

INTELLIGENT MINING MULTI DIMENSIONAL ASSOCIATION RULES FROM LARGE INCONSISTENT DATABASES

Sarjon Defit¹, Mohd Noor Md Sap²

¹Faculty of Computer Science Universiti Putra Indonesia (UPI) "YPTK" Padang,
West Sumatera, Indonesia.

Tel.: 62-751-776666 Fax: 62-751-71913 Email: sarjon_d@hotmail.com

²Faculty of Computer Science and Information System Universiti Teknologi Malaysia
KB. 791 80990, Johor Bahru, Malaysia.

Tel.: 60-7-5532419 Fax: 60-7-5566155 Email: mohdnoor@fksm.utm.my

Abstract:

The widespread use of computer applications, database technologies and data collection techniques have resulted in the accumulation of large amounts of data in databases. This has generated an urgent need for new techniques that can intelligently and automatically transform the processed data into useful information and knowledge. In this paper, we propose an intelligent method for mining multi dimensional association rules from large inconsistent databases. It is called Intelligent Mining Association Rules (IMAR). The proposed IMAR was experimented and studied using three domain data sets. It includes Australian Credit Card (ACC), Jakarta Stock Exchange (JSX), and Cleveland Heart Diseases (CLEV) data sets. The results of this study show that IMAR is a promising method for mining multi dimensional association rules from large inconsistent databases intelligently and accurately, and IMAR is a promising method for solving complex data mining problems.

1. Introduction to Data Mining

The widespread use of computer applications, database technologies and data collection techniques have resulted in the accumulation of large amounts of data in databases [1, 2, 3, 4, 5, 27]. For instance, thousands of databases have been used in business management, government administration, banks, stock markets and super markets. It creates demands for analyzing these data and extracting them into useful knowledge [3, 4, 5, 6, 27]. Obviously, using traditional methods of data analysis is far beyond human capabilities to analyze large amounts of data [3, 4, 5, 7, 8, 27]. This has generated an urgent need for new techniques that can intelligently and automatically transform the processed data into useful information and knowledge [1, 2, 3, 4, 5, 7, 8, 27].

Knowledge Discovery in Databases (KDD) and data mining are two techniques for discovering rules from large amounts of data in databases [9, 10, 11, 27]. Knowledge Discovery in Databases (KDD) is defined as the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [7,9, 27]. In Zaiane's opinion [7, 27], the Knowledge Discovery in Databases (KDD) process comprises several steps including (i) data cleaning, (ii) data integration, (iii) data selection, (iv) data transformation, (v) data mining, (vi) pattern evaluation, and (vii) knowledge representation. Although data mining is only one step of the Knowledge Discovery in Databases (KDD) process, it is the central part of KDD process.

Data mining is described as an effort to understand, analyze and eventually make use of the large amounts of data available [3, 4, 5, 9, 12, 27]. It is defined as the automatic extraction of patterns, rules, previously unknown and potentially useful from large amounts of data in databases and using it to make crucial business decisions [12, 27].

Generally, data mining is divided into descriptive and predictive data mining [7, 10, 12, 13, 27]. The former is conducted to describe general properties of existing data and to create meaningful subgroups such as demographic cluster while the latter is used for forecasting explicit values based on available data [7,10, 12, 13, 14, 27]. In a view that differs from Zaiane's, Sang, J.L. and Keng, S. [2, 27] classified data

mining into two main classifications based on (i) databases, i.e. relational databases, object oriented databases, web databases, multimedia databases, spatial databases and time series databases, and (ii) knowledge to be discovered, i.e. association rule, classification, characterization, cluster, and prediction.

Based on these classifications, association rule is a data mining method that is classified as knowledge to be discovered. It is a data mining method used to discover interesting rules or relationships among attributes in databases. It has attracted great attention in research communities in recent years since it has wide application in solving business problems and it is easier to understand even among non data mining experts.

2. Background of the Problem

This section provides the background of the problems. It includes association rules, data cleaning and data transformation as given in the following sub sections.

2.1 Association Rules

Association rule is a data mining method that is used for discovering knowledge from large amounts of data in databases. It is a rule in the form of [3, 4, 5, 6, 7, 15, 16, 17, 18, 27].

$$X_1, X_2, \dots, X_m \rightarrow Y_1, Y_2, \dots, Y_n \quad [S, C] \quad (1.1)$$

where X_i and Y_i are items. The S and C are support and confidence of rules respectively. It has become more and more popular since its introduction in 1993. Today, it is still one of the most popular methods in data mining [3, 4, 5, 6, 15, 16, 17, 18, 27].

A number of promising association rule methods have been studied and developed. These include (i) a close method [19, 27], (ii) a closet method [20, 27], (iii) online analytical mining association rule [6, 27], (iv) mining association rules with multiple Minimum Item Support [15, 27], and (v) discovery of knowledge at multiple level concepts [9, 27]. These methods have given great advantages to their users when generating rules from large amounts of data. However, association rule methods still present several weaknesses which need further improvement. The first weakness is related to the accuracy of the association rule method [6, 9, 15, 19, 20, 27]. The association rule method should portray the contents of the database accurately. The noise and uncertainty should be reduced elegantly. It is one of the important tasks in data preparation in order to identify which data in databases are inaccurate or missing values. The second drawback has to do with the usefulness of the association rule method [6, 15, 19, 20, 27]. The association rule method should be useful for certain applications and generate association rule from data which are represented at higher levels concepts. The raw data should be transformed from raw data into higher levels concepts. It allows users to generate association rules deeply and view database contents at different abstraction levels. The third weakness is in the identification of interesting rules [9, 15, 19, 20, 27]. The association rule method should generate more interesting rules accurately. The generated rule should be identified in order to reduce the number of interesting rules without loss of information. Fourth, there is also some weakness in the use of prior and domain knowledge [6, 9, 15, 19, 20, 27]. The association rule method should generate interesting rules intelligently. It should us to generate association rules from large amounts of data in databases intelligently and automatically. Lastly, intelligent hybrid association rule method [6, 9, 15, 19, 20, 27]. The association rule method should generate knowledge from various domains and solve complex data mining problems. It should be able to generate more interesting rules and reduce the number of rules without loss of information.

**BORANG PESANAN
(Order Form)**

Penyelenggara Penerbitan,
Fakulti Sains Komputer dan Sistem Maklumat
Universiti Teknologi Malaysia,
81310 UTM Skudai,
Johor.

(u.p.: Cik Najmah bt. Hj. Shamsuddin)

Sila kirimkan pesanan jurnal seperti yang dicatatkan di bawah kepada saya seperti alamat yang diberi. Bersama-sama ini disertakan wang kiriman pos/cek/draf bank no.: atas nama Bendahari, Universiti Teknologi Malaysia bernilai RM (Ringgit Malaysia :)

Nama/Name:

Organisasi/
Organization

Alamat/ :
Address

Bandar/ City :

No.Telefon atau Fax:
(Telephone or Fax Number)

1. Sila hantarkan kepada saya **..... salinan Jurnal Teknologi Maklumat, Jilid 14 Bil.1 (Jun 2002) (RM 15.00 senaskah)
2. Sila hantarkan kepada saya **..... salinan Jurnal Teknologi Maklumat, Jilid 14 Bil.2 (Dis 2002) (RM 15.00 senaskah)
3. Sila hantarkan kepada saya **..... salinan Jurnal Teknologi Maklumat, Jilid 15 Bil. 1 (Jun 2003) (RM 15.00 senaskah)

Perhatian :

Sila tandakan pada kotak pesanan yang berkaitan.

** Sila nyatakan jumlah salinan.

Tambahkan belanja pos sebanyak RM2/= setiap naskah kepada jumlah harga kiriman pos, cek atau draf bank.

Mustahak : Sila sertakan borang ini apabila membuat tempahan

2.2 Data Cleaning

Data cleaning is a method employed for handling uncertain data that is contained in large amounts of data in databases. It has become a crucial step in Knowledge Discovery in Databases (KDD) process and data mining [21, 22, 23, 24, 27].

A number of promising data cleaning methods have been studied and developed. For instance, listwise deletion, pairwise deletion, mean imputation, maximum likelihood [21, 27], and data cleaning based on rough set [22, 27]. These data cleaning methods have given great advantages in order to create a complete and consistent data. However, these data cleaning methods need the following further improvements. First, it requires the use of prior and domain knowledge [21, 22, 27]. The data cleaning method should identify and fill the missing values intelligently. It should allow us to identify and fill the incomplete data in databases intelligently. Second, an intelligent hybrid data cleaning method could be incorporated to improve the method [21, 22, 27]. The data cleaning method should be able to handle the missing values more efficiently. It should be able to create a more powerful complete data as a source for the next process, i.e. data transformation and rule generation.

