OPTIMIZED FEATURE CONSTRUCTION METHODS FOR DATA SUMMARIZATION OF RELATIONAL DATA

FLORENCE SIA FUI SZE

THESIS SUBMITTED IN FULLFILMENT FOR THE DEGREE OF MASTER OF SCIENCE

SCHOOL OF ENGINEERING AND INFORMATION TECHNOLOGY
UNIVERSITI MALAYSIA SABAH
2014

ABSTRAK

KAEDAH PEMBINAAN ATRIBUT YANG DIOPTIMUMKAN UNTUK MERUMUSKAN DATA BERJADUAL

Terdapat banyak pendekatan telah dibangunkan untuk mendapat pengetahuan (iaitu maklumat yang berguna) daripada data yang disimpan di dalam pangkalan data berjadual. Penggabungan Dinamik Atribut Hubungan (Dynamic Aggregation of Relational Attributes - DARA) algoritma merupakan salah satu pendekatan diperkenalkan untuk merumuskan data yang disimpan di dalam jadual sasaran yang mempunyai hubungan satu-ke-banyak dengan data yang disimpan di dalam jadual bukan sasaran melalui proses transformasi daripada data hubungan perwakilan ke ruang vektor perwakilan dan proses pengelompokan digunakan untuk mengumpulkan data berdasarkan persamaan ciri-ciri yang terdapat di dalam data. Hasil rumusan data akan dijadikan sebagai input data kepada mana-mana algoritma pengelasan untuk melaksanakan tugas klasifikasi. Klasifikasi merupakan satu tugas yang biasanya dilakukan untuk memperoleh pola dalam data yang boleh digunakan untuk pengelasan data yang baru. Di dalam DARA, ketepatan pengelasan data yang diperolehi daripada tugas klasifikasi boleh terjejas disebabkan oleh ketepatan deskriptif rumusan data, DARA. Ketepatan deskriptif rumusan data DARA adalah sangat dipengaruhi oleh perwakilan rekod bukan sasaran dalam bentuk model ruang vektor. Pembinaan attribut telah menunjukkan kemampuan untuk memperkayakan perwakilan rekod bukan sasaran dan dengan itu, meningkatkan ketepatan deskriptif rumusan data. Tetapi kaedah pembinaan attribut yang digunakan di dalam DARA adalah tidak begitu berkesan kerana DARA tidak meneroka semua perwakilan rekod yang mungkin berpotensi tinggi untuk dihasilkan. Di dalam tesis ini, kaedah pembinaan attribut baru diperkenalkan dan persoalan sama ada ketepatan deskriptif rumusan data boleh mendapat manfaat daripada kaedah pembinaan attribut baru disiasat. Rangka kerja yang dicadangkan melibatkan pengunaan algoritma genetik serta beberapa jenis kaedah pemarkahan attribut untuk mengoptimasikan proses pembinaan attribut. Tesis ini juga membentangkan kajian berkaitan dengan kaedah untuk meningkatkan ketepatan deskriptif algoritma DARA melalui perumusan data secara gandaan. Keputusan empirik menunjukkan bahawa ketepatan pengelasan dapat ditingkatkan dan dengan itu, ketepatan deskriptif rumusan data boleh mendapat manfaat daripada kaedah yang dicadangkan. Kaedah tersebut menyediakan ruang carian yang lebih luas untuk mendapatkan cara perwakilan yang lebih relevan bagi mewakili rekod di dalam jadual bukan sasaran.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AV	Attribute-Value
CE	Total Cluster Entropy
CRISP-DM	Cross-Industry Standard Process for Data Mining
DARA	Dynamic Aggregation of Relational Attributes
DBI	Davies Bouldin Index
DM	Data Mining
FC	Feature Construction
FE	Feature Extraction
FK	Foreign Key
FLFCWOS	Fixed Length Feature Construction without Substitution
FLFCWS	Fixed Length Feature Construction with Substitution
FS	Feature Selection
GA	Genetic Algorithm
IG	Information Gain
ILP	Inductive Logic Programming
KDD	Knowledge Discovery in Database
k-NN	k-Nearest Neighbour
NB	Naïve Bayes
P_c	Probability of Crossover
PK	Primary Key
PKDD	Principles and Practice of Knowledge Discovery in Databases
P_m	Probability of Mutation
SQL	Structured Query Language
TF-IDF	Term Frequency-Inverse Document Frequency
VLFCWOS	Variable Length Feature Construction without Substitution
VLFCWS	Variable Length Feature Construction with Substitution
WEKA	Waikato Environment for Knowledge Analysis

LIST OF PUBLICATIONS

Sia, F., & Alfred, R. (2012). Evolutionary-based feature construction with substitution for data summarization using DARA. *Data Mining and Optimization (DMO), 2012 4th Conference*, 53-58.

