THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION

INSTITUTE OF SCIENCE AND TECHNOLOGY

MINING COMPLEX ASSOCIATION RULES BASED ON INTER-CLUSTER COMMUNICATION

MASTER'S THESIS

Iman ALİ F. HASSE

THE DEPARTMENT OF INFORMATION TECHNOLOGY

THE PROGRAM OF INFORMATION TECHNOLOGY

FEBRUARY 2017

THE UNIVERSITY OF TURKISH AERONAUTICAL ASSOCIATION

INSTITUTE OF SCIENCE AND TECHNOLOGY

MINING COMPLEX ASSOCIATION RULES BASED ON INTER-CLUSTER COMMUNICATION

MASTER'S THESIS

Iman A. F. Hasse

1406050054

THE DEPARTMENT OF INFORMATION TECHNOLOGY

THE PROGRAM OF INFORMATION TECHNOLOGY

FEBRUARY 2017

Türk Hava Kurumu Üniversitesi Fen Bilimleri Enstitüsü'nün 1406050054 numaralı Yüksek Lisans öğrencisi, Iman A. F. Hasse belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı Mining Complex Association Rules Based on Inter-Cluster Communication: Bilgi Erişim Teknikleri ile Kan Şekeri Seviyesinin Tahmini başlıklı tezini, aşağıda imzaları olan jüri önünde başarı ilesunmuştur.

Supervisor: Assist. Prof. Dr. Shadi AL SHEHABI .....

Türk Hava Kurumu Üniversitesi

Jury Members: Assist. Prof. Dr. Tansel DÖKEROĞLU .....

Türk Hava Kurumu Üniversitesi

Prof. Dr. Ahmet COŞAR .....

Orta Doğu Teknik Üniversitesi

Assist. Prof. Dr. Shadi AL SHEHABI .....

Türk Hava Kurumu Üniversitesi

Thesis Defense Date: 01.02.2017

THE UNIVERSITY OF TURK HAVA KURUMU

THE DEPARTMENT OF INFORMATION TECHNOLOGY

Yüksek Lisans Tezi olarak sunduğum, Mining Complex Association Rules Based on Inter-Cluster Communication: Bilgi Erişim Teknikleri ile Kan Şekeri Seviyesinin Tahmini adlı çalışmamın, tarafımdan akademik etik ve kurallara aykırı düşecek bir yardıma başvurmaksızın yazıldığını ve yararlandığım kaynakların kaynakçada gösterilenlerden oluştuğunu, bunlara atıf yapılarak yararlanılmış olduğunu belirtir ve bunu onurumla doğrularım.

01.02.2017

Iman A. F. Hasse

ACKNOWLEDGEMENTS

I wish to express my warm gratitude to my supervisor, Assist. Prof. Dr. Shadi AL SHEHABI, for the guidance and encouragement to achieve and to reach the level of knowledge and experience where I am at right now.

I owe my most sincere gratitude to my husband, Dad and Mom. And I would like to say "without you, I wouldn't make it here, at this place at this time, thanks". And also, I would like to thank my brothers, and friends for the love and the support. Thank you all, and may Allah bless you all.

FEBRUARY 2017 **Iman A. F. Hasse**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Mining Complex Association Rules Based on Inter-Cluster Communication

Hasse, Iman

Master, Department of Information Technology

Thesis Supervisor: Assist. Prof. Dr. Shadi AL SHEHABI

FEBRUARY-2017

Data mining can be understood as the process of examining large amount of structured or unstructured information for the reason of building small and useful summarized information. That information helps will be used later in helping businesses and experts to make - for example - knowledge driven decisions, increasing revenue, and may be future trends. In research, the term data mining as it widely referred to as knowledge discovery from data, describes all techniques and procedures used to extract and summarize large amount of information and yet addresses all challenges and complexities faced by mining such transformation. For example, building associations rules (the task of looking and searching for relationships between variables), clustering (the task of identifying groups of data with similar properties), and classification (the task of searching for new undiscovered patterns). Therefore, the idea of this study is to find the relationships between different subspaces namely viewpoints, these subspaces describe the same set of data and represent it in multi-dimensional space. Therefore, we propose a novel approach, called MCARIC, for extracting complex association rules from two viewpoints based on numeric multi-viewpoint models (such as multi-clustering). When we implement our mining approach we can note that less number of complex association rules is extracted when compared to conventional symbolic methods. Indeed, symbolic methods are not able to extract association rules directly from viewpoints. This study also shows that the proposed approach works on binarized and non-binarized datasets, which is not feasible using symbolic methods.

BIR KAMPÜS ÇEVRE IPV6 DİNAMİK YÖNLENDİRME PROTOKOLÜ UYGULANMASI

Hasse, Iman

Yüksek Lisans, Bilişim Teknolojileri Anabilim Dalı

Tez Danışmanı: Yrd. Doç. Dr. Shadi AL SHEHABI

ŞUBAT-2017

Veri madenciliği kısa ve yararlı özet bilgi oluşturma amacı ile büyük miktarda yapılandırılmış veya yapılandırılmamış bilgi yığınlarını inceleme süreci olarak anlaşılabilir. Bu destekleyici bilgiler daha sonra işletmelere ve uzmanlara yardım amacıyla kullanılacaktır – örneğin – bilgi odaklı olarak alınan kararlar, gelir artırma ve belki de geleceğin trendleri. Araştırmada, veri kümelerinden bilgi keşfi olarak da bilinen veri madenciliği, büyük miktarda bilgiyi seçerek almak ve özetlemek için kullanılan tüm teknikleri ve prosedürleri tanımlamaktadır hal böyle olunca böylesi bir dönüşüm madenciliği ile karşılaşılan tüm zorlukları ve karmaşıklıkları ele almaktadır. Örneğin, ilişkilendirme kurallarının belirlenmesi (başka bir deyişle değişkenler arasında ki ilişkiyi bulmaya çalışma ve araştırma görevi), kümeleme (benzer niteliklere sahip olan veri gruplarını belirleme görevi) ve sınıflandırma (yeni keşfedilmemiş örüntüleri araştırma görevi).

Bu nedenle bu çalışmanın amacı bakış açıları olarak adlandırılan farklı altuzaylar arasında ki ilişkiyi bulmaktır, bu altuzaylar aynı veri kümesini tanımlar ve onu çok-boyutlu bir uzayda temsil eder. Dolayısıyla biz, sayısal çoklu-bakış açısı modellerine (çoklu-kümeleme gibi) dayanan iki bakış açısından yararlanarak karmaşık ilişkilendirme kurallarını bulmak için, MCARIC denilen yeni bir yaklaşım öneriyoruz. Veri madenciliği anlayışımızı uyguladığımızda geleneksel sembolik metotlar ile karşılaştırıldığında daha az sayıda karmaşık ilişkilendirme kuralının seçip çıkarılmış olduğunun farkına varabiliriz. Gerçekte, sembolik metotlar ilişkilendirme kurallarını bakış açılarından doğrudan seçip çıkartma kapasitesine sahip değildir. Bu çalışma aynı

zamanda önerilen yaklaşımın ikili ve ikili--olmayan (binarized and nonbinarized) veri kümeleri üzerinde çalıştığını göstermektedir, ki bu da sembolik metotları kullanarak uygulanamamaktadır.

CHAPTER ONE

GENERAL INTRODUCTION

1.1 Research Methodology

Generally, the aim of knowledge discovery is to "extract implicit, previously unknown, and potentially useful information from data". The increasingly and notable rapid growth of the data currently available automatically, the process of automated knowledge extraction techniques became rather important and required to valorize the ever-growing amounts of data stored in databases, for example, or any information system. Data mining can be understood as the process of examining large amount of structured or unstructured information for the reason of building small and useful summarized information, or in contrast, is Knowledge Discovering in Databases (KDD). Symbolic methods are the most commonly used methods in knowledge extraction from databases, but these methods come with drawbacks, mainly, association rules generation which is an expensive process and produce a huge number of rules. Thus, generating many of redundant rules that may be contained in other rules and this problem makes the rules selection as a complex task. As well, the symbolic methods consider data and its characteristics to be equal in terms of importance, so it could extract imprecise knowledge. Therefore, some attempts have been proposed to extract effective knowledge using the Unsupervised Numerical Classification Models (Clustering Models), which are characterize by their ability to data collection and data separation from each other. But these methods also extract imprecise knowledge because they are working on high dimensional data description space. To cope with this problem, multi

viewpoint approach has been proposed by dividing the high dimensional data description space into different subspaces each subspace is called a viewpoint. Each viewpoint can be represented by a clustering model. This approach makes the knowledge extraction to be more precise than the global one (one description space). Then Inter-Cluster communication mechanism can be used for conserving the relationship between the subspaces.

The study presented in this thesis proposes a novel approach for mining complex association rules based on Inter-Cluster communication mechanism and shows that it generates less and more useful association rules when compared to conventional symbolic methods. The idea basically relies on finding the relationships between different subspaces, namely viewpoints, which describe the same data set and represent it in a multi-dimensional space. To realize a viewpoint, a clustering model is needed where number of objects can be gathered into a group (i.e. viewpoint) based on some similarities among them. Such mechanism is important especially when we deal with a huge database, since we can use the presenter of these data instead of scanning all the data. As a consequence, knowledge extraction from clusters will be more accurate. In this research, we will explain association rules as a strategy of knowledge extraction, but not by the traditional ways as they are expensive, first, and will generate a huge amount of information, second. Another problem has been addressed where simple association rules based on Inter-Cluster mechanism are extracted between an attribute from subspace with another attribute from another subspace. This study suggests another solution which is instead of mining simple association rules, complex association rules mining between subspaces should be used.

1.2 Objectives

1. We want to extract knowledge from databases in an accurate manner by using some methods (symbolic or numeric).
2. We need to explain or interpret some phenomena, but this interpretation may be expensive and difficult to understand by using the actual methods used for knowledge extraction.

3. For that we are trying to reduce the cost and simplify the interpretation of some phenomena by proposing a novel approach based on numeric multi-viewpoint models to give precise interpretation.
4. The aim of this research is to propose a novel approach based on numeric multi-viewpoint models (such as multi-clustering) to extract complex and important association rules between different clusters in order to discover the relationships between item sets belong to different clusters.

1.3 Organization of the thesis

This thesis is divided into six chapters, chapter 2 will discuss the concept of data, its types and the process methods. Also, we will discuss the concept of viewpoint in addition to that we will talk about the concept and goals of data mining and the most important techniques which are used for knowledge discovery such as clustering and association rules. At chapter 3, we will display the most important clustering algorithms and the data clustering concept to reach the correct effective extraction of knowledge. Also, in chapter 4 we will discuss the concept of association rules and the used techniques in its evaluation such as support and confidence. As well as, we will display some symbolic methods algorithms that used in association rules extraction. In chapter 5 we discuss the previous proposed algorithms that extract the symbolic association rules and we will suggest digital tools to extract the complex association rules.

1.4 Literature Review

A new generic multi-topographic neural network model was presented by Shadi Al Shehabi and Jean-Charles Lamirel in [1], with the main area of application of the proposed method is clustering and knowledge extraction from documentary data. The model has many characteristics with the most important being the generalization and communication mechanisms between the topographies. And it was shown how the generalization and communication techniques can be exploited within the NG and SOM models. As well as, and based on the measures of original quality and coherency, the evaluation of generalization mechanisms. Furthermore, the presented evaluation

includes a secondary result which proofs that the generalization mechanisms can expressively decreases the familiar border effect of the SOM map [1].

Jean-Charles Lamirel and Shadi Al Shehabi later proposed a new approach which extends the numerical models scope by providing them the knowledge extraction abilities. The strength of the proposed model can be realized through the generalization and communication mechanisms between topographies. These two mechanisms permit the simple rule extraction to be accomplished whenever a single or multiple viewpoints of the same data are considered. The extraction process of an association rule is naturally dependent on the measurement of the original quality, which assess the numerical classification model as it acts as a natural symbolic classifier, like the Galois lattice [2]. The original quality measures which are Recall and Precision have been proposed to assess the cluster analysis quality. The new proposed measure was developed from the Galois lattice theory and from the domain of Information Retrieval (IR) [3]. The results of the first experiment on patent dataset show how their measures can be used in order to compare viewpoint-oriented classification techniques, such as MultiSOM, with global cluster analysis technique, such as WebSOM [4].