2.3 Data Transformation

Data transformation is a method employed to transform raw data in large amounts of data in databases into higher levels concepts. It has become an important step in Knowledge Discovery in Databases (KDD) process and data mining. A number of data transformation methods have been studied and developed [9, 24, 25, 26, 27]. Some examples are (i) the automatic generation of conceptual hierarchy for numerical attributes [9, 27], (ii) the chimerge [25, 27], (iii) the Chi² [25, 27], and (iv) the DISC method [26, 27]. These data transformation methods have given great advantages in transforming raw data into higher levels concepts. However, these methods need the following further improvements. First, the accuracy of the data transformation method can be improved [9, 25, 26, 27]. The data transformation should be able to transform raw data into higher levels concepts accurately. The missing values should be filled elegantly in creating a transformed data more accurately. Second, use of prior domain knowledge is required [9, 25, 26, 27]. The data transformation method should be able to transform raw data into higher levels' concepts intelligently. It should enable us to transform raw data into transformed data intelligently and accurately. Lastly, the use of intelligent hybrid data transformation may enable the transformation of raw data in large databases into higher levels concepts to be generated more efficiently [9, 25, 26, 27].

In order to cater these problems, an intelligent method is proposed for mining multi dimensional association rules from large inconsistent databases.

3. IMAR General Architecture

In this section, the Intelligent Mining Association Rules (IMAR) general architecture is illustrated in figure 3.1. Figure 3.1 shows the general architecture of the Intelligent Mining Association Rules (IMAR). Generally, Intelligent Mining Association Rules (IMAR) consists of three main phases including preprocessing, processing and post processing. The first phase has two processes including data cleaning and data transformation while the main process of processing is rule generation. The first two phases, i.e. preprocessing and processing, comprise two main steps including training and running steps. The training step is conducted for generating neural network knowledge based of data cleaning, data transformation and association rules. The generation of complete data, transformed data and interesting association rules are conducted in the running step. Next, the generated interesting rules are applied in real world problems in creating crucial business decisions. The brief explanations of Intelligent Mining association rules (IMAR) are given in the following sections.

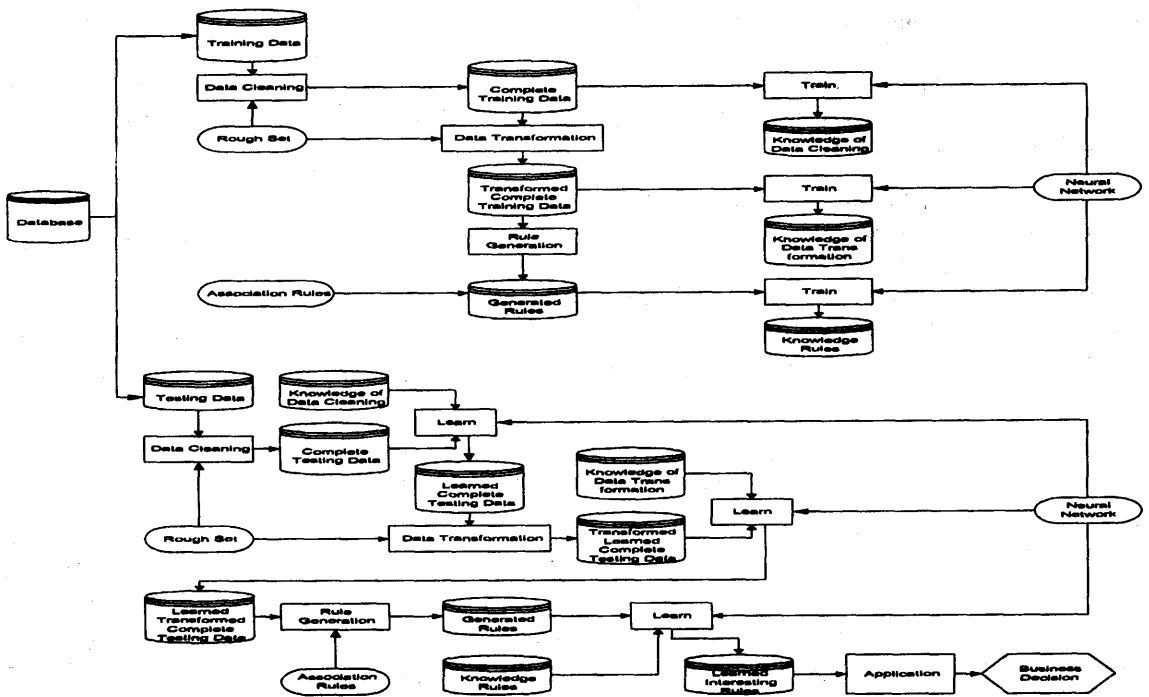


Figure 3.1: The Intelligent Mining Association Rules (IMAR) General Architecture

3.1 IMAR Phases

In the following sub sections, the IMAR phases is described. It includes preprocessing and processing phases.

3.1.1 Preprocessing

Preprocessing is the first phase of Intelligent Mining Association Rules (IMAR). It consists of two processes including data cleaning and data transformation.

3.1.1.1 Data Cleaning

In this section, the general architecture of data cleaning is illustrated in figure 3.2.

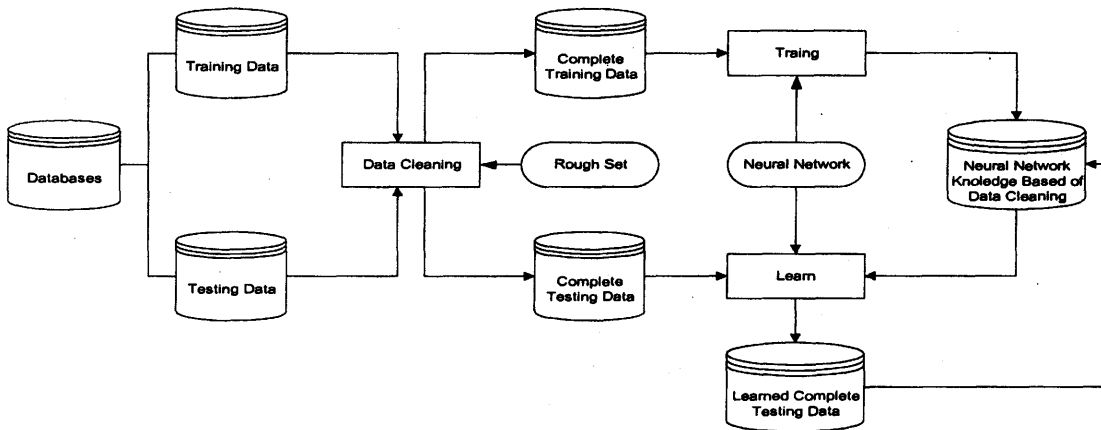


Figure 3.2: The Data Cleaning General Architecture

Figure 3.2 shows the general architecture of data cleaning. Generally, it consists of two steps including training and running steps. The training step is conducted for creating the neural network knowledge based of data cleaning while the running step for generating learned complete data. First, the data stored in databases may be incomplete. It is split into training and testing data. The incomplete training data are filled using basic data cleaning algorithm as described in [27]. It is used for creating complete training data. The incomplete and complete training are then merged for creating the basic structure of neural network knowledge based of data cleaning. It consists of two parts including the condition and action parts. The condition part is replaced with incomplete training data while the complete training data is located at the action part. Next, the merged data is then split into the training and testing data sets. These data are then trained using neural network described in [27] for creating trained data. The trained data consists of two parts including target output and trained output. The target output is replaced with the complete training data obtained using basic data cleaning algorithm while the trained output is replaced with the complete training data obtained using the neural network. Next, the condition part of testing data set is merged with the trained output part of trained testing data set in creating the neural network knowledge based of data cleaning. The merged data is saved as neural network knowledge based of data cleaning. Second, the incomplete testing data are filled using basic data cleaning algorithm described in [27] for creating complete testing data. Next, merge the incomplete and complete testing data in creating the basic structure of the new neural network knowledge based of data cleaning. The condition part is replaced with the incomplete testing data while the complete testing data is located at the action part. The merged data and neural network knowledge based of data cleaning are then assumed as the testing and training data sets respectively. These data sets are then learned using the neural network described in [27] in order to create the learned testing data set. It consists of two parts including the target and learned output. The target output is replaced with the complete testing data obtained using the basic data cleaning algorithm while the learned output is replaced with the complete testing data is obtained using the neural network. Next, merge the condition part of testing data set with the learned output part of the learned testing data set for creating a new neural network knowledge based of data cleaning. The testing data set and new neural network knowledge based of data cleaning are then compared. When the condition part of testing data set and the new neural network knowledge based of data cleaning are matched, the action part of testing data set is replaced with the action part of the new neural network knowledge based of data cleaning. The results of this process generate the learned complete testing data. Lastly, add new neural network knowledge based of data cleaning into the previous neural network knowledge based of data cleaning. For more details, a description of the data cleaning method is given in [27].