Sia, F., Alfred, R., Leau, Y.B., & Tan, S.F. (2012). A Variable Length Feature Construction method for data summarization using DARA. *Computing and Convergence Technology (ICCCT), 2012 7th International Conference*, 881-887.

Sia, F., Alfred, R., Leau, Y.B., & Tan, S.F. (2013). A Random Length Feature Construction Method for Learning Relational Data using DARA. *International Journal of Information Processing and Management*, 4(3), 103-113.

Sia, F., Alfred, R., & Chin, K.O. (2012). Learning Relational Data Based on Multiple Instances of Summarized Data Using DARA. *Soft Computing Applications and Intelligent Systems. Second International Multi-Conference on Artificial Intelligence Technology, M-CAIT 2013*, 293-301.

CHAPTER 1

INTRODUCTION

1.1 Introduction

In recent years, there are large and complex data often generated and gathered in a relational database for real life application (Kavurucu *et al.*, 2011). The traditional data storage which is a single table in an attribute-value structure is impractical to store these data. A relational database has been widely used for real world data storage because it is capable to capture and describe complex data in a relational form (Kavurucu *et al.*, 2011; Xavier *et al.*, 2011). The increasing demand for tools to analyse data stored in the relational database in order to extract their useful information has drawn researchers' attention to the works in the field of Knowledge Discovery in Database (KDD), specifically in the data mining. KDD is often presented as a broad process of discovering knowledge in data where data mining is viewed as one of the phases. Data mining is a process of extracting patterns or models from large volume of data by using intelligent algorithms (Maimon and Rokach, 2010). Data mining task can be categorized into descriptive data mining which finds patterns that describe some relations that present in the data, whereas, predictive data mining finds patterns or models that map data to the target concepts of interest (Choudhary *et al.*, 2009).

Traditional data mining methods are only applied to extract patterns from a single data table, thus, many data mining approaches have been developed to extract patterns from the relational data in a relational database. In a relational database, data are stored in the multiple tables in which data in a table can be linked to data stored in other tables. The relationships that exist among tables need to be considered when extracting patterns. A summarization process can be performed on the relational data by integrating all associated records into a single table applicable

for extracting patterns using traditional data mining methods. But this way of summarization is not appropriate when an individual record stored in a table is related to many records stored in other tables (i.e. one-to-many relationship between tables) because the integration will produce a table that consists of multiple records that are associated to a single record in the original table, thereby causing loss of meaning and redundancies in the summarized data (Thakkar and Kosta, 2012). In order to address this issue, a data summarization algorithm called Dynamic Aggregation of Relational Attributes (DARA) has been developed to summarize data stored in the multiple tables that have one-to-many relationships (Alfred, 2007, 2009b).

DARA data summarization algorithm is one of the approaches introduced in order to extract patterns from data stored in relational databases. This algorithm transforms target records stored in the target table that has high number of one-to-many relationship with the non-target records stored in the non-target table into a vector space representation (Alfred, 2007, 2009b). In DARA algorithm, a target record stored in a target table is a unique labelled object and the non-target records stored in a non-target table are objects in which multiple objects from this non-target table can be linked to a single object stored in a target table (Alfred, 2010). The summarization process takes place by clustering the records in the vector space model into groups where multiple records in non-target table associated to a particular record in target table. The identification number of clusters will be used to form a new constructed feature that will be embedded into the target table and used to perform the most common predictive data mining task that is the classification task by using any traditional classification methods. The quality of extracted patterns or learned models is highly influenced by the representation of the input data (Piramuthu and Sikora, 2009; Shafti and Perez, 2009). In DARA, the predictive accuracy of the classification task is influenced by the performance of the data summarization process on the basis of the descriptive accuracy of the single instance of summarized data. The descriptive accuracy of the single instance of summarized data is highly dependent on the way of the non-target records are represented in the vector space model (Alfred, 2009c, 2010). A feature construction method can be ap-

plied to improve the descriptive accuracy of the single instance of summarized data by constructing a new set of features to represent the non-target records that contribute towards the representation of these records to describe the associated target records in the vector space representation for summarization process (Alfred, 2009c, 2010).