Also, and from their second experiment which takes part in the EICSTES EEC project, they represent an original Webometrics experiment that combines contents and links classification beginning from huge non-homogeneous web pages set. Thus, the experiments highlighted in their research shows that break-even points among different measures of Recall/Precision can be used to conclude an optimum number of clusters for representing the Web data [3][4].

Leeuwen and Galburun proposed in [5] a novel method that finds a set of non-redundant association rules that describe how two views, of the same dataset, are related. Two-view datasets are the datasets where their characteristics are divided into two distinct sets each of which offer dissimilar view on the same objects set. They presented the duty of finding non-redundant and slight sets of associations which pronounces how the association of the two views can be realized. To accomplish this, they proposed a

novel method where the sets of rules are utilized where for one view can be translated to another and vice versa. Their models are called translations tables of which each is consist of both bidirectional and unidirectional rules that span both views and offer lossless translation from one of the views to the opposite one. As well as, they presented a score based on the Minimum Description Length (MDL) standard for the purpose of being able to estimate different translation tables and accomplish model selection. Next, in order to find models according to this score, they introduced three TRANSLATOR algorithms. The first one is parameter-free algorithm and iteratively adds the rule that progresses compression. The other two algorithms use heuristics to attain better trade-offs between runtime and compression. The assessment on real-world data shows that the only modest numbers of relations are needed to distinguish the two-view structure existing in the data, while the gotten translation rules are easily interpretable and deliver awareness into the data. T-Select, just as the symbolic methods, still consider data and its characteristics to be equal in terms of importance, thus it fails to isolate the important and strong relationships from the weak ones.

CHAPTER TWO

DATA REPRESENTATION AND KNOWLEDGE DISCOVERY IN DATABASE

2.1 Introduction

The amount of data that is been collected on daily bases by advanced information systems has increased hugely in the last years. So, in order to analyze this huge amount of data, a process has been suggested known as data mining or knowledge discovery in the databases (KDD). We will focus on the most trendy (popular) pattern discovery method in regards to data mining which is association rule mining at this chapter. In addition, we will focus on the binary data and it's processing for the purpose of clustering analysis and knowledge discovery to extract the complex association rules. Finally, we will illustrate the preprocessing of the numerical data through the data indexing process and we will show the representation method to that data for the clustering and knowledge discovery process to extract the numerical association rules.

2.2 Data Representation

There are many types of data like document data, binary data, textual data, sequential data, and so on. In this research, we will focus on textual data and binary data.

2.2.1 Textual Data

If we have N number of documents and the set of that documents contain M number of items, the natural representation of that documents will be through the documents matrix which dimensions of M*N. While M refers to the number of rows, N represents the number of columns. Before the indexation process, the elements always go back to their languages. For instance, the word of *jealousy* and *jealous* will be represented in one item which it is the same with one dimension in the document matrix. As we mentioned earlier, there are two methods to represent the document, the first one to mention is the Boolean model and when it comes to the second one it is the vector model [6]. At this chapter, we will discuss the vector model.

2.2.1.1 Term Weighting

Weight can be inserted into each document. The weight will depend on the number of term appearance in the document. The scheme of weight refers to the frequency of the term which symbolizes by $TF_{t,d}$. Consequently, these letters denote to the term and document correspondingly. The group of weights can be represented by a vector with one factor separately for each term. Towards this end, a weight of that term will be assigned to every term within the document which completely vary from one term to anther as it represents the number of how many times that particular term appeared in the document. We'd like to calculate the score between t which refers to the term of the query and d which refers to the document that is based on the weight of t and d [6].

2.2.1.1.1 Inverse Document Frequency

The consideration which can make the whole terms in the document are equal in the significance make data mining process unusable and comprise a lot of errors. For instance, if we take Reuter agency documents about the industry of assurance, we can notice that the assurance mostly founded in each document. Due to that, there is a technique which works to decrease the term appearance effect which recurring in a document set. As the main impression, the terms weights with high frequency should be

decreased by these terms repetitions appearance in the set. The idea is accomplished through the term frequency reduction by the use of factor related with term frequency in the et. However, it is useless to use the document frequency as a set document which comprises the term. When we try to differentiate between the documents, we can use the statistics on the documents level better than using the statistics on the comprehensive set level during the use of collection frequency. The collection frequency could be identified as the term repetition appearance number in the set of the document [7]. Table 2.1 shows a simple example of the reason behind the frequency of preference document to collection frequency where the behavior of difference to document frequency and collection frequency can be notable. Particularly, the collection frequency value for "try" and "insurance" almost equal but the frequency value of the document is the difference. The weight of "insurance" word bigger than of "try" word in spite of "insurance" can be found mush less of "try".

Word	CF	DF
Try	10422	8760
Insurance	10440	3997

Table 2.1 The difference behavior of collection frequency and document frequency in Reuter's collection

The term of document frequency is utilized so as to find the weight as follows:

$$IDF_t = \log \frac{N}{DF_t} \tag{2.1}$$

Since N refers to the total number of documents contained in the collection, thus, the inverse document frequency of seldom term (less repetition) might be high while the inverse document frequency value of the frequent term is low. As displayed in Table 2.2, the inverse document frequency values document set with 806791 in Reuter's agency. In addition to different terms frequencies values that taken from those documents.

Term	DF_t	IDF_t
Car	18165	1.65
Auto	6723	2.08
Insurance	19241	1.62
Best	25235	1.5

Table 2.2 Inverse document frequency and document frequencies values of a set of terms that taken from Reuters agency document

2.2.1.1.2 Weighting Function

To realize the combined weights of each term in the document, the term frequency and inverse document can be merged. According to the following equation, the term will be weighted in the document:

$$TF - IDF_{t,d} = TF_{t,d} \times IDF_t \tag{2.2}$$

On the other hand, $TF - IDF_{t,d}$ explains term's weight found in the document. Thus, and through the presented equation (2.2), we can notice the following statements:

1. The weight is being high when the term repetition is in documents which are in a small number (this'll give these documents high distinguished potency).
2. The weight is being low when the term repetition is in a few documents.
3. The weight is being lower when the term found actually in the whole documents. Thus, each document can be denoted by the vector and the specialized one content with every item in the dictionary and in addition to its weight according to the (2.2) equation.

2.2.1.2 Vector Model

In radial space, the documents are represented by a model called vector model [6] [7]. The vector model represents basis in order to clustering and classifying data in addition to the data mining processes. This model uses the language term in order to represent data by the vector and this representation disregard of items from each other. Consequently, the vector stored items presence number in a document without interest of ordered or arranged between each other. Through this representation, we can found what characterize the phrase "Ahmed is faster than Ali" is similar to "Ali is faster than Ahmed". Thus, axiomatically we accomplish that the two documents similarity in contains if they are similar in their vectors.

2.2.2 Binary Data

Binary data can be thought as a special case of either multidimensional categorical data or multidimensional quantitative data. Its case, which is special regarding multidimensional categorical data, in which each of the categorical attribute might take one of two discrete values at most. It’s a special case of multidimensional quantitative data due to an ordering that exists between the two values as well. Moreover, binary data is also a representation of set wise data, in which each of the attribute is treated as a set element indicator. A value of 1 signifies that the element should be incorporated into the set. Such data is common in regards to market basket applications [8][9].

2.2.2.1 Discretization

Several data mining techniques, a specific type of classification algorithms in particular, require the data to be presented in categorical attributes form. Where also, some specific algorithms require the data to be in binary attributes form to be able to extract association rules. So, it’s often required that data stored in continuous or in discrete attribute form to be transformed to binary (the process of binarization) or to categorical attribute (the process of discretization) to, then after, be used. In addition,

and especially when a categorical attribute contains a huge number of values (categories) or when it contains values which rarely occurs, it's advisable, and in certain data mining tasks might be valuable, to minimize the number of categories through combining some of the values in groups. Since with feature selection, the discretization and binarization approaches are the ones which "generates the best result for the data mining algorithm which will be utilized to analyze the data". However, it's typically not advisable to implement those transformations directly at every given data mining tasks. As a result, discretization or binarization should only be administered only if it satisfies the criterion of "if applying such transformations is related to improving the performance (time or space) of the data mining task being considered or, to produce a better-quality results" [8][9].

Categorical Value	Integer value	x_1	x_2	x_3
Awful	0	0	0	0
Poor	1	0	0	1
OK	2	0	1	0
Good	3	0	1	1
Great	4	0	0	0

Table 2.3 Categorical to binary attributes conversion

Categorical Value	Integer value	x_1	x_2	x_3
Awful	0	0	0	0
Poor	1	1	0	0
OK	2	0	0	0
Good	3	0	1	0
Great	4	0	0	1

Table 2.4 Categorical to binary attributes conversion
(five asymmetric binary attributes)

2.2.2.2 Binarization

Binarization is a technique used to transform a categorical attribute into binarized form, and it can be understood as follows: If we assume that there're a m categorical values, then for each original value, assign a uniquely an integer that is within the interval of [0, m-1]. However, if the attribute is ordinal, then the order of the assignment must be maintained. It's important to take notice that if the attributes are originally represented using integers values, then attributes assignments shouldn't be within the intervals of those integers' values. Following this, the assigned m integers' values are converted into a binary number. The binary digits which are required to symbolize these integers is $n = [\log_2(m)]$ [9].

To exemplify, let consider the following categorical variable containing the values of {awful, poor, OK, good, great}, which should require three binary variables x_1, x_2 and x_3, the conversion presented in Table 2.3. However, such transformation could lead to forming unintended relationships among the transformed attributes. For instance, in Table 2.5, x2 and x3 are correlated attributes, since both of the attributes encode information about the same good value. Moreover, analysis of association requires asymmetric binary attributes, where the existence of the attribute of value=1 is deemed as important. With such problems of association, it is then, required to acquaint one binary attribute to each categorical value as illustrated in Table 2.4. However, when the resulting attributes' number is too large, in that case the technique which is defined below could be utilized in order to reduce it before binarization. Similarly, and in some cases, a single binary attribute should be replacing with two symmetric binary attributes, not one. We can think of a binary attribute which represents a person's gender, whether they are male or female. In the traditional association rule techniques in data mining, such information must be transformed into two symmetric binary attributes, one attribute is set to 1 only when the person is male, where the second is set to 1 only when the person is female. In such case, we can see that for a symmetric binary attribute require two bits of storage to represent one bit of information, which is somewhat inefficient.

2.3 Viewpoint Notion

The viewpoint building attitude involves in description space separating of the documents into different subspaces, each of which is consistent to unalike subsets of keyword.

The V set of the whole possible viewpoints which are delivered from the description space, represented by D, of a document set could be identified:

$$V = \{v_1, v_2, \dots, v_n\}, v_i \in P(D), with \bigcup_{i=1}^{n} v_i = D$$

Each of the v_i signifies a viewpoint, whereas $P(D)$ embodies the parts set of the description space of the documents D; the union of the different viewpoints institutes the documents description space.

The sum of viewpoint subsets which are delivered from V equation might have some overlapping ones. As well as fitting into the document structure, when the viewpoint resembles to different index vocabulary subsets related to document sub-fields which are different in turn. However, and regarding the framework for a documentary database, explicit viewpoints might be related to precise reference fields such as "title keywords", "indexer keywords", or "author" field. Furthermore, extracting opposite viewpoints might be possible through the use of complete document description space. The concept of viewpoint is more common than the one of document field. It's always likely to find a viewpoint which characterizes the description space utilized in the document field. As an example; we have a market-basket that contains many items like (milk, juice, bread, diaper, coke, eggs), and we will distribute those items into two sets which called views(viewpoints)

{milk, diaper, bread, eggs} ⟶ view1(viewpoint1)

{juice, coke} ⟶ view2(viewpoint2)

2.4 Data Mining and Knowledge Discovery in Databases (KDD)

Data mining plays an important part of knowledge discovery in databases (KDD), which for one reason to be considered is that data mining techniques are used in converting raw data into helpful information.

2.4.1 Data Mining Concept

Data mining can be defined as "*the nontrivial extraction of implicit, previously unknown, and potentially useful information from data*" [12] and "*the science of extracting useful information from large data sets or databases*" [13]. Generally, data mining is characterized as the process of analyzing data through the different standpoints as well as generating a concise summary of useful information. And technically, it's the process of finding correlations or patterns among dozens of fields that are in large relational databases [1].