3.1.1.2 Data Transformation

In this section, the general architecture of data transformation is illustrated in figure 3.3.

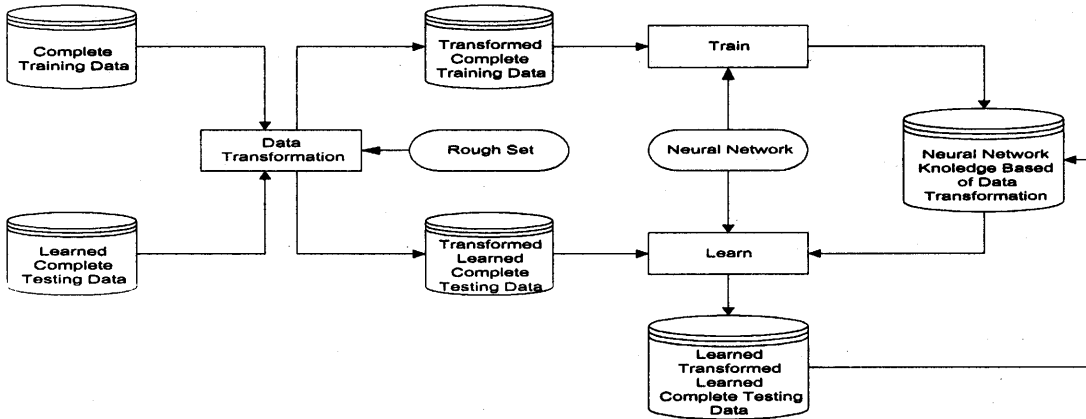


Figure 3.3: The General Architecture of Data Transformation

Figure 3.3 shows the general architecture of data transformation. Generally, it consists of two steps including training and running steps. The training step is conducted for creating neural network knowledge based of data transformation while the running step for generating learned transformed data. First, the complete training data are transformed using the basic data transformation algorithm as described in [27] in creating transformed complete training data. The complete and transformed training data are then merged for creating a basic the structure of neural network knowledge based of data transformation. It consists of two parts including condition and action parts. The condition part is replaced with the complete training data while the transformed complete training data is located at the action part. Next, the merged data is then split into training and testing data sets. These data are then trained using neural network described in [27] for creating trained data. The trained data consists of two parts including the target and trained output. The target output is replaced with the transformed complete training data obtained using the basic data cleaning algorithm while the trained output is replaced with the transformed complete training data obtained using the neural network. Next, the condition part of testing data set is merged with the trained output part of trained testing data set in creating the neural network knowledge based of data transformation. The merged data is saved as neural network knowledge based of data transformation. Second, the learned complete testing data are transformed using the basic data transformation algorithm described in [27] for creating transformed learned complete testing data. Next, merge the learned complete and transformed learned complete testing data in creating the basic structure of the new neural network knowledge based of data transformation. The condition part is replaced with the learned complete testing data while the transformed learned complete testing data is located at the action part. The merged data and neural network knowledge based of data transformation are then assumed as testing and training data sets respectively. These data sets are then learned using neural network described in [27] in creating the learned testing data set. It consists of two parts including target and learned output. The target output is replaced with the transformed learned complete testing data obtained using the basic data cleaning algorithm while the learned output is replaced with the transformed learned complete testing data is obtained using the neural network. Next, merge the condition part of testing data set with the learned output part of the learned testing data set for creating the new neural network knowledge based of data transformation. The testing data set and new neural network knowledge based of data transformation are then compared. When the condition part of testing data set and the new neural network knowledge based of data transformation are matched, the action part of testing data set is replaced with the action part of the new neural network knowledge based of data transformation. The results of this process produce the learned transformed learned complete testing data. Finally, add new neural network knowledge based of data transformation into the previous neural network knowledge based of data transformation. For more details, a description of the data transformation method is given in [27].

3.1.2 Processing

The second phase of Intelligent Mining Association Rules (IMAR) is processing. The main process of processing phase is rule generation. In this section, the general architecture of rule generation is represented in figure 3.4.

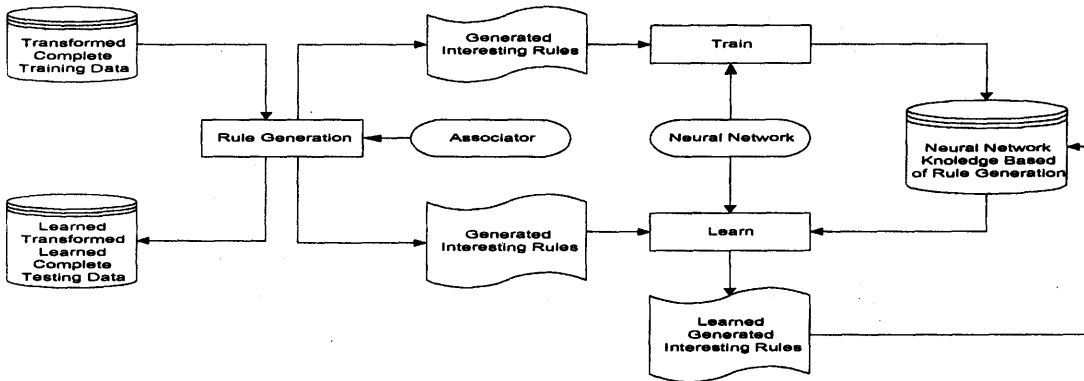


Figure 3.4: The General Architecture of Processing Phase

Figure 3.4 shows the general architecture of processing phase. Generally, it consists of two steps including training and running steps. The training step is conducted for generating the neural network knowledge based of association rules while the running step for generating learned interesting rules. First, the transformed complete training data are generated using the basic rule generation algorithm as described in [27] in generating interesting rules. The generated interesting rules consist of three parts including condition, action and performance of rules, i.e. support, confidence and interesting. The generated interesting rules are then coded into numerical values. Next the coded interesting rules are then split into two parts called the training and testing data sets. These data sets are then trained using neural network described in [27] for creating trained data. It consists of two parts including the target and trained output. The target output is replaced with the action part of testing data set obtained using the basic rule generation algorithm while the trained output is replaced with the action part of testing data set is obtained using the neural network. Next, the condition part of testing data set are then merged with the trained output part of testing data set in order to create neural network knowledge based of rule generation. Second, the learned transformed learned complete testing data are generated using the basic rule generation algorithm as described in [27] in generating interesting rules. The generated interesting rules are then coded into numerical values. Next, the coded interesting rules and neural network knowledge based of rule generation are taken as the testing and training data set. These data sets are then learned using neural network described in [27] in creating learned testing data set. It consists of two parts including the target and learned output. The target output is replaced with the action part of testing data set obtained using the basic rule generation algorithm while learned output is replaced with the action part of testing data set is obtained using the neural network. Next, merge the condition part of the testing data set with the learned output part of learned testing data set for creating new neural network knowledge based of association rules. The merged data and new neural network knowledge based of rule generation are then compared. When the condition part of testing data set and new neural network knowledge base of rule generation are matched, the action part of testing data set is replaced with the action part of the new neural network knowledge based of association rules. Learned interesting rules are derived from the results of this process. Finally, add new neural network knowledge based of rule generation into the previous neural network knowledge based of rule generation. For more details, a description of processing phase is described in [27].

3.2 IMAR Modules

Intelligent Mining Association Rules (IMAR) is designed using a combination of several intelligent techniques, i.e. rough set, association rules, neural networks and knowledge based. The purpose

of combining these intelligent techniques is to create a method for solving the data mining complex problems, i.e. data cleaning, data transformation and association rules. A brief explanation of IMAR modules is given in sections 3.2.1, 3.2.2 and 3.2.3.

3.2.1 Rough Set

Rough set is the first module of Intelligent Mining Association Rules (IMAR). It is used for supporting the preprocessing phases in both data cleaning and data transformation as presented in figure 3.5.