In this thesis, novel feature construction methods are proposed in order to improve the descriptive accuracy of the DARA algorithm. The aim of the research work in this thesis is to investigate whether or not the descriptive accuracy of the single instance of summarized data can benefit from the novel feature construction methods. The work carried out also involves the application of an evolutionary algorithm that is genetic algorithm (GA) on the novel feature construction methods in order to optimize the feature construction process and several feature scoring measures are evaluated in order to find the best relevant set of constructed features.

1.2 Problem Statement

DARA (Dynamic Aggregation of Relational Attributes) data summarization approach is an effective method to extract patterns from relational data by summarizing the data stored across multiple tables with one-to-many relationship in order to facilitate the classification task. In DARA data summarization process, it had found that the descriptive accuracy of the single instance of summarized data influences the predictive accuracy of the classification task. High descriptive accuracy of the single instance of summarized data is required in order to achieve high predictive accuracy of the classification task. The descriptive accuracy of the single instance of summarized data is highly dependent on the patterns generated to represent the records in non-target table that describe target records in the TF-IDF (*term frequency-inverse document frequency*) weighted frequency matrix (i.e. vector space representation) for summarization process. Therefore, more attention has been devoted to generate enriched patterns by enriching the representation of the records stored in the non-target table that associated to the records stored in the target table.

In DARA, feature construction has been applied to enrich the representation of records in the non-target table by constructing a new set of features to represent the records in the non-target table. The feature construction method is called "Fixed Length Feature Construction without Substitution" (FLFCWOS). It constructs a new set of features by randomly combining the attributes obtained from an original set of attributes given in the non-target table at fixed length without allowing an individual attribute to be combined more than once. Based on this method, given a set of attributes in the non-target table $\{F_1, F_2, F_3, F_4, F_5\}$, one could have a set of constructed features such as $\{F_1F_2, F_3F_4, F_5\}$ or $\{F_1F_3, F_2F_4, F_5\}$. However, this feature construction method does not explore all possible relevant constructed features subsets that may valuable to enrich the representation of records in non-target table.

1.3 Research Motivation

In DARA data summarization process, the application of feature construction to construct a new set of features from an original set of attributes given in the non-target table which will be used to generate patterns that represent the records in the non-target table has proved an improvement in the descriptive accuracy of the single instance of summarized data (Alfred, 2009c, 2010). This is the main motivation that leads to devoting our attentions on feature construction in relation to improve the descriptive accuracy of the single instance of summarized data. High predictive accuracy can be achieved by performing the classification task on the target table with a single instance of summarized data which has high descriptive accuracy.

In addition to that, it is known that many researches have focused on feature construction in order to achieve better data mining performance. This is due to its capability in augmenting the original features set of the raw data set and hence, it enhances the representation of data so as to highlight more informative patterns for data mining task. The highly informative patterns ease the mining process in which lower the computational cost and mining time. In fact, a good quality of data representation could result in high quality of discovered knowledge or patterns.

1.4 Novel Feature Construction Methods for Data Summarization

This research introduces novel feature construction methods to find potential relevant constructed features as many as possible to generate interesting patterns for summarizing records stored in the non-target table in order to improve the descriptive accuracy of the single instance of summarized data.

The novel feature construction methods are called the "Fixed Length Feature Construction with Substitution (FLFCWS)", "Variable Length Feature Construction without Substitution (VLFCWOS)", and "Variable Length Feature Construction with Substitution (VLFCWS)". FLFCWS constructs a new set of feature by combining the original attributes randomly at fixed length and it allows an individual attribute to be combined more than once. In VLFCWOS, a new set of features is constructed by combining the original attributes randomly at various lengths without allowing an individual attribute to be combined more than once. On the other hand, in VLFCWS, a new set of features is constructed by combining the original attributes randomly at various lengths but it allows an individual attribute to be combined more than once. Table 1.1 shows the possible set of features constructed based on the FLFCWS, VLFCWOS, and VLFCWS methods. Let $F = \{F_1, F_2, F_3, F_4, F_5, F_6\}$ represents a set of attributes given in a non-target table.

