in the future. With predictive models, data miners can explore alternative scenarios to determine which actions will produce the desired future outcome.

**Neural network** is a model for predicting or classifying cases using a complex mathematical scheme that simulates an abstract version of brain cells. A neural network is trained by presenting it with a large number of observed cases, one at a time, and allowing it to update itself repeatedly until it learns the task.

**Noise** is the difference between a model and its predictions. Sometimes data is referred to as noisy when it contains errors, such as many missing or incorrect values or when there are extraneous columns.

**Online analytical processing (OLAP)** enables users to analyze many layers of current and historical data. Though OLAP can tell you what is happening and what happened previously with your data, it can't tell you what will happen in the future.

**Pivot tables** are interactive tables that enable users to get different views of information by easily repositioning rows, columns, and layers of data.

**Predictive analytics** is a combination of advanced analytic techniques and decision optimization. Predictive analytics uses historical information to make predictions about future behavior, and then delivers recommended actions to the people and systems that can use them.

**Predictive modeling** is the process of creating models to predict future activity, behavior, or characteristics. For example, a predictive model may show which customers are most likely to churn in the future, based on the characteristics and actions of previous churners.

A request sent to a database for information based on specified characteristics or properties is called a **query**.

**Record** is a set of related data stored together. Also known as a row in spreadsheets or a case in statistics.

**Reporting** is the process of deploying or distributing the results of data analysis in a format that is comprehensible to the recipient.

**Return on investment (ROI)** is the value that is returned or obtained from various investments in technology, infrastructure, etc.

**Rule induction technique** is the process of automatically deriving decision-making rules for predicting or classifying future cases from example cases.

**Sequence detection** is the process of discovering the order of events in data. For example, use sequence detection to discover the order in which customers purchase certain products. Contrast with **association,** which reveals which products are purchased together.

**Text mining** is a process of analyzing textual information, such as, documents, e-mails, and call center transcripts to extract relevant concepts.

Data in a text format or other non-numerical format is called **unstructured.**

**Up-selling:** The practice of offering and selling more profitable products or services to existing customers than those they currently own or use.

**Web cubes:** An online, multi-layered display used to examine the relationships between symbolic data fields.

**Web mining:** The process of analyzing data from online activities including pay-per-click advertising and other marketing campaigns to discover relevant patterns and important behavioral insights.

**Training set:** In a dataset a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set. Usually a dataset is divided into a training set, a validation set (some people use 'test set' instead) in each iteration, or divided into a training set, a validation set and a test set in each iteration. The training set can be selected by applying a random filter to the data, e.g., select 20% of the points at random to generate the model and test against the remaining 80%.

**Overfitting:** Overfitting occurs when there are too many parameters or the wrong kind of parameters that has been overly trained to data but won't have good forecasting ability with new data i-e noise is being modeled instead of signal but one don't know it.

**Supervised and unsupervised learning:** Informally clustering is assignment of objects to classes on basis of observations about objects only. In unsupervised learning, classes are initially unknown and need to be "discovered" from the data. For example cluster analysis, class discovery, unsupervised pattern recognition etc. In supervised learning, classes are predefined and need a "definition" in terms of the data which is used for prediction. For example classification, discriminant analysis, class prediction, supervised pattern recognition etc.