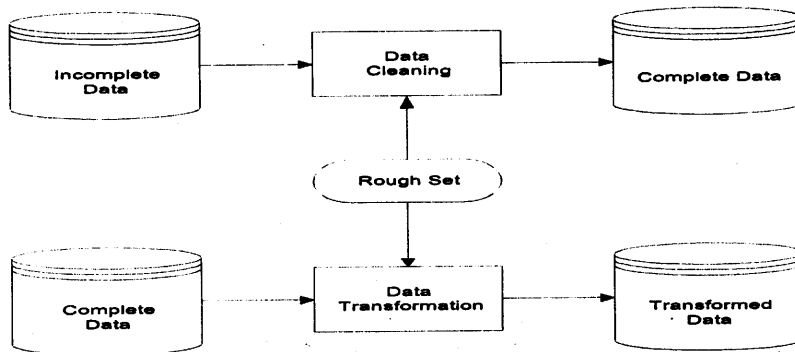


Figure 3.5: The Function of Rough In IMAR Data Cleaning and Data Transformation

Figure 3.4 shows the rough set which functions in supporting the IMAR data cleaning and data transformation process. It is conducted in both training and running steps. In the training step, the rough set is used for supporting the creation of neural network of data cleaning and data transformation, i.e. the creation of action part of neural networks knowledge based. In the running step, it is used for creating the target output of the complete data and transformed data.

In order to support the data cleaning and data transformation processes, the discern object concepts in both discernibility and indiscernibility matrix is described in [27]. First, in the data cleaning process, the discernibility concept is used for identifying which objects in the decision systems are discern. In other words, it is used for identifying the objects in the decision system which have the same decision values in order to cluster the same objects into one cluster. Then, the indiscernibility matrix is used for identifying which objects in one cluster are indiscern or have different conditional values. It is conducted for measuring the distance of objects in one cluster in order to find the most similar objects. Next, the incomplete data in decision system are replaced with the most similar objects. Second, in data transformation, the rough set is used for identifying which objects of transformed data are consistent or inconsistent using the discernibility matrix concept. An object in one cluster is considered as an inconsistent object when the object has the same conditional values when the decision is different. The decision values of inconsistent objects in one cluster are replaced with the frequent decision values in one cluster. For more details, the IMAR rough set, i.e., discern object, is described in [27].

3.2.2 Association Rule

The second module of IMAR is the association rule. It is used in the processing phase in order to generate interesting rules from databases as presented in figure 3.6.

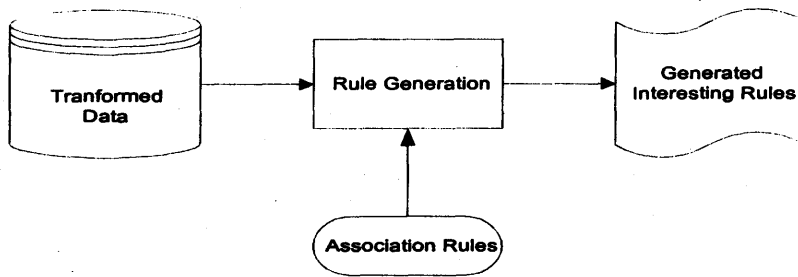


Figure 3.6: The Function of Association Rules in IMAR Processing Phase

Figure 3.5 shows the function of association rules techniques in order to generate interesting rules from databases. It is conducted in both training and running steps. In the training step, it is used for supporting the creation of the neural network knowledge based of association rules, i.e. the basic structure of neural networks knowledge based. In the running step, it is used for creating the target output, i.e. the action part, of interesting rules. In order to generate interesting rules, the basic concept of association rules is described in [27]. The generation of association rules in IMAR follows these steps: (i) generate clusters of item sets in databases. The item sets are clustered into one cluster when the support and interestingness of pair item sets are greater than or equal to Minimum support and minimum interestingness threshold, (ii) generate association rules from each cluster, and (iii) evaluate the generated rules in order to identify whether the generated rules are interesting or not. For more details, the IMAR association rules concept is described in [27].

3.2.3 Neural Networks Knowledge Based

In this module, two intelligent techniques, i.e. neural network and knowledge based are combined. It is used in the preprocessing and processing phases in both training and running steps. In the training step, it is used for creating the neural network knowledge based including the neural network of data cleaning, data transformation and association rules. In the running step, it is used for supporting IMAR in order to generate learned complete data, transformed data and interesting rules. The function of neural network knowledge based in training and running of IMAR is presented in figure 3.7a and 3.7b.