2.4.2 Process of KDD

As depicted in the Figure 2.1, which describes the role of data mining as an essential part of knowledge discovery in databases. Which consists, in general, of all processes of converting data into useful information. Which in turn consists of steps of transformation, from data preprocessing to post processing of data mining results [10].

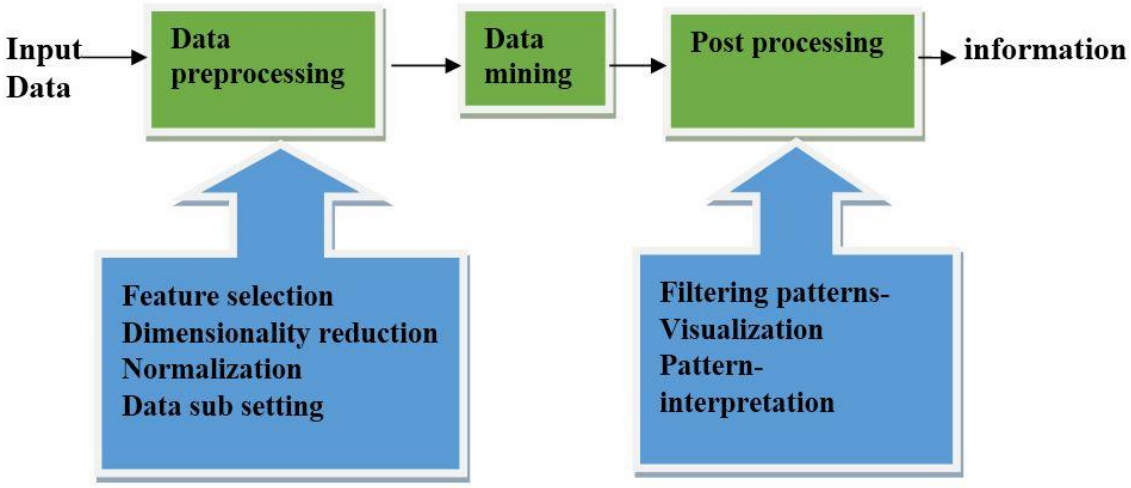


Figure 2.1 The process of KDD [10]

The data can be stored in different formats (flat files, spread sheets, or relational tables) and may stay in a centralized data repository or be distributed over multiple sites. The aim of preprocessing is to transform the input data into a proper format for subsequent analysis. The steps which is in data preprocessing include fusing data from multiple sources, cleaning data to remove noise and duplicate observations, and selecting features that are relevant to the data mining task at hand. Because of the many ways data can be collected and stored, data preprocessing is probably the most time-consuming step in overall knowledge discovery process. “Closing the loop” is the phrase used to refer the process of integrating data mining results into decision support system. Such integration requires a post processing step that ensures that just valid and useful results are incorporated into the decision support system. An example of post processing is visualization, which enables analysts to explore the data and the data mining results through a variety of viewpoints. Statistical measures or hypothesis testing methods could be implemented during postprocessor to get rid of wrong data mining results [10][11].

2.4.3 Data Mining Techniques

The data mining techniques have compatible methods in order to identify the forms of models. These compatible methods have a major role in knowledge discovery. Thus, the models which reflect the importance and the benefit of knowledge is considered as a part of KDD processes [10]. Therefore, in order to go properly in the compatibility operations, there are two mathematical operations one is statistical and the other is logical. The statistical methods have unavoidable impacts on the model whereas, the logical methods are significantly vital. In our study, we will significantly focus on the statistical methods due to they’re widely utilized in various significant applications for the field of data mining. The data mining depends heavily on well-known methods which are the statistical and machine learning methods such as clustering, association rules, and some other techniques [11].

2.4.3.1 Clustering

Clustering is the process of unifying objects into groups where each of which these objects are similar in some way. A cluster is a collection of objects which shares some similarities and are all “dissimilar “ to the objects found in other clusters [14]. Clustering can be considered the most important unsupervised learning problem; so, as every other problem of unlabeled data. The aim of clustering is to designate the intrinsic grouping in a set of unlabeled data. It will be discussed in details later.

2.4.3.2 Association rules

Given a set of transaction, find rules that will predict the occurrence of an item which is based on the occurrence of the other items in the transaction. If we considered the itemset on the left side of the figure below, which represents the market-basket purchases, a sample of generated association rules generated from such itemset is described below. Association rules will be explained in details in Chapter Four.

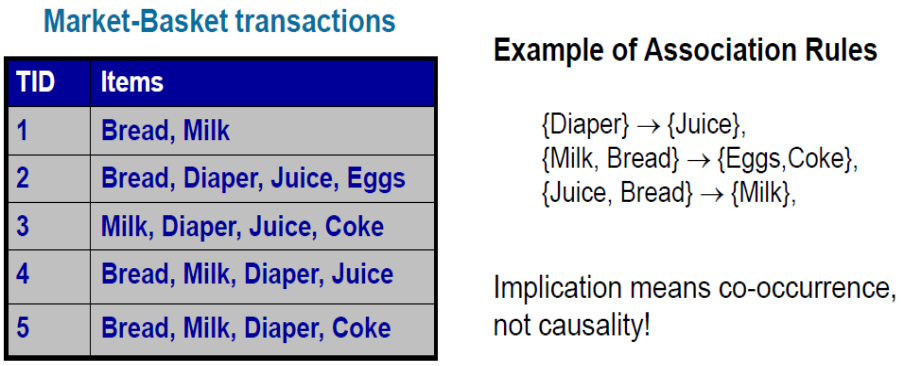


Figure 2.2 Association rule example

2.5 Conclusion

At this chapter, we have identified some data types which will be the center of our attention at this research and also the ways of processing. As well as, the concept of data mining has been discussed which considers the basic of knowledge extraction and the processing of data by data mining techniques where the clustering is the basic technique to be applied on the data during this research. Through the clustering, we will discuss the extraction of association rules that is one of the more important required figures.

CHAPTER THREE

CLUSTERING ANALYSIS

3.1 Introduction

Clustering can be understood as the process of searching and associating object or objects in groups, such that the objects which are in a group - have some or completely - similar (or associated) (these objects share common characteristics) to one another. And different or completely unrelated to the objects present in other groups [14]. In this chapter, we will present some clustering methods. Also, we will present some clustering evaluation measures in order to find the optimal one. This optimal model will be used later in this thesis for extracting the association rules.

3.2 Clustering Analysis

In most of the applications, the clustering is not defined correctly. Figure 3.1 clearly illustrates the cluster which displays 20 points and three different approaches to dividing it into clusters. Figures 3.1 (b) and 3.1 (d) shows an arbitrary data sample divided into two and to six groups, respectively. Nevertheless, the division that is available for the two large clusters into three sub-clusters is simply the tool for the human vision system. As well as, it might not reasonable to state that points of four clusters as illustrated in Figure 3.1 (c). The same graph also explains that the cluster definition is inaccurate and the accurate definition of it will be according to the nature of the data and the required results. The cluster derives the labels only through the data, whereas, the classification is considered supervised classification and the unnamed

figures been assigned to the class name which uses model been developed from the object with famous class. Furthermore, in the case of if the cluster was used without any qualification, the analysis of the cluster is considered unsupervised classification. As well as, the cluster is often called anatomy or partition and these terms are used for the styles which are outside the traditional frame for analyzing the cluster. However, there are other names calling the cluster such as unsupervised learning. In addition, the cluster is used to get a thought through to the available dataset. Thus, the cluster is always enough to give us a thought about data distribution within the dataset. We gave a small example of the cluster such as collecting the customers for gathering documents and shopping, for the purposes of adjusting the answers performance in search engines and to realize the models in the spatial or temporal deliveries to the diseases. In this chapter, we will exhibit on disparate forms for the clustering algorithms which contain central base to create the asset groups [14][15].

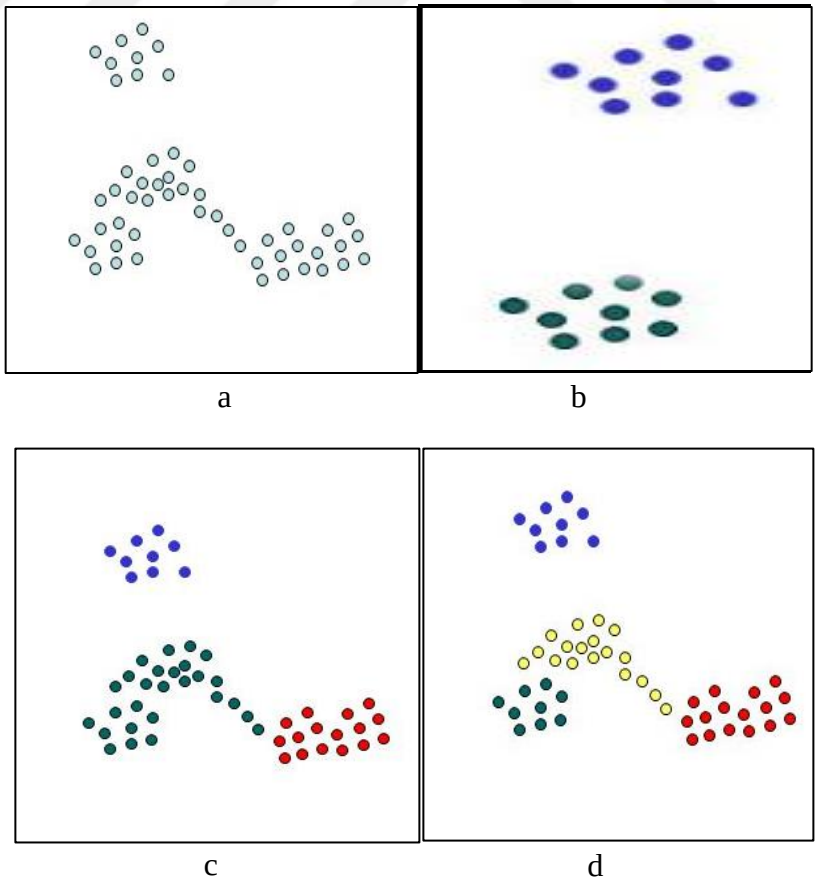


Figure 3.1 Different numbers of clusters of the same set of points

This includes the clustering methods which represent the data and it's perfectly and correctly. This procedure requires to inquiry and estimate for the clustering goals through the mathematical models which will be showed during this chapter. We will start by the perfect cluster in knowledge discovery. One of the most important methods in clustering and classifications is the artificial neural network which includes multi-gas, neural gas, self-organizing map, and multi self-organizing map. The artificial neural network and according to its topographical terminology which came from the relationship of the neurons that mimic in their mode of work to the thinking process of the human. In our study, we will work on a multi self-organizing map which has various learning functions that available in the data through the knowledge discovery. Also, it has weak to enough accuracy to invest it in knowledge discovery. As well as, this chapter will focus on use the unsupervised learning processes which represent the neuron combined which is specifically use in knowledge discovery to produce the composite association rules [14][15][16].

3.3 General Trends on Clustering Algorithms

The clustering algorithms depending on different styles can be categorized into multiple classifications as follows [15][16]:

3.3.1 Partitioning Approach

The algorithms of partitioning clusters work on partitioning data which consist of n elements within separated cluster k. Also, each cluster consists of at least one element and each element belongs to just one cluster. In order to get good clustering, the clustering methods work on dividing a dataset into primary partitions. Then, the work will be on enhancing the clusters quality through multiple iterations and some data elements will be moved from one cluster to an alternative in each iteration. Actually, the partitioning clustering help on directing but they are not guaranteed to get high quality in the partitioning time. The “k-means” algorithm is considered as one of the widely used and rather important among known partitioning algorithms, an example of

k-means is showed in Figure 3.2. Furthermore, k-means algorithm is useful to find k which is the circle of the cluster to labeling the data set. We can notice that algorithms of partitioning cluster need to identify prior to the number of clusters that must be initiated [15][16].