Table 1.1: Set of features constructed based on the novel feature construction methods

Feature Construction Method	Set of Constructed Features
FLFCWS	$F_1F_2F_3$, $F_2F_4F_3$, $F_4F_5F_6$
VLFCWOS	F_1F_2, $F_3F_4F_5$, F_6
VLFCWS	$F_1F_2F_3$, F_4F_3, $F_4F_5F_6$

The framework of these novel feature construction methods in summarizing relational data are described in greater detail in Chapter 3.

1.5 Research Objectives

The objectives of this research are to:

i. Propose a new framework for feature construction methods that construct a relevant set of features for summarizing data stored in the multiple tables with one-to-many relationship.

ii. Develop a genetic-based algorithm to optimize the process of constructing features in finding a relevant set of constructed features that best used to summarize relational data with one-to-many relationship.

iii. Determine the best framework for feature construction methods by evaluating the performance accuracy of the classification task.

1.6 Research Scopes

The scopes of this research are:

i. In designing the framework for the feature construction methods, three feature constructions will be developed to construct features for the Dynamic Aggregation of Relational Attributes (DARA) data summarization algorithm. These frameworks will accommodate algorithms called the Fixed Length Feature Construction with Substitution (FLFCWS), Variable Length Feature Construction without Substitution (VLFCWOS), and Variable Length Feature Construction with Substitution (VLFCWS).

ii. In optimizing the process of constructing features in finding a set of relevant constructed features that best used to summarize relational data with one-to-many relationship, a genetic algorithm will be designed to find the most relevant set of constructed features according to four feature scoring measures which including Total Cluster Entropy (CE), Information Gain (IG), Total Cluster Entropy coupled with Information Gain (CE-IG), and Davies-Bouldin Index (DBI).

iii. In evaluating the performance accuracy of the classification task based on the proposed framework for feature construction methods, Mutagenesis database

(Kristen and Wrobel, 2001) and Hepatitis database (PKDD 2005) will be used for data summarization process. The quality of the constructed features in summarizing relational data is evaluated based on the predictive accuracy of the classification task that is performed on the target table with a single instance of summarized data. The predictive accuracy is the percentage of instances in the test set that is correctly classified. This can be obtained by comparing the known class label of test set with the result of the classification task of the test set. The classification algorithms which include C4.5 decision tree, Naïve Bayes, and k-nearest neighbor implemented in WEKA are used to perform the classification task.

1.7 Research Contributions

This research contributes towards an area of Computer Science field that is Artificial Intelligent, specifically in data mining which is mainly on processing data that stored in the multiple tables with one-to-many relationship in order to improve the performance accuracy of extracting patterns from relational data. The contributions of this research work are presented as below:

Three novel feature construction methods, namely Fixed Length Feature Construction with Substitution (FLFCWS), Variable Length Feature Construction without Substitution (VLFCWOS), and Variable Length Feature Construction with Substitution (VLFCWS) have been developed for DARA data summarization algorithm to summarize data stored in the multiple tables with one-to-many relationship. The existing feature construction method applied in DARA data summarization algorithm constructs a set of features by combining original attributes at fixed length without allowing an individual attribute to be combined more than once causes many of the possible combinations of attributes that may valuable have not yet discovered. The novel feature construction methods proposed in this thesis implement different strategies in constructing a set of features which enables the combination of attributes at various lengths and allowing an individual attributes to be combined more than once and this will provide a wider combination of features space.

- A genetic algorithm is applied in each novel feature construction method where Information Gain (IG), Total Cluster Entropy (CE), Total Cluster Entropy coupled with Information Gain (CE-IG), and Davies-Bouldin Index (DBI) were used as the feature scoring measures to find the best relevant constructed set of features.

- The predictive accuracy of the classification task performed on the target table with a single instance of summarized data can be improved by constructing a new set of features using the novel feature construction methods to represent records in the non-target table for data summarization process. This has shown that the application of the novel feature construction methods in summarizing relational data with one-to-many relationship is capable to improve the descriptive accuracy of the single instance of summarized data. An analysis of the results obtained from the experiments showed that the Variable Length Feature Construction without Substitution (VLFCWOS) method is the best framework for feature construction method in DARA data summarization algorithm.