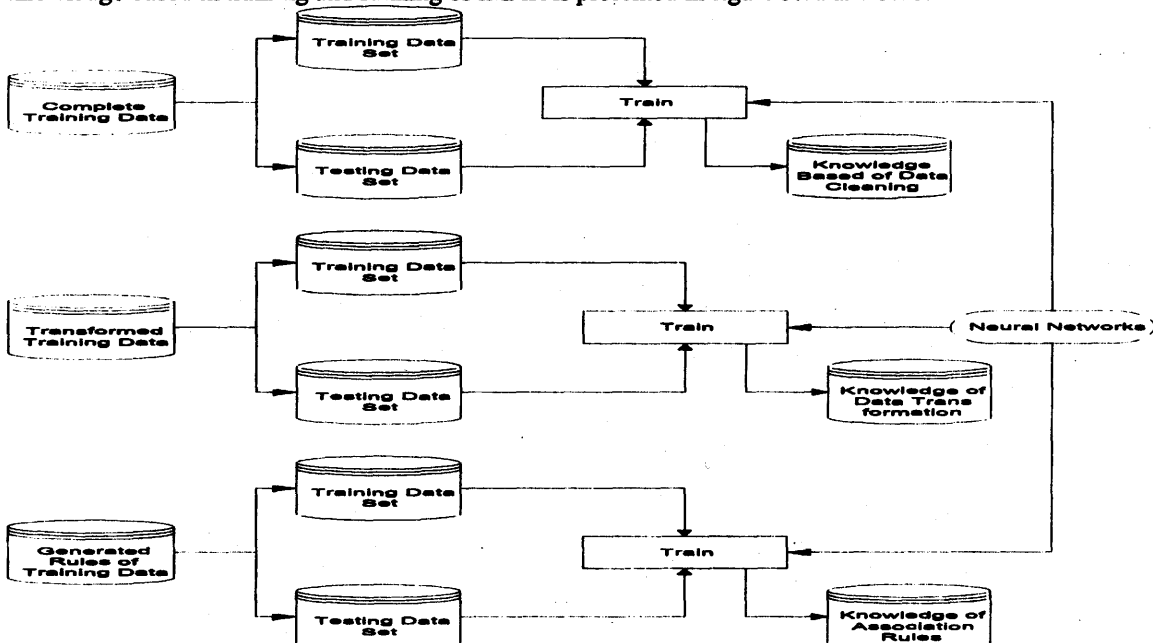


Figure 3.7a: The Function of Neural Network and Knowledge Based in

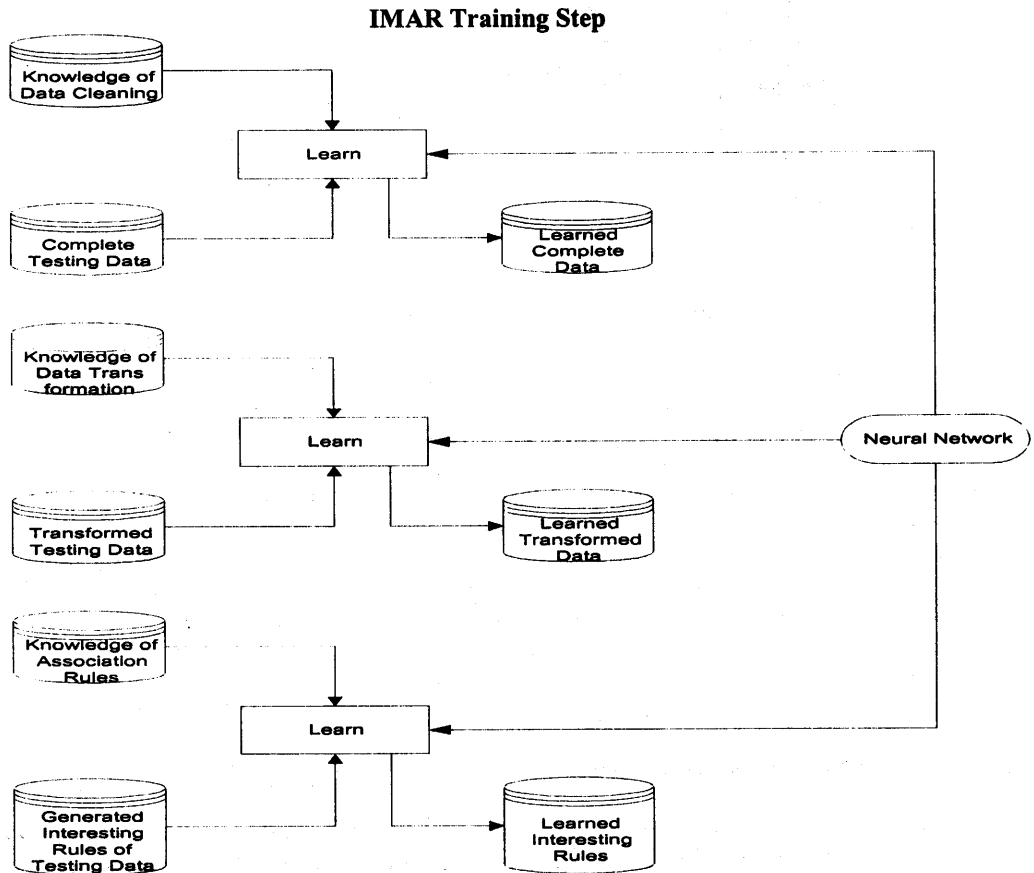


Figure 3.7b: The Function of Neural Networks and Knowledge Based in IMAR Running Step

Figure 3.7a shows the function of neural network in order to create the neural network knowledge based of data cleaning, data transformation and association rules. In order to create the neural network knowledge based, the training data is split into training and testing data sets. These data are then trained using neural networks in order to generate trained data. The trained data is saved as the neural network knowledge based. The neural network knowledge based consists of two parts namely the condition and action parts. The condition part is replaced with the condition part of testing data sets while the action part with trained data.

Figure 3.7b shows the function of neural network knowledge based in order to generate learned complete data, transformed data and interesting rules. In order to generate learned data, the testing data set and neural network knowledge based are learned using neural network. Then, the learned data is compared with the testing data set using the matching process algorithm described in [?] in order to generate learned data output.

4. DISCUSSION AND COMPARISONS

In this chapter, the Intelligent Mining Association Rules (IMAR) method and performances are discussed. In section 4.1, the Intelligent Mining Association Rule (IMAR) method is presented. The Intelligent Mining Association Rules (IMAR) performance including preprocessing and processing phases is given in section 4.2.

4.1 IMAR Method

The association rule is a method for discovering association rules from large amounts of data in databases. It has become a popular data mining method for discovering knowledge and has been applied successfully in a wide range of business problems. A number of promising association rules methods have been studied and developed, i.e. Closet, Close, MSAppriori, online analytical of association rules and meta rule guided mining association rules methods. These association rules methods have given great advantages for discovering knowledge from large databases. However, these association rules methods still have some limitations and need further improvements as described in section 2.1. In order to cater to these association rule limitations, in this research, we have proposed an intelligent method for mining multi dimensional association rules from large inconsistent databases. It is called Intelligent Mining Association Rules (IMAR). A comparison of IMAR with previously proposed association rules methods is given in table 4.1.

Table 4.1: A Comparison of IMAR With Previously Proposed Association Rules Methods

Researchers	Method	Model	Techniques
Pasquier, N. and Bastide Y. et.al.	Close		Association Rules
Jian, P. and Jiawei, H.	Close		Association Rules
Bing, L. and Wynne, H. et.al.	MSAppriori		Association Rules
Hua, Z.	Online Analytical of Association Rules		Association Rules
Yongjian, F.	Progressively Deepening and Meta Rule Guided Mining		Association Rules, Statistic
Sarjon, D	IMAR		Rough Set, Neural Networks, Knowledge Based, Association Rules, Statistic

Table 4.1 shows a comparison of IMAR with previously proposed association rule methods in both models and techniques used. This study shows that most previous association rule methods, i.e. Closet, Close, MSAppriori, online analytical of association rules, generate association rules directly from raw data in databases and use a single technique, i.e. statistics. On the other hand, the association rule

method proposed by Yongjian Fu generate association rules from transformed data and this method is developed using statistical and association rule techniques. The statistical technique is used for transforming raw data into transformed data while the association rules is used for generating association rules from transformed data contained in databases. Compared with previously proposed association rules methods, the IMAR method is proposed for creating a method which is able to solve complex data mining problems, i.e. data cleaning, data transformation, and association rules generation. In order to achieve the purposes, IMAR has been designed through three main phases, i.e. preprocessing, processing and post processing. The processing phase is conducted for (i) creating a complete databases, and (ii) transforming raw data into transformed data. The generation of interesting rules is done in the processing phase while the generated interesting rules are applied into real world problems for making a crucial business decision in the post processing phase. In order to achieve these purposes, the IMAR has been developed using a combination of several techniques, i.e. rough set, neural network, knowledge based, association rules and statistics. The rough set technique is used for supporting the preprocessing phase in both data cleaning and data transformation while the association rules is used for supporting the processing phase in order to generate interesting rules. The neural network knowledge based, i.e. a combination of neural network and knowledge based, is used in both preprocessing and processing phases for (i) creating knowledge based, represented as the neural network knowledge based, which consists of knowledge based of data cleaning, data transformation and association rules, and (ii) supporting the IMAR process in order to generate learned complete data, transformed data and interesting rules.

In the preprocessing phase, the researcher has proposed two intelligent methods, i.e. intelligent data cleaning and data transformation methods. First, data cleaning is a method for handling the missing and uncertain data in databases. It is an important step in the Knowledge Discovery in Databases (KDD) process and data mining. A number of data cleaning methods have been studied and developed, i.e. mean imputation, maximum likelihood and data cleaning based on rough set methods. These data cleaning methods have handled the missing and uncertain data contained in databases elegantly. However, these data cleaning methods still have some weaknesses and need further improvements as described in section 2.2. In order to improve the ability of data cleaning methods, in this research, the researcher has proposed an intelligent data cleaning method for handling the missing and uncertain data in databases intelligently and elegantly. A comparison of IMAR data cleaning with previously proposed data cleaning methods is given in table 4.2.

Table 4.2: A Comparison of IMAR Data Cleaning With Previously Proposed Data Cleaning Methods

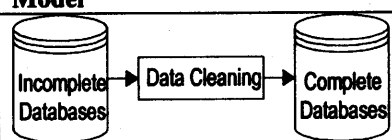
Researchers	Method	Model	Techniques
Little And Shanker	Listwise Deletion, Pairwise deletion, mean imputation, maximum likelihood		Statistic

Table 4.2: A Comparison of IMAR Data Cleaning With Previously Proposed Data Cleaning Methods (Cont)

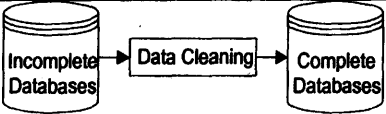
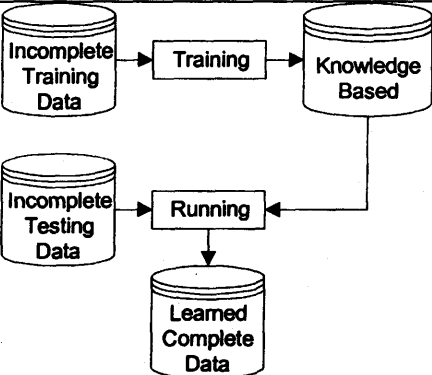
Felix, R	Rule Induction from Inconsistent and Incomplete Data Using Rough Set	 <pre> graph LR A[Incomplete Databases] --> B[Data Cleaning] B --> C[Complete Databases] </pre>	Rough Set
Sarjon, D.	IMAR Data Cleaning	 <pre> graph TD A[Incomplete Training Data] --> B[Training] B --> C[Knowledge Based] D[Incomplete Testing Data] --> E[Running] C --> E E --> F[Learned Complete Data] </pre>	Rough Set, Neural Network, Knowledge Based, and Statistic

Table 4.2 shows a comparison of IMAR data cleaning with previously proposed data cleaning methods. This study concludes that, most of previous data cleaning methods directly handle the incomplete data from databases and use a single technique which is either the statistic or rough set. Compared with previously proposed data cleaning methods, IMAR data cleaning was developed using a combination of several intelligent techniques, i.e. rough set, neural network, knowledge based, and statistic. In order to generate more accurate complete data, IMAR was designed through two main steps, i.e. training and running steps. The first step is conducted for creating neural network knowledge based of data cleaning and then the created neural network knowledge based of data cleaning is used for supporting the running step in order to generate learned complete data intelligently.