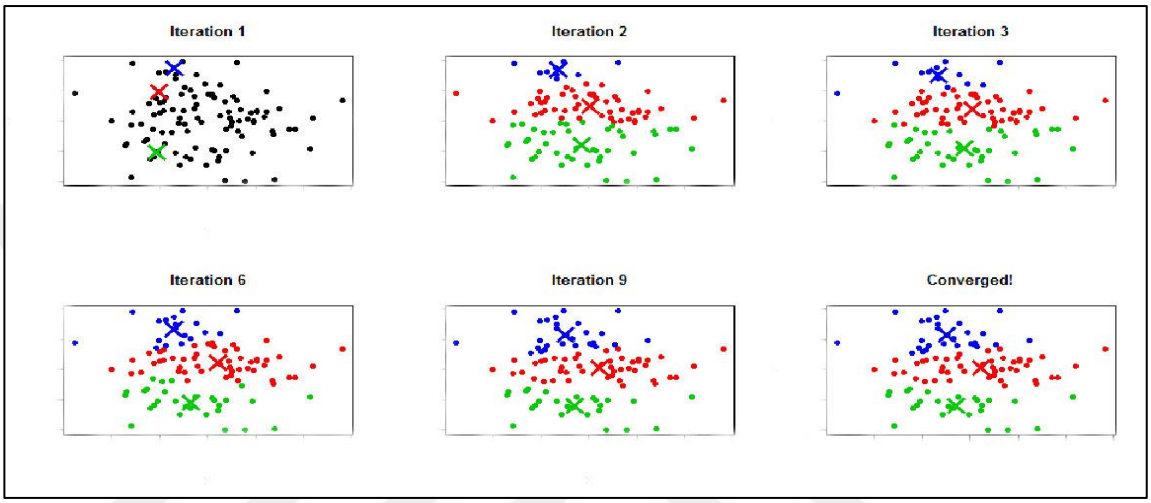


Figure 3.2 k-means algorithm implementation

3.4 Unsupervised Neural Network Algorithms

At this step, we will talk about some unsupervised algorithms such as NG and SOM which we will talk about them in details in the next step. The word unsupervised refers to the clustering process and this type is the basic of our work at this research because we do not have previous idea about classes of data or inputs which can be generated by the network and from it we will implement the knowledge extraction. As we mentioned earlier, it includes many techniques including the following:

3.4.1 Self-organizing Map

Kohonen self-organizing map is considered as one of the widely adopted and rather important approach which associated by unsupervised learning approaches used to organize data with different dimensions in neural network form [17]. The approach depends on the categorization and drops data of limited dimension network. The

approach used effectively and progressively in many applications and especially in mining texts field. Furthermore, the self-organizing map has numerous components of neural processing which known as neurons and be structured according to specific network forms. As well as, self-organizing map be in the form of the vector with one direction always organized as the map of rectangular dimensions with two, three or six dimensions as explained in Figure 3.3.

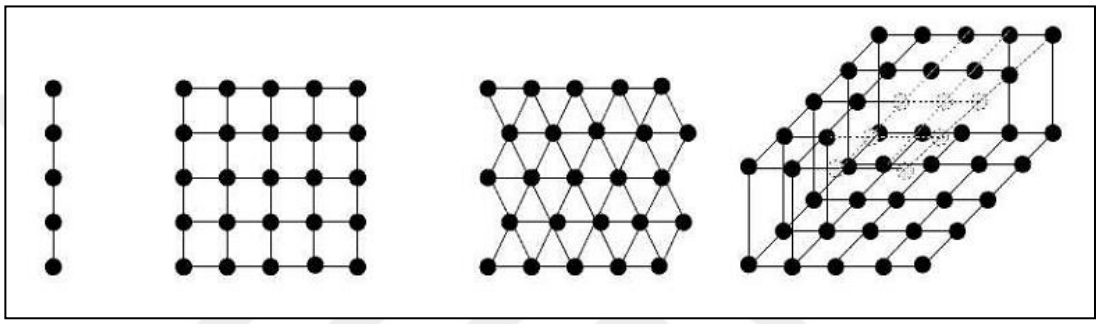


Figure 3.3 Self-organizing map patterns

Then, each c unit will denote to W_c vector where $W_c \in \Re^n$. Also, these vectors are characterized as to have the same dimension number [16][17]. In order to the conditioning the neuron, the distance on this network is used as a standard $r = a_{km}$ where the neuron $s = a_{ij}$ epitomizes the winner. This distance is "Manhattan". The complete algorithm to Self-Organizing Map is showed in Figure 3.4.

Self-Organizing Map Algorithm

---

1. *Initialize the weight vectors,* $W_{c_i} \in \mathcal{R}^n$
2. t=0
3. *for epoch=1 to* N_{epochs}do
4. *for input =1 to* N_{inputs}do
5. t=t+1
6. for x=1 to X do

7. *compute distance* d_k, $s(x)=\text{argmin}_{c\in A}||x-W_c||$
8. *end for*
9. *compute BMU for current input,* $d_c(t)=min_k$, $d_k(t)$
10. *for k=1 to k do*
11. $\Delta W_r = \in (t).h_{rs}.(x - W_i), r = 1, 2, \dots, N$
12. *end for*
13. *end for*
14. *end for*

Figure 3.4 Self-Organizing Map algorithm

$$d(r,s) = |i-k| + |j-m| \quad (4.1)$$

The set A is prepared with neural c_i that its number is $N = N_1.N_2, A = \{c_1, c_2, \dots, c_N\}$ within the vector that it support $W_{c_i} \in \Re^n$ that randomly chosen and $N_1$and N_2 represent two rectangular dimensions [16][17]. Also, to prepare the time medium with primary value where h_{rs} is the consequence neighboring neuron which is used in order to define the relative strongest to the condition neurons of the neural networks which has numerous kinds. The Gaussian consequence and Bubble consequence is showed in Figure 3.5, which be given in the relation:$h_{rs} = \left(\frac{-d(r,s)^2}{2\sigma^2}\right)$.

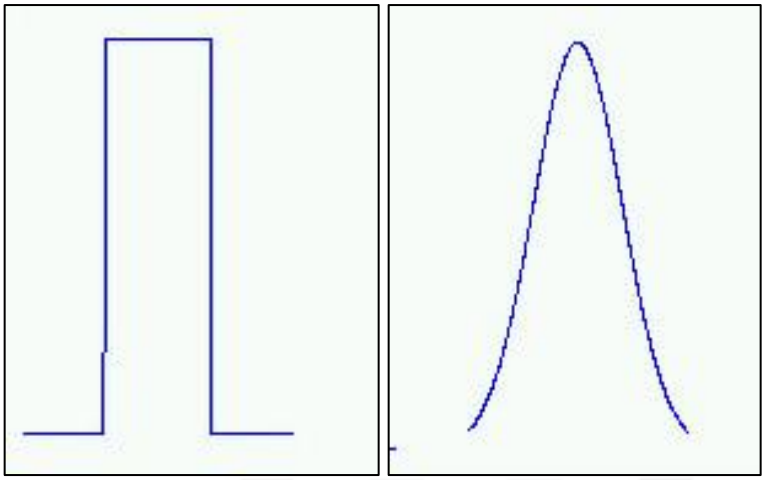


Figure 3.5 Neighboring neuron consequence (Bubble and Gaussian)

It is given $\sigma(t)$ and average of learning network $\in (t)$ according to the following relations:

$$\sigma(t) = \sigma_i \left({}^{\sigma_f}/_{\sigma_i}\right)^{{}^{t}/_{t_{max}}}, \in (t) = \in_i \left({}^{\in_f}/_{\in_i}\right)^{{}^{t}/_{t_{max}}}.$$

It must choose the suitable primary values medium magnitudes $\sigma_i, \in_i$ and the final values $\sigma_f, \in_f$. It is tiny state to the text data that exist in the space within n dimensions, where (a) denotes to the network initial case (from b to f) medium cases after learning repetitions group, (g) is the network final case and (h) Voronoi which agree with the final cases (Voronoi region is the set points space where these points are near to neuron which agrees with other neurons) where:

$$\sigma_i = 10, \sigma_f = 0.01, \in_i = 0.5, \in_f = 0.005, t_{max} = 40000, N_1 = N_2$$

3.4.2 Multi Self-Organizing Maps (MultiSOM)

The MultiSOM is an extension to the original SOM algorithm and been suggested by [18]. This model uses numerous points of the view and each single one is denoted by single SOM map, all of which are used for the reason of improving the quality and the divisions of points sets required in data analysis. And to decrease the

noise that it absently generates the classification approaches. All the analysis views have protected by using the connections machines between maps. The benefit of multi-views analysis which introduced by MultliSOM and compared with SOM is clearly showing the accurate mining process such as patent analysis. Also, the generalization mechanism is considered a practical method been supplied by MultliSOM [18][19]. However, special care must be taken into consideration to the widely-known problems which connected by the structure of the design that called SOM called as the border effect. The border effect means the units of network edges do not extend to outside as should be towards the radical values of data. The neural networks of SOM are not necessary to close to the data structure because of the network structure [18]. In our research, we will focus only on inter-clustering communication mechanism.

3.5 Inter-Cluster Communication

Inter-cluster communication mechanism makes it possible to focus on important relationships between different topics which belong to different subspaces [1]. The inter-cluster communication process between clusters operates in three step:

First, the original activity is directly set on a source cluster, directly associated to the data. Then, the transmission from the source to target cluster is by itself based on two transmission steps; the first is from the activated source cluster to its associated data, whereas the second transmission step is from the activated data to the target cluster [1].

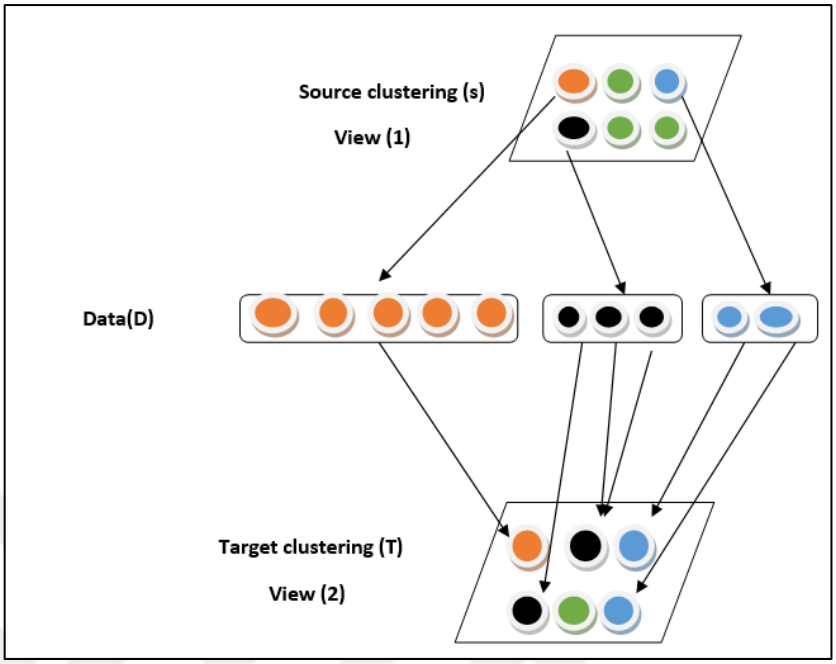


Figure 3.6 Inter-Cluster communication [1]

3.6 Evaluation of Clustering Approaches

As we mentioned previously, there are three clustering approaches. In order to know which one of the clustering approaches is suitable for some type of data, we must use many of mathematical functions. Moreover, the number of clusters which are generated by using these approaches cannot be generated perfectly by using only these mathematical functions. The best model which represents the optimal number of clusters can be chosen by using these methods to represent data and carry out the knowledge discovery processes. There are numerous methods which are used to assess the clustering results and choose the best and the optimal model [20][3]. Two ways of these methods will be mentioned through this thesis.