1.8 Thesis Organization

This thesis consists of seven chapters. The structure of each chapter is briefly described as below:

a. Chapter 1

   This chapter provides an overview of the research work in this thesis. It introduces briefly on the research background and problem addressed. It also presents the objectives, scopes, and contributions of the research. The organization of the chapters in this thesis is outlined at the end of this chapter.

b. Chapter 2

   This chapter presents the literature related to the research work of this thesis. It begins by presenting an overview of the Knowledge Discovery in Database (KDD). A well-known KDD methodology is presented to describe the KDD process. Then, it presents the most important phase in KDD process that is the data preparation phase. It also presents the core process of the KDD process that is

the data mining phase. This is followed by presenting the main concern of the work in this thesis that is the data mining in relational database. It describes the relational model of the relational database and reviews the existing data mining approaches to extract patterns from data stored in the multiple tables in a relational database. This chapter is concluded by underlying Dynamic Aggregation of Relational Attributes (DARA) data summarization algorithm as an effective data mining approach when dealing with data stored in the multiple tables that have one-to-many relationship between tables.

c. Chapter 3

This chapter presents the methodology and framework used in the research work. It begins by presenting a framework for Fixed Length Feature Construction with Substitution (FLFCWS) method and a brief description on the process of summarizing relational data based on the features constructed using FLFCWS method. This is followed by presenting a framework for Variable Length Feature Construction without Substitution (VLFCWOS) method and a brief description on data summarization process based on the features constructed using VLFCWOS method. Lastly, a framework for Variable Length Feature Construction with Substitution without Substitution (VLFCWOS) method and a brief description on the data summarization using VLFCWOS method are presented. This chapter is concluded by summarizing this chapter.

d. Chapter 4

This chapter presents the feature construction in summarizing relational data using DARA. It begins by describing the DARA data summarization process in which a relational data representation is transformed into a vector of patterns representation in order to construct a TF-IDF (*term frequency-inverse document frequency*) weighted frequency matrix. This data transformation is required in order to summarize or cluster data stored in the relational databases. It also illustrates the application of feature construction to construct a set of features to enrich the representation of the patterns used in constructing a TF-IDF weighted frequency matrix. Then, novel feature construction methods are presented for DARA data summarization process in order to construct a set of features that

can further enrich the data representation. It outlines the novel feature construction methods involved in summarizing relational data. Next, the development of genetic algorithms for constructing features based on the novel feature construction methods is presented. This followed by presenting the feature scoring measures used to find the best relevant constructed set of features. This chapter is concluded by presenting the subsequent step need to be taken to complete the DARA data summarization process that involves the novel feature construction methods.

e. Chapter 5

This chapter presents the experimental setup and design. This chapter begins by outlining the purposes of conducting experiment. Next, the datasets used for data summarization are described. Then, the settings of the DARA algorithm for summarizing data and the parameter settings of genetic algorithms for constructing features based on the novel feature construction methods are outlined. There are three classifiers of WEKA are used to perform the classification task on the target table with a single instance of summarized data. A brief description of each of these classifiers is presented. Lastly, the experimental procedures which involve the process from summarizing the datasets by applying the genetic-based novel feature construction methods until the target table with a single instance of summarized data is fed to the classifiers so as to perform the classification task are demonstrated. This chapter is concluded by summarizing this chapter.

f. Chapter 6

This chapter presents the experimental results and the analysis of the results. The predictive accuracy results collected from the classifiers are tabled. The predictive accuracies of the classification tasks on the target table with a single instance of summarized data based on the genetic-based novel feature construction methods are compared with the genetic-based existing feature construction method for every classifier. It also presents the comparison of the predictive accuracies of the classification task with respect to the feature scoring measure used. This is followed by presenting the discussion based on the

comparative study results. Next, the comparison of the predictive accuracies of the classification task among of the genetic-based novel feature construction is presented. The discussion based on the evaluation of the results obtained from this comparison is also presented. Lastly, this chapter is concluded by answering the question addressed in this thesis and underlying the best framework for feature construction process in summarizing data that are stored in the multiple tables that have one-to-many relationship.