Second, data transformation is a method for transforming raw data into transformed data or discretized data. It is a crucial step in Knowledge Discovery in Databases (KDD) process and data mining. It allows us to extract knowledge from different abstraction levels and generate simpler and association rules which are easier to understand even for non data mining experts. A number of promising data transformation methods have been studied and developed, i.e. automatic generation of conceptual hierarchy, ChiMerge, Chi², DISC algorithm, and equal binning frequency. These data transformation methods could transform raw data into transformed data efficiently and effectively. However these data transformation methods still have some weaknesses and need further improvements as described in section 2.3. In order to reduce these weaknesses in data transformation, in this research, an intelligent data transformation method is proposed for transforming raw data into transformed data intelligently and accurately. A comparison of IMAR data transformation with previously proposed data transformation methods is given in table 4.3.

Table 4.3: A Comparison of IMAR Data Transformation With Previously Proposed Data Transformation Methods

Researchers	Methods	Model	Techniques
Yongjian, F.	Automatic Generation of Conceptual Hierarchy		Statistic
Kerber	ChiMerge		Statistic
Hua, L. and Rudy, S	Chi ²		Statistic
Wesley, W.C. and Kuorong, C.	DISC Algorithm		Statistic
Nguyen, H.S.	Equal Binning Frequency		Statistic
Sarjon, D.	IMAR Data Transformation		Rough Set, Neural Network, Knowledge Based, Statistic

Table 4.3 shows a comparison of IMAR data transformation with previously proposed data transformation methods. This study shows that most of previously proposed methods use a single technique, i.e. statistic, for transforming raw data into transformed data. These previously proposed data transformation methods also directly transform raw data into transformed data without handling the missing and uncertain data contained in databases in order to generate transformed data accurately. Compared with previously proposed data transformation methods, IMAR data transformation method was developed using a combination of several intelligent techniques, i.e. rough set, neural networks, knowledge based, and statistic. In order to generate more accurate and consistent transformed data, IMAR data transformation was designed through two steps, i.e. training and running steps. The first step is conducted for creating neural network knowledge based of data transformation and then the created neural network knowledge based of data transformation is used for supporting the running step in order to generate learned transformed data.

4.2 IMAR Performance

The proposed Intelligent Mining Association Rules (IMAR) was experimented and studied using three domain data sets. It includes Australian Credit Card (ACC), Jakarta Stock Exchange (JSX), and Cleveland Heart Diseases (CLEV) data sets. In the following section, the IMAR performances in both preprocessing and processing phases are discussed.

4.2.1 Preprocessing Phase

In this sub section, the IMAR performances in the preprocessing phase in both IMAR data cleaning and data transformation performances are described.

4.2.1.1 IMAR Data Cleaning Performance

In this research, the IMAR data cleaning performance is measured based on three differences error measurements, called Mean Difference Error (MDE), Mean Relative Error (MRE) and MSE (Mean Square Error) which are defined as follows

$$MDE = \frac{1}{n} \sum_{i=1}^n (va_i - vp_i) \quad (4.1)$$

$$MRE = \frac{1}{n} \sum_{i=1}^n \frac{(va_i - vp_i)}{vp_i} \quad (4.2)$$

$$MSE = \left[\frac{1}{n} \sum_{i=1}^n (va_i - vp_i)^2 \right]^{1/2} \quad (4.3)$$

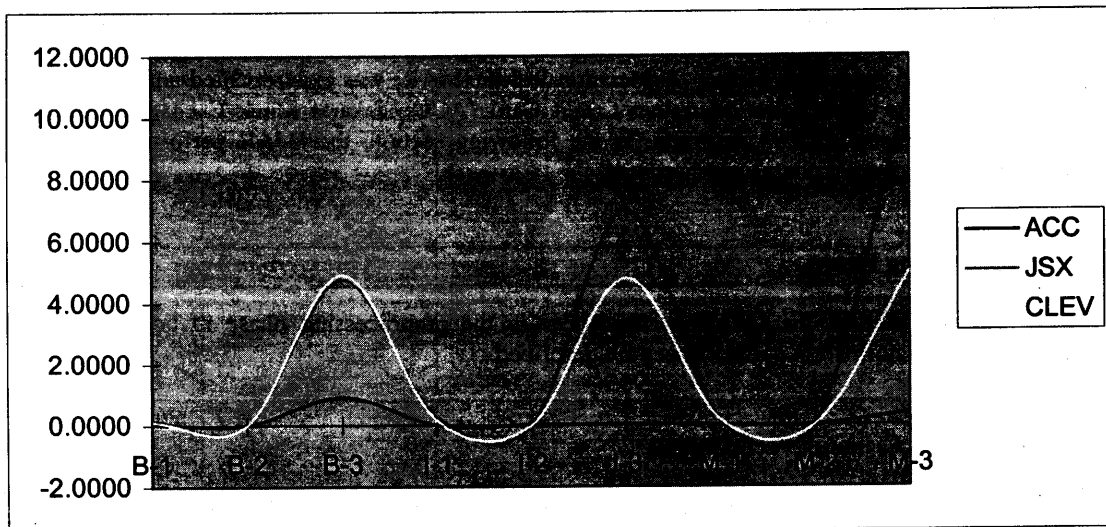
where

va_i and vp_i = the observed and predicted values of output respectively
 n = the number of observations

The performance of IMAR data cleaning compared with mean imputation data cleaning method proposed by Little and Shanker as described in section 2.2 is given in the following table 4.4 and figure 4.1.

Table 4.4: The Performance of IMAR Data Cleaning Compared with Mean Imputation Data Cleaning Method

Data Set	Basic Data Cleaning			IMAR Data Cleaning			Mean Imputation		
	MDE	MRE	MSE	MDE	MRE	MSE	MDE	MRE	MSE
ACC	0.037	0.0026	0.921	0.004	0.0013	0.0160	0.0127	0.0013	0.4033
JSX	0.107	0.0013	4.764	0.260	0.0012	7.000	0.108	0.0032	10.987
CLEV	0.156	0.0001	4.911	0.138	0.0011	4.795	0.1436	0.0011	4.9932



where

- B-1 : The Mean Difference Error of Basic Data Cleaning
- B-2 : The Mean Relative Error of Basic Data Cleaning
- B-3 : The Mean Square Error of Basic Data Cleaning
- I-1 : The Mean Difference Error of IMAR Data Cleaning
- I-2 : The Mean Relative Error of IMAR Data Cleaning
- I-3 : The Mean Square Error of IMAR Data Cleaning
- M-1 : The Mean Difference Error of Mean Imputation
- M-2 : The Mean Relative Error of Mean Imputation
- M-3 : The Mean Square Error of Mean Imputation

Figure 4.1: The Performance of IMAR Data Cleaning Compared with Mean Imputation Data Cleaning Method (Cont)

Table 4.4 and figure 4.1 show the performance of IMAR data cleaning compared with mean imputation data cleaning method. For example, in the JSX data set, the Mean Difference error (MDE) of IMAR data cleaning and mean imputation are equal to 0.2600 and 0.1080 respectively, while the Mean Relative Error (MRE) are equal to 0.0012 and 0.0032 respectively. The Mean Square Error (MSE) of the IMAR data cleaning method is equal to 7.000 while the MSE for the mean imputation method is 10.987. Based on these experimental results, we conclude that these data cleaning methods have offered a better results for handling uncertain and missing data. However, IMAR data cleaning offers several advantages as given in the following table 4.5.