3.6.1 Evaluating of Clustering Based on The Characteristics of Distribution Data

Assume that we have a group of generated clusters from clustering methods have been applied to a group of documents, then the recall is (*Rec*) and the precision is (*Pre*ss) to a particular feature (P) in the cluster can be stated as follows [3]:

$$Prec(p) = \frac{|c_p^*|}{|c|} \tag{4.3}$$

The precision in the cluster denotes the ratio between the document number which contains the feature in the cluster and between the documents number which been found in the same cluster. The precision criterion dealings with the extent of the generated cluster of clustering which mean that data which been found in one cluster have the same features.

$$Prec(p) = \frac{|c_p^*|}{|c_p^*|} \tag{4.4}$$

Whereas, the recall of the cluster represents the ratio between the documents that contain the feature in the cluster and between the whole clusters that have the numerous feature. Thus, the recall measures to what extent the cluster contents which have autonomous features from one cluster to another.

$$c_p^* = \{d \in c | W_d^p \neq 0\} \tag{4.5}$$

Where c_p^* denotes to the restriction of the set c to the members of the set which having the features p. Therefore, the measurement of the clustering precision can be expressed as follows:

$$P = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{S_c} \sum_{p \in S_c} \frac{c_p^*}{|C|} = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{S_c} \sum_{p \in S_c} Prec(p) \tag{4.6}$$

Where S_c denotes to the set of the features that are proper to the cluster c which can be found as:

$$S_c = \{p \in d, d \in c | \bar{W}_c^p = Max_{c' \in C}(\bar{W}_{c'}^p)\} \tag{4.7}$$

Where $\bar{C}$ denotes the irregular set of clusters which are extracted from the entire C clusters:

$$\bar{C} = \{c \in C | S_c \neq \emptyset\} \tag{4.8}$$

$$\bar{W}_c^p = \frac{\sum_{d \in c} W_d^p}{\sum_{c' \in C} \sum_{d \in c'} W_d^p} \tag{4.9}$$

Where W_d^p denotes to the features weight p for the element d. The measurement of recall clustering could be expressed as follows:

$$R = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} \frac{|c_p^*|}{|c_p^*|} = \frac{1}{|\bar{C}|} \sum_{c \in \bar{C}} \frac{1}{|S_c|} \sum_{p \in S_c} Rec(p) \tag{4.10}$$

Whenever the precision increase, the document topics which belong to the same cluster will be closed to one. This helps the user to clarify the content of the cluster. Thus, the comprehensive of the mentioned cluster will be measured by the recall. Also, the recall evaluates the existence feature in an individual cluster. Finally, the measurements of recall and precision will be used to compare between different clustering approaches.

3.7 Conclusion

In this chapter, we presented a set of clustering methods which are known as unsupervised methods. We have focused on the unsupervised neural network methods, such as Self-Organizing Map (SOM) and its alternative which is MultiSOM. SOM is applied on one dataset or one viewpoint whereas MultiSOM is applied on several viewpoints of the same set of data. This method provides Inter-Communication mechanism which is used for conserving the overall view of the viewpoints. We will use MultiSOM as a basis for extracting association rules between viewpoints but one of drawbacks of the clustering method is how we can find its optimal number of clusters for representing a given dataset. Therefore, we presented some measures for evaluating the clustering method and in turns it can find the optimal number of clusters.

CHAPTER FOUR

ASSOCIATION RULES EXTRACTION BASED ON SYMBOLIC METHODS

4.1 Introduction

In this chapter, association analysis will be explained in details which has great advantage in extracting the important relationships that existed within huge set of data. It is possible to represent the shown relationships on the form of association rules to set of frequent items as we will notice later. As well as, at this part of the thesis, we will display one of different methods to extract the data which is the symbolic methods that consider the most common to extract knowledge but it is so expensive when process the huge size of the databases [21]. In another hand, the digital methods proved the high efficiency when processing this size of data in knowledge extraction. Thus, in our thesis we will depend on the digital methods in order to evaluate and analyze the extracted knowledge as we will discuss it in the fifth chapter.

4.2 Association Rules Definition

A group of data and a group of graphical data will be taken and each graphical element contains on a number of elements been taken from that sets of elements. Association subsequences is an opposite process to dataset that return models or associations which found between the set of items. These models can be presented by 72% of records which contain items A, B, C, and in addition, the items of D and E. The case with 72% of records is called the confidence factor of the association rule. The A, B

and C items in this rule are forming the reverse part of the items D and E part. This means the rule form of the figure $\{A, B, C\} \rightarrow \{E, D\}$. In any one of these parts the association rules can include any number of items [22].

4.2.1 Association Rules Evaluation

The measurement of interestingness of different association rules is needed to quantify the dependency among rules, and it is calculated between and for each rule with the consequent and antecedent ones. One of the common rule assessment measures, which are commencing by support and confidence are given as follows:

Tid	Items
1	ABDE
2	BCE
3	ABDE
4	ABCE
5	ABCDE
6	BCD

Table 4.1 Dataset example

Sup	*Rsup*	Itemsets
3	0.5	ABD,ABDE,AD, ADE, BCE, BDE, CE, DE
4	0.67	A, C, D, AB, ABE, AE, BC, BD
5	0.83	E, BE
6	1.0	B

Table 4.2 Frequent itemsets extracted with *min. sup* = 3, with relative minimum support of 50%

4.2.1.1 Support

The definition of rule's *support* is the number of transactions which consists of X and Y, and can be represented as:

$$sub(X \rightarrow Y) = sub(XY) = |t(XY)| \tag{5.1}$$

However, the fraction of transaction which consists of both X and Y (i.e. empirical joint probability of itemset's) defines the relative support. And it is comprising the following rule:

$$sub(X \rightarrow Y) = P(XY) = rsub(XY) = \frac{\sup(XY)}{|\boldsymbol{D}|}$$

The attention is typically given to frequent rules with $sub(X \rightarrow Y) \geq minsup$, where $minsup$ is a threshold specified by the user. When minimum support is determined as a fraction, following relative support is implicated. An important note to consider is that the relative support is a symmetrical measure, since:

$sub(X \rightarrow Y) = sub(X \rightarrow Y)$.

To give an example of rule assessment measures, let's consider the binary dataset D depicted in Table 4.1, which is written in transactional form. It comprises of six transactions over a set of five items $\mathcal{T} = \{A, B, C, D, E\}$. The set regarding all of frequent itemsets by $minisup = 3$ which is listed at Table 4.2. The table demonstrates the support as well as relative support which are for each of frequent itemset. The association rule $AB \rightarrow DE$ that is derived from the itemset ABDE has got support $\sup(AB \rightarrow DE) = \sup(ABDE) = 3$, and its relative support is:

$rsupAB \rightarrow DE = \sup(ABDE)\ /|\boldsymbol{D}| = 3/6 = 0.5.$

4.2.1.2 Confidence

The confidence of a rule is defined as the conditional probability of which a transaction includes the consequent Y that is included in the antecedent X:

$$conf(X \rightarrow Y) = P(Y|X) = \frac{P(XY)}{P(X)} = \frac{rsup(XY)}{rsup(X)} = \frac{\sup(XY)}{\sup(X)}$$

Rule			Conf
A	$\rightarrow$	E	1.00
E	$\rightarrow$	A	0.80
B	$\rightarrow$	E	0.83
E	$\rightarrow$	B	1.00
E	$\rightarrow$	BC	0.60
BC	$\rightarrow$	E	0.75

Table 4.3 Rules confidence

Typically, the interest is only given to rules with high confidence value, with $X \rightarrow Y \geq minconf$; where $minconf$ is a threshold specified by the user. Confidence isn't a symmetric measure if compared to support measure, since by the definition it's conditional as it depends on the antecedent.

Table 4.3 demonstrates association rules generated from dataset in Table 4.1, along with their calculated confidence values. If we consider the rule $A \rightarrow E$ from the same table, the confidence value is $\sup(AE)\,/sup(A) = 4/4 = 1.0$. We can also notice the symmetry regarding confidence, observe the $E \rightarrow A$ rule which has confidence value of $\sup(AE)/\sup(E) = 4/5 = 0.8$.

Much care should be given to understand the goodness for each extracted rule. For example, if we consider the rule of $E \rightarrow BC$ with confidence of $P(BC \backslash E) = 0.06$, which can be interpreted as: given E, we've probability of 60 % chance of finding BC. Whereas if we calculated the unconditional probability of BC from $P(BC) = 0.67$, such value can be interpreted as: that E, actually, has a deleterious impact on BC [25].

4.3 Symbolic Methods

The symbolic methods are the most common used methods of effective knowledge extraction from the huge databases. Many methods such as Apriori and Close can be used which through we can extract the association rules [22].

4.3.1 Apriori Algorithm

All possible itemsets in its quest are enumerated by the brute force approach to be able to determine the ones that are frequent. As a result of such approach, a huge waste of resources will occur and most importantly is the possibility of generating numerous itemsets which are not frequent. Let $X, Y \subseteq \mathcal{T}$ be any of two itemsets, with if $X \subseteq Y$, then $\sup(X) \geq \sup(Y)$. Such condition can be abbreviated as follows: (1) if X is frequent, in that case any of subset $X \subseteq Y$ is frequent as well, and (2) if X isn't frequent, in that case ay subset $Y \supseteq X$ can't be frequent. These two properties are utilized by the Apriori algorithm in order to improve significantly the brute-force technique. For the reason that it utilizes a level-wise/breadth-first exploration of the itemset search space as it also prunes all of supersets related to any frequent candidate, which is true since there would be no superset of an infrequent itemset that could be frequent. It abstains from generating any candidate which has got an infrequent subset as well. In addition, the Apriori algorithm enhances the I/O complexity significantly as instead of counting the support for a single itemset, it utilizes a breath-first fashion to explore the prefix tree, and then calculates the support value for each valid candidate of size k which comprise level k in the prefix tree [21][22][23].

If we considered the dataset example shown in Table 4.1, and with specifying $minsup = 3$, Figure 4.2 demonstrates the itemset search space for the Apriori method arranged as a prefix tree, where connection between any two itemsets occur if and only if one of the itemset is a prefix and immediate subset of the other. Each node in the graph represents an itemset along with its support value. For example, the node $AC(2)$ can be understood as itemset with support of $\sup(AC) = 2$. Apriori enumerates the candidate patterns in a level-wise manner, as it is depicted in the Figure 4.3, which shows the power of pruning the search space via the two Apriori properties as well. For instance, once we determine that AC is infrequent, we could prune any itemset which has AC as a prefix, that is, the entire substrate under AC could be pruned. Similarly, for *CD*. The extension *BCD* from *BC* could be pruned too, due to it has got an infrequent subset, namely *CD* [22][23].

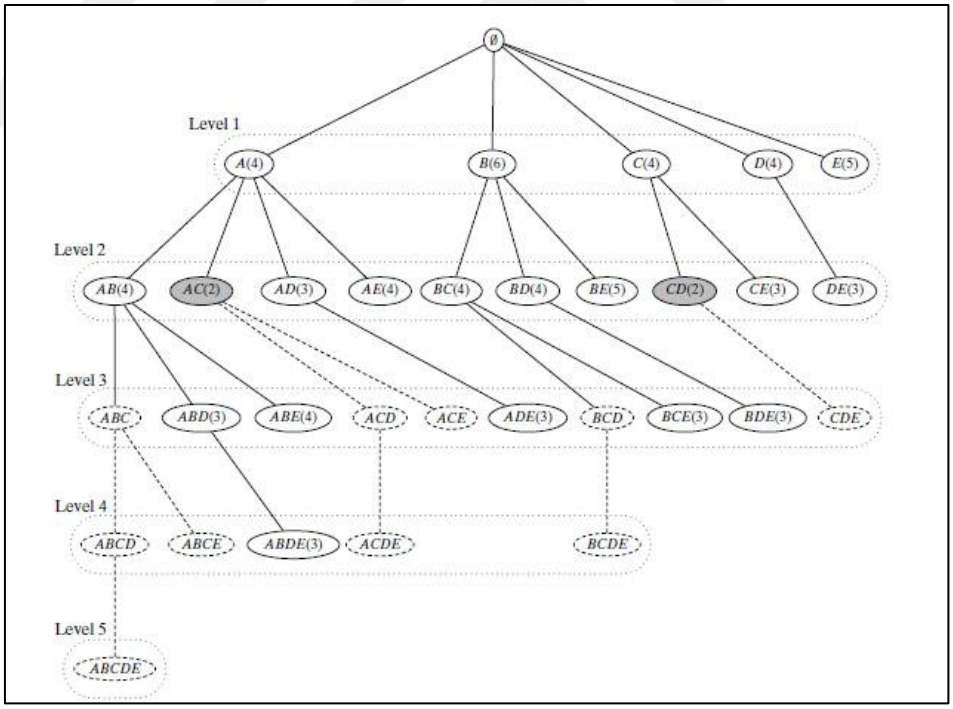


Figure 4.1: Apriori: prefix search tree and effect of pruning. Shaded nodes signify infrequent itemsets, on the other hand dashed nodes and lines signify all of the pruned nodes as well as branches. Solid lines signify frequent itemsets.