g. Chapter 7

This chapter presents the study of a technique to learn relational data based on multiple instances of summarized data and its implication to the predictive accuracy of the classification task. This chapter begins by presenting the proposed method to generate multiple instances of DARA summarized data using the selected single instance of summarized data which is obtained using the best set of features constructed based on the feature construction methods. Next, the experimental setups and results are presented. The predictive accuracy of the classification task using the proposed method and the tradition techniques is compared. Last section concludes this chapter by presenting the findings of this study.

h. Chapter 8

This chapter concludes the research work of this thesis. The summary of the research processes, findings, limitations, and future directions are outlined.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter reviews on literatures related to the domain of interest of this research. The first section presents an overview of Knowledge Discovery in Database (KDD) field and a well-known CRISP-DM model to describe the KDD process. The second section presents the importance of data preparation in discovering knowledge and tasks involved in preparing data for knowledge discovery process. This followed by a description of the common approaches and a brief review of the corresponding existing methods used to perform the data preparation task. The third section describes the core process of KDD, called data mining which including the two broad groups of tasks, namely descriptive data mining task and predictive data mining task. The third section describes the relational model of the relational database and provides a review on the existing approaches used to perform the data mining task on the data stored in multiple tables in a relational database. Last section concludes this chapter by underlying the main focus of this research.

2.2 Knowledge Discovery in Database

A tremendous amount of data are being generated every day from many fields such as medical and health care industry, science and engineering, education, finance, banking, business, and almost every other aspect of daily life. The advances in computer and database technology have enabled this large amount of data to be collected and stored in the database easily. The data in the database often contain knowledge which can be very beneficial in facilitating important decision making. In order to extract the knowledge, an analysis process is required to be performed on the data. For example, in banking field, the consumer historical data stored in the database of loan can be analysed to identify the characteristic of good and bad loan

applicant that valuable for better loan application approval. In the past few years, the data analysis relies heavily on human to analyse data manually. However, it is expensive, time consuming, and impractical when the amount of data has far exceeded the human analytic ability. The data accessibility provided by the database system, such as automated match and retrieval, are also not adequate for analysis process.

Therefore, Knowledge Discovery in Database (KDD) has emerged to automate the works of analysing and extracting knowledge implicitly present in the large volume of data. KDD is an interdisciplinary field that involved numerous fields such as machine learning, pattern recognition, databases, statistics, artificial intelligence, knowledge acquisition, data visualization, and high performance computing. KDD is widely defined as a "*nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*" (Fayyad, 1996; Fayyad *et al.*, 1996).

In other words, KDD is a process of extracting patterns which are valid on new data with certainty, previously unknown, beneficial in achieving the knowledge discovery goal, and comprehensible to the user. The term pattern (i.e. knowledge) is an expression represented in some languages that describes a subset of data or a model that applicable to the data subset (Fayyad, 1996).

KDD process comprises numerous steps, each of which involve many decisions to be made by user and application of techniques from any of the abovementioned fields to accomplish a particular discovery task. The KDD process is interactive in which any changes made at each step will affect the performance of task in the subsequent step and also iterative, meaning that moving backward and forward between several steps is allowed to make any changes for better performance of task at later step. As the KDD research field continues evolved, there are several different KDD process models are currently available as the

guidelines to execute knowledge discovery task systematically. The models are differing from one another in their number of step and scope of task at specific step but they shared some common tasks which start with understanding the application domain, preparing data set, extracting knowledge, evaluating the extracted knowledge, and putting the discovered knowledge into use (Maimon and Rokach, 2010).

One of the well-known KDD process models is CRISP-DM (Cross-Industry Standard Process for Data Mining) model (Wirth, 2000; Azevedo, 2008; Maimon and Rokach, 2010). CRISP-DM model was first developed in the late 1990s by a large consortium of European Companies. It has been broadly applied in industrial real knowledge discovery projects. This model consists of six steps which are business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Chapman *et al.*, 2000).

The procedures in each of the six steps within CRISP-DM model are briefly described as below:

1. **Business Understanding**

   This step initiates the knowledge discovery process by understanding the objectives and requirements of the project from a business perspective. Based on these, the knowledge discovery problem is determined and a preliminary project plan for achieving the objectives is prepared.

2. **Data Understanding**

   This step starts with collecting an initial data and proceeds with performing tasks that enable to become familiar with the data. These followed by getting initial insights into the data, identifying the quality problem of data, and detecting the interesting data subsets.