Table 4.5: The Advantages and Disadvantages of IMAR Data Cleaning

Advantages	Disadvantages
i) It can handle the missing and uncertain data elegantly and intelligently ii) It gives more accurate values iii) It decreases the error including MDE, MRE and MSE	i) It is limited to the use on numeric data only

Table 4.5 shows the advantages and disadvantages of the IMAR data cleaning method. This study shows that the IMAR data cleaning can (i) handle the missing and uncertain data in databases intelligently and elegantly, (ii) handle the missing values accurately, and (iii) decrease the error measurements including mean differences error (MDE), Mean Relative Error (MRE) and Mean Square Error (MSE). Although the IMAR data cleaning method has great advantages, this method is still limited to the use of handling the

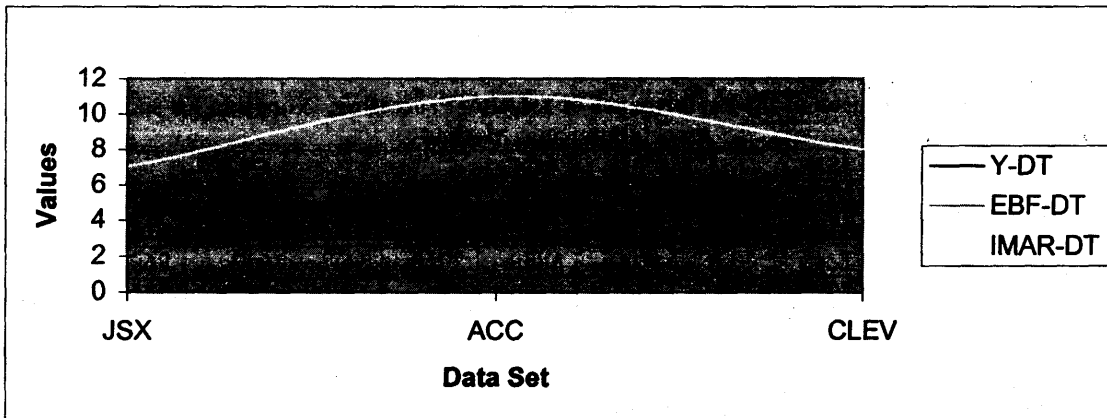
missing values from numerical data. It still needs further improvement for handling uncertain and missing data from other data types, i.e. categorical and text data.

4.2.1.2 IMAR Data Transformation Performance

In this research, the performance of IMAR data transformation is evaluated based on the number of generated classes of transformed data. A comparison of the performance three data transformation methods, i.e. (i) IMAR, (ii) Yongjian Fu's and (iii) the equal binning frequency data proposed by Nguyen, HS is given in table 4.6 and figure 4.2.

Table 4.6: A Comparison of the Performance (i) IMAR Data Transformation, Yongjian Fu's Data Transformation and Basic Data Transformation Method

Data Sets	The Generated Number of Classes (Mean)		
	Yongjian Fu's Data Transformation	Basic Data Transformation	IMAR Data Transformation
JSX	3	14	7
ACC	3	14	11
CLEV	3	14	8



where:

Y-DT = Yongjian Fu's Data Transformation
 EBF-DT = Basic Data Transformation
 IMAR-DT = IMAR Data Transformation

Figure 4.2: A Comparison of the Performance (i) IMAR Data Transformation, Yongjian Fu's Data Transformation and Basic Data Transformation Method

Table 4.6 and figure 4.2 show a comparison of IMAR data transformation, Yongjian Fu's data transformation and basic data transformation method. It is show here that, different methods generate different number of classes. For instance, Yongjian Fu's data transformation method generates the number of classes depending on the threshold level of the number of classes, i.e. 2, 3, ..., n. On the other hand, the basic data transformation generates each attribute in databases into a static number of classes. In other words, it is possible for all attributes in databases to have the same number of classes. Compared with Yongjian Fu's and basic data transformation, the IMAR data transformation could transform raw data in databases into different number of classes.

Based on these experimental results, it is concluded that the IMAR data transformation method offers several advantages as given in table 4.7.

Table 4.7: The Advantages and Disadvantages of the IMAR Data Transformation

Advantages	Disadvantages
i) It can transform raw data into transformed data accurately and intelligently. ii) It can transform raw data into various number of classes of transformed data	i) It limits itself to transforming raw data from numerical data. ii) it cannot transform raw data into higher levels.

Table 4.7 shows the advantages and disadvantages of IMAR data transformation. This method can (i) transform raw data into transformed data accurately and intelligently, and (ii) transform raw data into transformed data with different number of classes. Although IMAR data transformation could transform raw data into transformed data more accurately, this method still has the following limitations (i) it needs further improvement for transforming raw data into higher levels, and (ii) it is limited to transforming raw numerical data and needs further study for transforming raw data from other data types.

4.2.2 Processing Phase

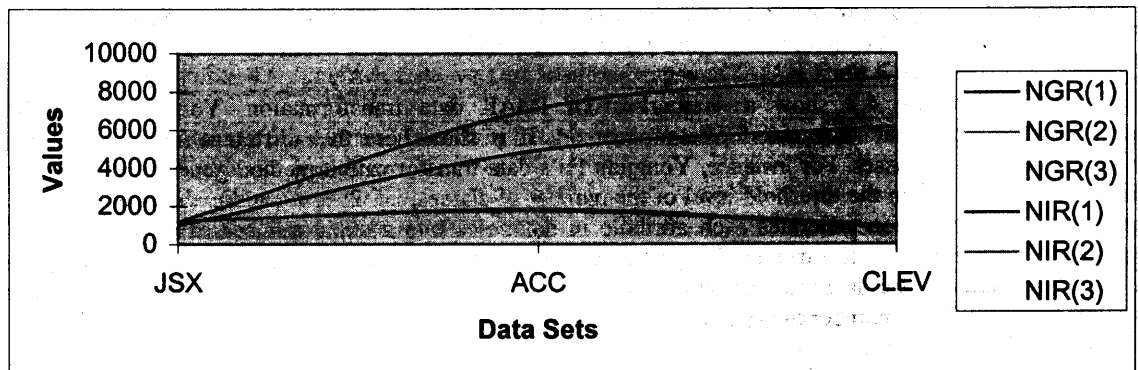
In this research, the processing performance of the Intelligent Mining Association Rules (IMAR) is measured using the number and accuracy of generated interesting rules. A comparison of processing performance of IMAR, basic association rule and Yongjian Fu's association rule methods is given in table 4.8 and figure 4.3.

Table 4.8: A Comparison of the Performance of IMAR's Processing, Yongjian Fu's and Basic Association Rules Method

Methods	JSX			ACC			CLEV		
	NGR	NIR	Acc	NGR	NIR	Acc	NGR	NIR	Acc
BAR	1181	1004	90.13	7120	4941	78.71	8781	6161	69.95
Yongjian Fu's	1240	1236	99.5	1879	1758	96.7	953	860	90.4
IMAR-AR	202	189	91.4	232	231	99.6	272	271	99.8

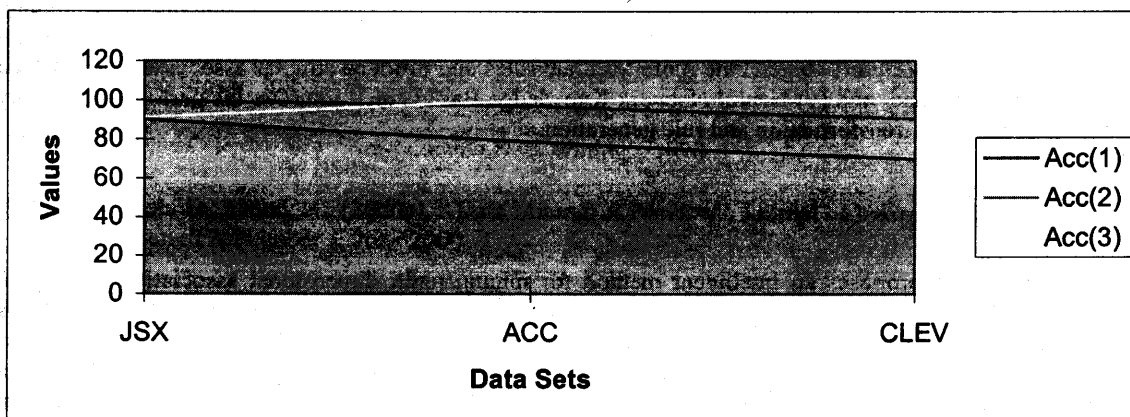
where:

BAR	= Basic Association Rules
Yongjian Fu's	= Association Rules Method Proposed by Yongjian Fu
IMAR-AR	= Intelligent Association Rules
NGR	= Number of Generated Rules
NIR	= Number of Interesting Rules
Acc	= Accuracy

**Figure 4.3a: A Comparison of The Performance of IMAR's Processing, Basic Association Rules and Yongjian Fu's Association Rules Methods**

where:

- NGR(1) = Number of Generated Rules Using Basic Association Rules
- NGR(2) = Number of Generated Rules Using Association Rules Method Proposed by Yongjian Fu
- NGR(3) = Number of Generated Rules Using IMAR-Association Rules
- NIR(1) = Number of Generated Interesting Rules Using Basic Association Rules
- NIR(2) = Number of Generated Interesting Rules Using Association Rules proposed By Yongjia
- NIR(3) = Number of Generated Interesting Rules Using IMAR-Association Rules



where:

- Acc(1) = Accuracy of Generated Interesting Rules Using Basic Association Rules
- Acc(2) = Accuracy of Generated Interesting Rules Using Association Rules Proposed by Yongjian Fu
- Acc(3) = Accuracy of Generated Rules Using IMAR-Association Rules

Figure 4.3b: A Comparison of The Performance of IMAR’s Processing, Basic Association Rules and Yongjian Fu’s Association Rules Methods

Table 4.8, figure 4.3a and 4.3b show a comparison of the performance of IMAR association rule with Yongjian Fu’s association rule methods. For example, using basic association rule, Yongjian Fu and IMAR association rule, the generated rules which form the JSX data set are equal to 1181, 1240 and 202 while the number of interesting rules are equal to 1004, 1236, and 189 rules respectively. The accuracy of generated interesting rules for the basic association rule and Yongjian Fu and IMAR association rule are equal to 90.3, 99.5 and 91.4 % respectively. To illustrate another example, the ACC data set shows that, the number of generated rules using basic association rules and Yongjian Fu and IMAR association rules are equal to 7120, 1879 and 232 respectively while the generated interesting rules are equal to 4941, 1758 and 231 respectively. The accuracy of generated interesting rules are equal to 78.71, 96.7 and 99.6 respectively. These results here show that IMAR’s processing offers several advantages as given in table 4.9.

Table 4.9: The Advantages and Disadvantages of IMAR’s Processing

Advantages	Disadvantages
i) it can mine multi dimensional association rules from large databases accurately and intelligently	i) it cannot mine more complex association rules
ii) It can reduce the number of generated interesting rules without loss of information	

Table 4.9 shows the advantages and disadvantages of IMAR's processing. This study shows that IMAR's processing could (i) mine multi dimensional association rules from large inconsistent data intelligently and accurately, (ii) reduce the number of generated interesting rules without loss of information, and (iii) generate simpler association rules which are easier to understand. Although IMAR's processing could reduce the number of generated interesting rules with higher accuracy, it should be extended for generating association rules from other types of rules since it is , at this moment, limited to generating rules from first predicates calculus form.

The IMAR performances including IMAR's preprocessing and processing have been studied and analyzed. This study shows that (i) IMAR could generate interesting rules from large inconsistent databases intelligently and accurately, (ii) IMAR could reduce the number of generated interesting rules without loss of information and improve the performance accuracy, (iii) IMAR could solve complex data mining problems, i.e. data cleaning, data transformation and association rules, (iv) the preprocessing phase in IMAR offers the opportunity to generate more accurate rules and produce simpler association rules which are easier to understand, and (v) Neural Network Knowledge Based could improve the performance ability of data cleaning, data transformation and rule generation.

5. Conclusion

This study proposes an intelligent method for mining multi dimensional association rules from large inconsistent databases. This method is called Intelligent Mining Association Rules (IMAR). IMAR is developed using a combination of intelligent techniques, i.e. rough set, neural networks, knowledge based and association rules, and statistic. A combination of these techniques is proposed for creating a data mining method that is able to handle complex data mining problems, i.e. data cleaning, data transformation and association rules. In order to achieve these objectives, the IMAR was designed through three main phases (i) preprocessing, i.e. data cleaning and data transformation, (ii) processing, i.e., rule generation, and (iii) post processing. The first phase is conducted for (i) handling missing and uncertain data contained in databases, and (ii) transforming raw data into transformed data or discretized data. The generation of interesting association rules is conducted in the second phase, i.e. rule generation. Next, the generated interesting association rules are applied into real world problems for creating crucial business decisions.

The results of this study show that (i) Intelligent Mining Association Rules (IMAR) is a promising method for mining multi dimensional association rules from large inconsistent databases intelligently and accurately, (ii) IMAR is a promising method for solving complex data mining problems, (iii) neural network knowledge based can improve the performance accuracy in order to support IMAR specifically during preprocessing and processing phases, and (iv) the processing phase in IMAR, i.e. data cleaning and data transformation provided the opportunity to generate more accurate association rules and to produce simpler association rules which are easier to understand.

BIBLIOGRAPHY

- [1] Ming-Syan, C., Jiawei, H., et.al., (1996). "Data Mining : An Overview From a Database Perspective" Proceeding of IEEE Transaction on Knowledge and Data Engineering, Vol 8, No 6, December 1996
- [2] Sang, J.L., and Keng, S., (1998). "Data Mining – A Review", 1998
- [3] Sarjon, D., Mohd, N., (2002a). "Mining Association Rules Using Rough Set and Association Rules Methods", Proceeding of International Conference on Artificial Intelligence in Engineering and Technology 2002 (ICAIET'02), Kota Kinabalu, Sabah, Malaysia, June 17-18, 2002
- [4] Sarjon, D., Mohd, N., (2002b). "Mining Multiple Level Association Rules Using Rough Set and Association Rule Methods". Proceeding of International Conference on Artificial Intelligence and Soft Computing 2002 (ASC'02), Banff, Canada, July 17-19, 2002

- [5] Sarjon, D., Mohd, N., (2002c). "Association Rules Using Rough Set and Association Rules Methods", Proceeding of Pacific Rim International Conference on Artificial Intelligence 2002 (PRICAI'02), Tokyo, Japan, August 18-20, 2002.
- [6] Hua, Z., (1998). "On-Line Analytical Mining of Association Rules", MSc Thesis Simon Fraser, December 1998.
- [7] Goebel, M., and Gruenwald, L., (1999). "A Survey of Data Mining and Knowledge Discovery Software Tools", Proceeding of ACM SIGKDD, Volume 1, Issue 1, Page 20, June 1999
- [8] Wei, W., (1999). "Predictive Modeling Based on Classification and Pattern
- [9] Yongjian, F., (1996). "Discovery of Multiple Level Rules From Large Databases", PhD Thesis Simon Fraser, July 1996
- [10] Zaiane, O.R., (1998). "Introduction Data Mining", 1998
- [11] Sarjon, D., Mohd, N., (2000a). "Data Mining: A Preview", Journal of Information Technology, Volume 12, Number 1, June 2000
- [12] Sarjon, D., Mohd, N., (2000b). "Predictive Data Mining Based on Similarity and Clustering Method", Journal of Information technology, Volume 12, Number 2, December 2002.
- [13] Olaru, C., and Wehenkel, L., (1999). "Data Mining", 1999
- [14] Jiawei, H., (1999). "Data Mining", in J. Urban and P. Dasgupta (Eds) Encyclopedia of Distributed Computing, Kluwer Academic Publisher, 1999
- [15] Bing, L., Wynne, H., et.al., (1999). "Mining Association Rules With Multiple Minimum Support", Proceeding of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-99), August 15-18, 1999
- [16] David, W.C., Vincent, T. Ng., et.al., (1996). "Efficient Mining of Association Rules in Distributed Databases", Proceeding of IEEE Transactions on Knowledge and Data Engineering, Vol 8, No 6, December 1996
- [17] Megiddo, N., and Srikant, R., (1998). "Discovery Predictive Association Rules", 1998
- [18] Hipp, J., Guntzer, U., et.al., (2000). "Algorithms For Association Rule Mining – A General Survey and Comparison", Proceeding of ACM SIGKDD, Vol 2, Issue 1, Page 58, July 2000
- [19] Pasquier, N., Bastide, Y., et.al., (1997). "Discovering Frequent Closet Itemsets For Association Rules", 1997.
- [20] Jian, P., Jiawei, H., et.al (1997). "Closet: An efficient Algorithm For Mining Frequent Closet Itemsets", 1997.
- [22] Felix, R., and Ushio, T., (1999a). "Rule Induction From Inconsistent and Incomplete Data Using Rough Sets", IEEE 1999
- [23] Agarwal, S., Keller, A.M., et.al., (1995). Flexible relation: An Approach For Integrating Data From Multiple, Possibly Inconsistent Data", IEEE 1995
- [24] Dung, P.M., (1996). "Integrating Data From Possibly Inconsistent databases", 1996

- [25] Huan, L., and Rudi, S., (1998). " Chi2: Feature Selection and Discretization of Numeric Attributes", 1998.
- [26] Wesley, W.C., and Kuorong, C., (1997). " Abstraction of High Level Concepts From Numerical Values in Databases", 1997.
- [27] Sarjon, D., (2002). " Intelligent Mining Multi Dimensional Association Rules From Large Inconsistent Databases", PhD Thesis, Universiti Teknologi Malaysia.