For candidate generation, the EXTEND PREFIX TREE procedure is use prefix-based extension. Given two frequent *k*-itemsetsX_a and X_b with a common k − 1 length prefix, which is, given twosibling leaf nodes with a common parent, we generate the (k + 1)-length candidate $X_{ab} = X_a \cup X_b$. This candidate is kept merely if it hasn't infrequent subset. Lastly, if a*k*-itemsetX_a hasn't got any extension, it is pruned from the prefix tree, and we recursivelyprune any of its ancestors with no *k*-itemset extension, in order that in$C^{(k)}$ all leaves are at level k. The whole process is repeated for the next level if new candidates were included. This process proceeds until new candidates are not added anymore [22][23].

D	A	B	C	D	E
1	1	1	0	1	1
2	0	1	1	0	1
3	1	1	0	1	1
4	1	1	1	0	1
5	1	1	1	1	1
6	0	1	1	1	0

(a) Binary Database

t	i(t)
1	ABDE
2	BCE
3	ABDE
4	ABCE
5	ABCDE
6	BCD

(b) Transaction database

x	A	B	C	D	E
t(x)	1 3 4 5	1 2 3 4 5 6	2 4 5 6	1 3 5 6	1 2 3 4 5

(c) Vertical Database

Figure 4.2: An example database

Figure 4.1 exemplifies the Apriori algorithm on the example dataset that is from Table 4.1 utilizing *minsup* = 3, see Table 4.2. All of the candidates $C^{(1)}$ are frequent (see Figure 4.2 a). All of the pairwise combinations will be taken into consideration during extension, because all of them share the empty prefix $\emptyset$ as their parent. These comprise the new prefix tree$C^{(2)}$ which is illustrated in Figure 4.2 b; since E hasn't got any prefix-based extensions, it's erased from the tree. Following support computation $AC(2)$ and $CD(2)$are removed (shown in gray) because they're infrequent. In Figure 11c the next level prefix tree is demonstrated. The candidate BCD is pruned as a result of the infrequent subset CD presence. All candidates that are at level 3 are frequent. Lastly,

$C^{(4)}$ (demonstrated in Figure11d) has got only one candidate $X_{ab} = ABDE$, that is generated from $X_a = ABD$ and $X_b = ABE$ since this's the only pair of siblings. After this step process of mining stops, due to no more extensions are possible [22][23]. See Figure 4.3.

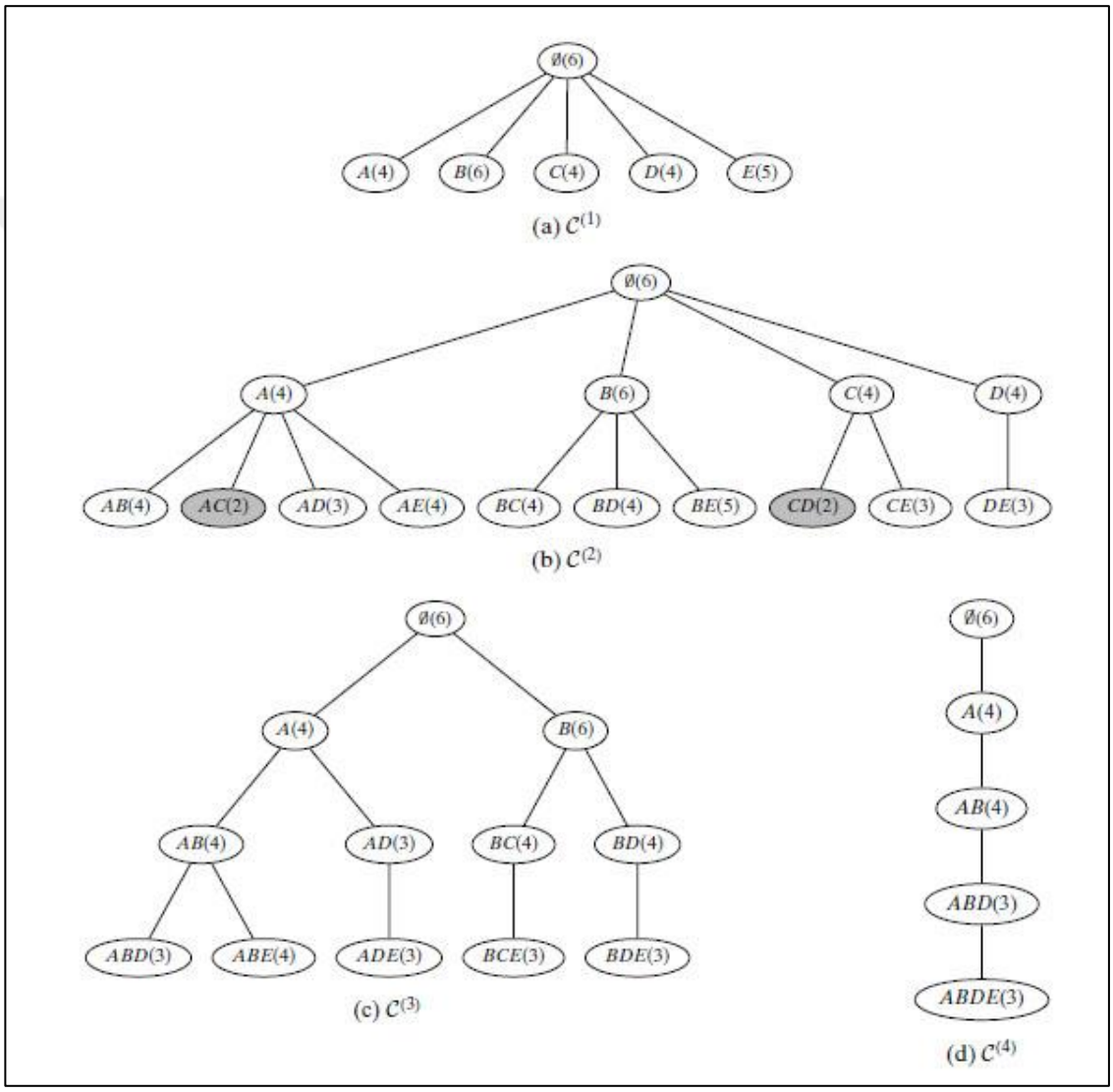


Figure 4.3: Itemset mining using Apriori algorithm.

Figure 4.4 shows the pseudo-code of the Apriori algorithm. Let $C^{(k)}$ symbolizes the prefix tree which encompassing all of the candidate *k*-itemsets. The algorithm starts through inserting the single items into an initially empty prefix tree so to populate$C^{(1)}$. Then it computes the support values for all current set of candidates (the while loop (lines 5-11)) which are in level *k* via the COMPUTESSUPPORT procedure, then for each transaction, it generates *k*-subsets that are in the database **D**, then it increases the support value of the corresponding candidate in $C^{(k)}$for each of such subset if it exists. As a consequence, the database is scanned once per level and the supports values of all candidates' k-itemsets are adjusted accordingly. Following, we eradicate any infrequent candidate (line 9). Prefix tree's leaves that survive comprise the set of frequent *k*-itemsets$\mathcal{F}^{(k)}$, that are utilized to generate the candidate (k + 1)-itemsets for the subsequent level (line 10) [21][22][23].

APRIORI Algorithm

APRIORI (D, $\mathcal{T}, minsup$):

1 $\mathcal{F} \leftarrow \emptyset$

2$C^{(1)} \leftarrow \{\emptyset\}$// *Initial prefix tree with single items*

3 $\boldsymbol{foreach}\ i \in \mathcal{T}\ \boldsymbol{do}$ *Add* i *as child of* $\emptyset$ *in* $C^{(1)}$ *with* $\sup(i) \leftarrow 0$

4 $k \leftarrow 1$// *k denotes the level*

5 $\boldsymbol{while} C^{(k)} \neq \boldsymbol{do}$

6 COMPUTESUPPORT ($C^{(k)}, \boldsymbol{D}$)

7 $\boldsymbol{foreach}$ *leaf* $X \in C^{(k)}\boldsymbol{do}$

8 $\boldsymbol{if}\ \sup(X) \geq minsup\ \boldsymbol{then}\ \mathcal{F} \leftarrow \mathcal{F} \cup \{(X, \sup(X))\}$

9$\boldsymbol{else}$ *remove* X *from* $C^{(k)}$

10 $C^{(k+1)} \leftarrow EXTENDED\ PREFIX\ TREE\ (C^{(k)})$

11 $k \leftarrow k + 1$

12 $\boldsymbol{return}\ \mathcal{F}^{(k)}$

$COMPUTESUPPORT\ \left(C^{(k)}, \boldsymbol{D}\right)$:

13 $\boldsymbol{foreach}\ \{t, \boldsymbol{i}(t)\} \in \boldsymbol{D}\ \boldsymbol{do}$

14 $\quad \boldsymbol{foreach}\ k - subset\ X \subseteq \boldsymbol{i}(t) \boldsymbol{do}$

15 $\quad\quad \boldsymbol{if}\ X \in C^{(k)} \boldsymbol{then}\ \sup(X) \leftarrow (X) + 1$

EXTENDEDPREFIXTREE $\left(C^{(k)}\right)$:

16 $\boldsymbol{foreach}\ leaf\ X_a \in C^{(k)} \boldsymbol{do}$

17 $\quad \boldsymbol{foreach}\ leaf\ X_b \in SIBLING\ (X_a), such\ that\ b > a\ \boldsymbol{do}$

18 $\quad\quad X_{ab} \leftarrow X_a \cup X_b$

// if there are any infrequent subsets prune candidate

19 $\quad\quad \boldsymbol{if}\ X_j \in C^{(k)},\ \boldsymbol{for\ all}\ X_j \subset X_{ab}, such\ that\ \left|X_j\right| = |X_{ab}| - 1\ \boldsymbol{then}$

20 $\quad\quad\quad$ Add X_{ab}as child of X_a with $\sup(X_{ab}) \leftarrow 0$

21 $\quad\quad \boldsymbol{if}\ no\ extension\ from\ X_a \boldsymbol{then}$

22 $\quad\quad$ remove X_a, and all ancestors of X_a with no extension, from $C^{(k)}$

23 **return** $C^{(k)}$

Figure 4.4 Apriori algorithm

The worst-case performance complexity performs by Apriori method is still $O(|\mathcal{T}|.|\boldsymbol{D}|.2^{|\mathcal{T}|})$ for the simple reason of all of the itemsets might be frequent. As a consequence, result of the pruning regarding the search space there's much lower cost compared to other algorithms. Yet, Apriori requires $O(|\mathcal{T}|)$ database scans in terms of I/O cost compared to $O(O^{|2\mathcal{T}|})$ scans performed by the brute-force method. In practice, it solely requires l database scans, where l is longest frequent itemset's length [22][23].

FP-Growth is an improved version of Apriori algorithm for discovering frequent itemsets and it was proposed by Han in [24]. FPGrowth proved to be very fast and yet to be memory efficient for the reason that it uses an internal structure in its computations known as FP-Tree.

The input to FP-Growth takes the input of transaction datasets (i.e. database), and uses a specified threshold (min support) to extract association rules. Transaction database can be understood as a set of transaction, each of which contains a set of items. It's important to understand that FP-Growth accepts or allows items to appear twice in the same transaction, and more importantly, it assumes that items are sorted in lexicographical order in all transactions. As an example, we can take the following transaction database, shown in Table 4.4, it consists of five transactions (from t1 to t5) and of five different itemsets (1,2,3,4, and 5).

Transaction id	Items
t1	{1, 3, 4}
t2	{2, 3, 5}
t3	{1, 2, 3, 5}
t4	{2, 5}
t5	{1, 2, 3, 5}

Table 4.4 Transaction database example

If we applied FP-Growth at the afore\"mentioned database, the generated results by FP-Growth are shown in Table 4.5 along with their support values (the presented results are generated with minsup of 40%). It discovers all the items that occur more frequently than others, and such frequency are decided upon the specified minsup threshold. Meaning, mining all itemsets which appear in at least minsup transactions in the transaction database.

Itemsets	Support
{1}	3
{2}	4
{3}	4
{5}	4
{1, 2}	2
{1, 3}	3
{1, 5}	2
{2, 3}	3
{2, 5}	4
{3, 5}	3
{1, 2, 3}	2
{1, 2, 5}	2
{1, 3, 5}	2
{2, 3, 5}	3
{1, 2, 3, 5}	2

Table 4.5 Association rules generated by FP-Growth

In the results, for example, the itemset {2, 3 5} has a support of 3 because it appears in transactions t2, t3 and t5 [22][23].

4.3.1.1 Frequent Itemset Generation

To enumerate the total possible itemsets list, a lattice structure might be utilized, such as the lattice structure of the itemset of $I = \{a, b, c, d\}$ shown in Figure 4.5. Usually, the number of frequent itemsets which can primarily be generated from a dataset of k items can be up to $2^k - 1$. Due to k may be huge in different applications. The itemsets search space which requires to be searched is exponentially large [22][23].

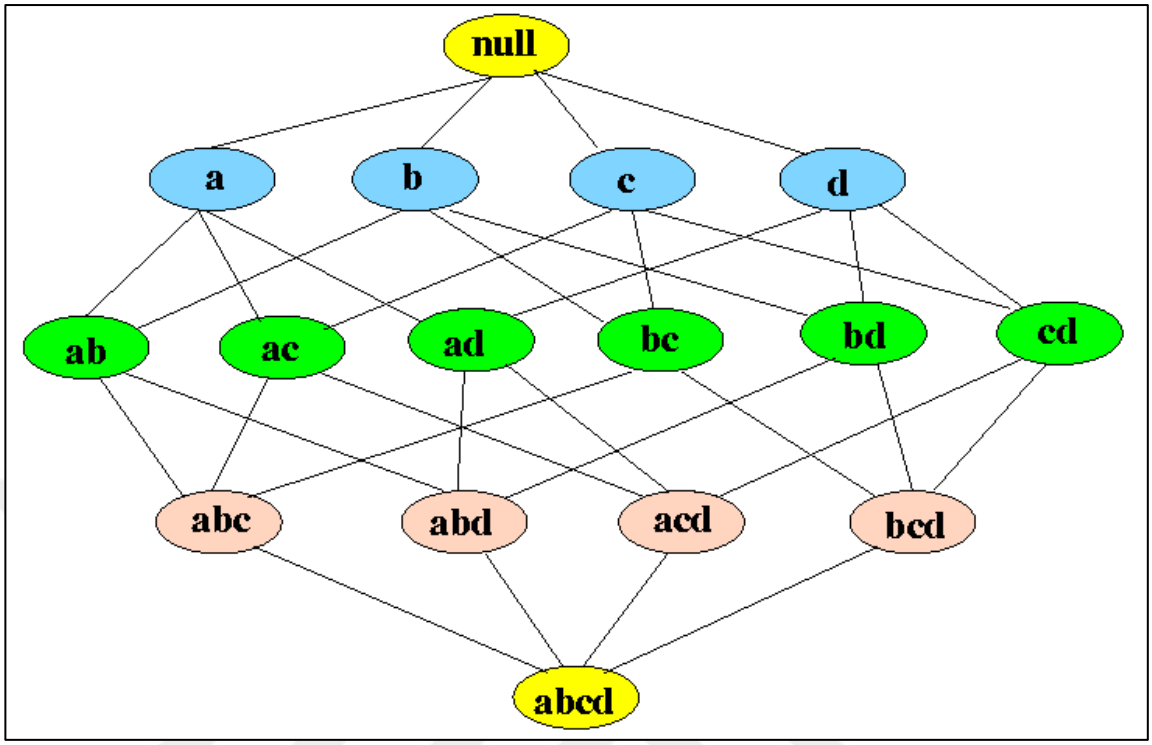


Figure 4.5 Itemset Lattice

There're many methods in order to decrease the computational complexity in regards to frequent itemset extraction, such as:

1- Decrease the candidate number itemsets (M). The Apriori standard has been clarified in the following section.
2- Decrease the number of comparisons. In lieu of comparing each of the candidate itemset against each transaction, we're able to decrease the number of comparisons through using data structures which are more advanced.

Definition of order feature: Let I bea set of items, and $J = 2^I$ be the power set of I. A measure f is monotone (or upward closed) if $\forall X, Y \in J : (X \subseteq Y) \Rightarrow f(X) \leq f(Y)$, that means which if X is a subset of Y, then $f(X)$ mustn't exceed $f(Y)$. On the other hand, f is anti-monotone (or downward closed) if $\forall X, Y \in J : (X \subseteq Y) \Rightarrow f(Y) \leq f(X)$, that means which if X is a subset of Y then $f(Y)$ must exceed $f(X)$ [22][23]. Any measure

that have got an anti-monotone feature could be contributed as directly into the mining algorithm in order to effectively clip the exponential search space regarding candidate itemsets, (as demonstrated in figure 14 generate frequent itemsets through utilizing Apriori method).

4.3.1.2 Association Rules Extraction

This section demonstrates how to extract association rules effectively out of a provided frequent itemset [22][23]. Each of the frequent k-itemset, Y, can produce up to $2^k - 2$ of association rules if we ignore all the rules with empty antecedents/consequent ($Y \rightarrow \emptyset$ *or* $\emptyset \rightarrow Y$). Also, if we partitioned the itemset *Y* into two non-empty subsets, the association rule that could be extracted is *X* and *Y*−*X*, such that $X \rightarrow Y - X$ satisfies the confidence threshold. However, the rules generated from such method must satisfies the support threshold for the reason of they are being generated from a frequent itemset. For instance, consider *X= {1, 2, 3}* to be a frequent itemset, thus, there're six association rules candidate which can be generated through *X: {1,2}→{3}, {1,3}→{2}, {2,3}→{1}, {1}→{2,3}, and {2}→{1,3}, {2}→{1,3}, {3}→{1,2}.* Since each of their support is identical to the support regarding X, the rules must satisfy the support threshold.

Calculating confidence value for any association rule doesn't entitle any further scans of the transaction dataset. Let's, for example, take the following rule {*1, 2} → {3}* which is extracted from the frequent itemset denoted by *X = {1, 2, 3}*, thus, the confidence can be calculated by *σ({1, 2, 3}) / σ ({1, 2}).* And due to *{1, 2, 3}* is frequent, the anti-monotone features of support ensure which *{1,2}* must be frequent, as well. Further, and because of the fact that support counts for both itemsets were calculated during the process of frequent itemset generation, it is not necessary to reread the whole of the dataset once again [22][23].

4.3.2 Close Algorithm (Charm)

Mining frequent closed itemsets requires performing closure checks, which is defined as whether $X = c(X)$. And for the reason that performing direct closure examination on dataset can be quite expensive, as it requires the verification of which X is the largest itemset that is common to all of the tids in $t(X)$, that is $X = \cap_{t \in t(X)} \ i(t)$. But, we'll – instead – define a vertical tidset intersection which is based on the method known as CHARM, for the purpose of applying more effectual closed itemsets closure check. Considering the following collection defines as IT-pairs $\{(X_i, t(X_i))\}$, three properties hold as follows [22][25][26]:

Property (1) if $t(X_i) = t(X_j)$,then $c(X_i) = c(X_j) = c(X_i \cup X_j)$, that implies which we could replace each occurrence of $X_i$with$X_i \cup X_j$ and prune the branch under X_j due to its closure is identical to the closure of $X_i \cup X_j$.

Property (2) if $t(X_i) \subset t(X_j)$, then $c(X_i) \neq c(X_j)$ but $c(X_i) = c(X_i \cup X_j)$, that means which we could replace each occurrence of X_i with $X_i \cup X_j$, and yet we can't prune X_j since it generates a different closure. Notice that if $t(X_i) \supset t\ (X_j)$ then we easily interchange the role of $X_i$and X_j.

Property (3) $t(X_i) \neq t\ (X_j)$, then $c(X_i) \neq c(X_j) \neq c(X_i \cup X_j)$. In this situation, we can't remove either $X_i$or X_j, since each of them generates a different closure.

From the pseudo-code for Charm algorithm. As an input, it takes all the frequent single items with their tidsets. Also, and at this stage, the set of all closed itemsets, denoted by C, is empty. For any given IT-pairs set defined as $P = \{(X_i, t(X_i))\}$, the method first starts by sorting the pairs in ascendant fashion according to their support values. Then, for each of the item X_i will be extended through all other items X_j in the sorted order. Only after, we will apply the three properties defined above so for branches to be pruned wherever it is possible. First of all, we ensure which $X_{ij} = X_i \cup X_j$ is frequent, with

checking the cardinality of $\boldsymbol{t}(X_{ij})$. If yes, following we check properties 1 and 2 (lines 8 and 12). Take into consideration that whenever we replace $X_i$with $X_{ij} = X_i \cup X_j$, in the current set P we make sure to do so. Also, the new set P_i, solely when property 3 holds do we add the new extension X_{ij} to the set P_i (line 14). If the set P_i isn't empty, then we could make a recursive call to Charm. Eventually, if X_i isn't a subset of any of the closed set Z by the same support, we could safety add it into the set of closed itemsets, C (line 18). For mining frequent itemsets the Charm algorithm is demonstrated in Figure 4.6 [22][25][26].

CHARM Algorithm

$initial\ Call: C \leftarrow \theta, P \leftarrow \{(i, \boldsymbol{t}(i)): i \in \tau, \sup(i) \geq minsup\}$

$\boldsymbol{CHARM}\ (P, minsup, C)$:

1 Sort P in increasing order of support (i.e., through increasing $|\boldsymbol{t}(X_i)|$)

2$\boldsymbol{foreach}(X_i, \boldsymbol{t}(X_i)) \in P, \boldsymbol{do}$

3 $P_i \leftarrow \emptyset$

4$\boldsymbol{foreach}(X_i, \boldsymbol{t}(X_i)) \in P, with\ j > i\ \boldsymbol{do}$

5$X_{ij} = X_i \cup X_j$

6 $\boldsymbol{t}(X_{ij}) = \boldsymbol{t}(X_i) \cap \boldsymbol{t}(X_j)$

7 $\boldsymbol{if}\ sub(X_{ij}) \geq minsup\ \boldsymbol{then}$

8 $\boldsymbol{if}\ \boldsymbol{t}(X_i) = \boldsymbol{t}(X_j)\boldsymbol{then}$ // $property1$

9$Replace\ X_i\ with\ X_{ij}\ in\ P\ and\ P_i$

10 $Remove\ (X_j, \boldsymbol{t}(X_j))\ from\ P$

11 $else$

12 $\boldsymbol{if}\ \boldsymbol{t}(X_j) \subset \boldsymbol{t}(X_j)\boldsymbol{then}$ //$Property\ 2$

13 $Replace\ X_i\ with\ X_{ij}\ in\ P\ and\ P_i$

14$\boldsymbol{else}$ //$Property\ 3$

15$P_i \leftarrow P_i \cup \{X_{ij}, \boldsymbol{t}(X_{ij})\}$

16 $\boldsymbol{if}\ P_i \neq \emptyset\boldsymbol{then}\ CHARM\ (P_i, minsup, C)$

17 $\boldsymbol{if}\ \nexists\ Z \in C, such\ that\ X_i \subseteq Z\ and\ \boldsymbol{t}(X_j) = \boldsymbol{t}(Z)\boldsymbol{then}$

18 $C = C \cup X_i$ // $Add\ X_i\ to\ closed\ set$

Figure 4.6 Charm algorithm

4.3.2.1 Generating Sets of Closed Repetitive Items

The frequent itemsets that generated from the set of items by using close method is called closed frequent itemsets.

The definition of closed frequent itemsets: the itemsets is a set of closed items, if it is closed and the column is closed and has column bigger or equal to the threshold of the specified column.

Where at the previous example at Figure 4.5, let's assume that the threshold of the column is 40% which means equal to 2, the set *{b,c}* is a closed frequent itemsets because its column is equal to 60% which means it is equal to 3. The closed frequent itemsets in the network is appointed by highlighted node. So, the closed frequent itemsets is generating as showed in Figure 4.7.

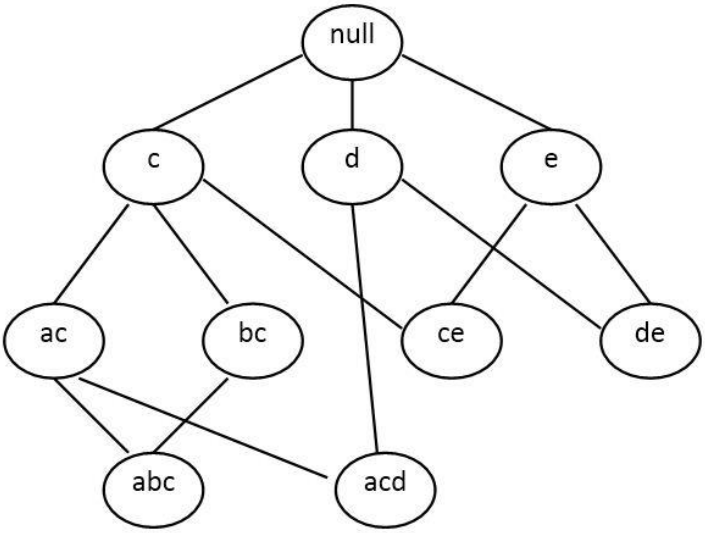


Figure 4.7 The network of closed frequent itemsets

4.3.2.2 Generating Association Rules According to Close Algorithm

Closed frequent items method can generate different types of association rules, such as the optimal association rules and approximation association rules). In more details, the generators set as follows [25][26]:

If *FCI* is a set of closed frequent itemset, and if *f* is a set of closed frequent items. Let's assume that FG_f is the set of frequent generators and FG_f is the set of frequent generators in the equivalence row f:

- **Find the optimal association rules:** the generic bases of finding the optimal association rules is:

$$GB = \{r : g \Rightarrow (f \backslash g) | f \in FCI \wedge g \in FG_f \wedge g \neq f\} \quad (6\text{-}1)$$

- **Find the approximate association rules:** the informative bases in order to find the approximate association rules is:

$$IB = \{r : g \Rightarrow (f \backslash g) | f \in FCI \wedge g \in FG \wedge g'' \subset f\} \quad (6\text{-}2)$$

g" is the close of g set

- **Transitive reduction:** the transitive reduction for IB is defined as follows:

$$RIB = \{r : g \Rightarrow (f \backslash g) \in IB | g'' \text{ is maximal proper subset of } f \text{ in } FCI\} \quad (6\text{-}3)$$

g" is the maximum set completely contained in *f* and be frequently close (FCI)

- **Minimal Non-Redundant rules (MNR):** the non-redundant minimal sets of can be defined as follows:

$$MNR = GB \cup IB \quad (6\text{-}4)$$

- **Transitive Reduction of Minimal Non-Redundant(RMNR): RMNR can found as follows:**

$$RMNR = GB \cup RIB \quad (6\text{-}5)$$

4.4 Conclusion

In this chapter, we have illustrated the concept of knowledge extraction from huge quantities of data by analyzing the participating between its features. Also, we explained the common methods in order to complete this process which is the symbolic methods that depends on analyzing the data of knowledge and study the more important features through the quantity of the repetition and then generate multiple types of association rules which defined by the symbolic methods. The Apriori method which through we can generate great sets of association rules in spite of that it is small database and this leads to generate many extra rules and because of this the close method come in use that depends on the close concept in order to generate association rules without losing any information.

CHAPTER FIVE

RULES EXTRACTION BASED ON NUMERIC MULTI-VIEWPOINT MODEL

5.1 Introduction

We have seen in the previous chapter some defects of the symbolic methods such as producing a huge number of the association rules that may cost a lot to extract the important rules. These defects can be solved through the use of numeric methods. As we stated in Chapter 4, the clustering methods maintain the important relationships between the data by putting similar data into one cluster. Therefore, we can take advantage from this property for extract important association rules. There is a precedent numeric method that extracts simple association rules, but this method produces a lot of rules and for some data types it can't extract any rule. So, we try to extract complex association rules which give more impact information and figure out the strong relationship between variables in one association rule.

5.2 Simple association rules extraction based on Inter-cluster communication

It has been (as we explained in the literatures review) proposed a method to extract simple association rules between an attribute from one subspace with another attribute from another subspace [2]. This algorithm (see Figure 5.1) is based on two criteria, which are Precision (Prec) and Recall (Rec):

PRECESION CRITERION **(of an attribute):** measures to what extent the attribute to be common between the data within the cluster.

RECALL CRITERION **(of an attribute):** to what extent single attributes are associated with single clusters.

Algorithm 1: Extracting Simple Association Rules

Input: Two-view points, first view point v_1, second view point v_2

Output: simple association rules

$\forall p_1 \in Pc^*, \forall p_2 \in Pc`^*$ and $c \in v_1, c` \in v_2$

1. **If** ($Rec(p_1) = Rec(p_2) = Prec(p_1) = Prec(p_2) = 1$) Then Test_Rule_Type;
2. El**seIf** ($Rec(p_1) = Rec(p_2) = Prec(p_2) = 1$) Then Test_Rule_Type;
3. **ElseIf** ($Rec(p_1) = Rec(p_2) = Prec(p_1) = 1$) Then Test_Rule_Type;
4. **ElseIf** ($Rec(p_1) = Rec(p_2) = 1$) Then Test_Rule_Type;

where Test_Rule_Type procedure is expressed as:

5. **If** ($Extent_{v1}(p_1) \subset Extent_{v2}(p_2)$) Then: $p1 \rightarrow p2$
6. **If** ($Extent_{v2}(p_2) \subset Extent_{v1}(p_1)$) Then: $p2 \rightarrow p1$
7. **If** ($Extent_{v1}(p_1) \equiv Extent_{v2}(p_2)$) Then: $p1 \leftrightarrow p2$

Figure 5.1 Extracting simple association rules

In this algorithm (figure 5.1), $Extent_{vj}(p_i)$ represents the set of data that contains p_i in the viewpoint v_j.

Indeed, this strategy produces a lot of simple association rules and huge numbers of relationships between two variables, so in our thesis we propose new algorithms for solving this problem, basing on the same criteria and basing on the closure computation, by producing complex association rules.

5.3 MCARIC: Complex Association Rules Based on Inter-cluster mechanism

In this chapter, we propose a new set of algorithms, called MCARIC approach, for extracting complex association rules from two viewpoints based on clustering models. The set of algorithms, that we have proposed, depend on two criteria which are recall and precision then it combines a closure computation (which is done by a symbolic method) and a cluster (which is generated by a clustering method) to make a hybrid method for

extracting useful knowledge. Each algorithm represents a specific category of association rules. The goal of defining these categories is to select the rules. These different categories don't depend on the confidence and the support to find the important rules.

In the first category (see figure 5.2), we can extract the most important rules when **recall=1** and **precision** = **1** for proper attributes (see equation 4.7). These rules contain only the attributes which are found in the same cluster and shared with all data in the cluster in both viewpoints.

Algorithm 1: Category 1

Input: Two-view dataset D, first view clustering C, second view clustering $C`$

Output: set of association rules

1: For each $c \in C$

2: Find P_c^*

3: For each $c` \in C`$

if $c \cap c` \neq \phi$

4: Find $P_{c`}^*$

% create A_{rp}^c the set of peculiar features from c such that their recall and

% precision are 1

5: $A_{rp}^c = \{a \mid a \in P_c^*, precision(a) = 1, recall(a) = 1\}$

% create $B_{rp}^{c`}$ the set of peculiar features from $c`$ such that their recall and

% precision are 1

6: $B_{rp}^{c`} = \{b \mid b \in P_{c`}^*, precision(b) = 1, recall(b) = 1\}$

% association rules extraction

7: if $|A_{rp}^c| \neq 0$ and $|B_{rp}^{c`}| \neq 0$

8: ExtensionTest $(A_{rp}^c, B_{rp}^{c`})$

End if

9: End if

10 : End for

11: End for

Function ExtensionTest (A, B)

if Extent(A) ⊂ Extent(B) then A → B
else if Extent(B) ⊂ Extent(A) then B → A
else if Extent(A) ≡ Extent(B) then A ↔ B
else if conf (A → B) > conf (B → A) then A → B
else if conf (A → B) < conf (B → A) then B → A
else A ↔ B

Figure 5.2 Algorithm 1: Category 1

In the second category, the **precision = 1**, so we are looking for the associations between common proper properties of all data in the clusters in both viewpoints and these attributes are presented in at least two clusters of each viewpoint.

Algorithm 2: Category 2

Input: Two-view dataset D, first view clustering *C*, second view clustering *C`*

Output: set of association rules

```
1:    For each c ∈C
2:        Find P_c^*
3:        For each c` ∈C`
              if c ∩ c` ≠ ϕ
4:                Find P_c`^*
                  % create A_p^c the set of peculiar features from c such that their precision are 1
5:                A_p^c = {a |a ∈P_c^* , precision(a) = 1, recall(a) ≠ 1}
                  % create B_p^c` the set of peculiar features from c` such that their precision are 1
6:                B_p^c` = {b |b ∈P_c`^* , precision(b) = 1, recall(b) ≠ 1}
                  % association rules extraction
7:                if |A_p^c| ≠ 0 and |B_p^c`| ≠ 0
8:                    ExtensionTest (A_p^c, B_p^c`)
                  End if
9:            End if
10:       End for
11 :  End for
```

Figure 5.3 Algorithm 2: Category 2

In the third category, we want to find associations between exclusive attributes in clusters in both viewpoints. These attributes are not present in more than one cluster, where **recall = 1.**

Algorithm 3 Category 3

Input: Two-view dataset D, first view clustering *C*, second view clustering *C`*

Output: set of association rules

For each $c \in C$

  Find P_c^*

  For each $c` \in C`$

    if $c \cap c` \neq \phi$

      Find $P_{c`}^*$

      % create A_p^c the set of peculiar features from *c* such that their recall are 1

      $A_r^c = \{a \mid a \in P_c^*, precision(a) \neq 1, recall(a) = 1\}$

      % create $B_r^{c`}$ the set of peculiar features from *c`* such that their recall are 1

      $B_p^{c`} = \{b \mid b \in P_{c`}^*, precision(b) \neq 1, recall(b) = 1\}$

      % association rules extraction

      if $|A_r^c| \neq 0$ and $|B_r^{c`}| \neq 0$

        ExtensionTest $(A_r^c, B_r^{c`})$

      End if

    End if

  End for

End for

Figure 5.4 Algorithm 3: Category 3

Finally, the fourth category is composed of three consecutive steps:

- Find a subset of the dataset which is shared in two clusters in both viewpoints.
- Find two subsets of attributes, from each viewpoint, that are shared between the previous subset of data.
- Finally, check whether the two subsets of attributes are closed itemsets or not. If they are closed, then find the associations between the two subsets of attributes.

This category injects a symbolic computation into a numeric one for extracting compact association rules (see figure 5.5).

Algorithm 4 Category 4

Input: Two-view dataset D, first view clustering *C*, second view clustering *C`*

Output: set of association rules

```
For each c ∈C
    For each c` ∈C`
        if c ∩ c` ≠ ϕ
            Find D_cc` the set of data d that are shared between c and c`, D_cc` = c ∩ c`
            D_cc` = { d |d ∈c, d ∈c` }
            % create A_s the set of attributes from the first view which are shared with
            % all data of D_cc`
            A_s = {a |a ∈d, ∀d ∈D_cc`, a ∈V_1}
            % create B_s the set of attributes from the second view which are shared
            % with all data of D_cc`
            B_s = {b |b ∈d, ∀d ∈D_cc`, b ∈V_2}
            % association rules extraction
            if |A_s| ≠ 0 and |B_s| ≠ 0
                if h (A_s)=A_s and h (B_s)=B_s
                    ExtensionTest (A_s, B_s)      % Closed itemsets
                Else
                    ExtensionTest (A_s, B_s)      % non-closed itemsets
            End if
        End if
    End for
End for
```

Figure 5.5 Algorithm 4: Category 4

5.4 Results and Discussion:

For the evaluation of our results, we will present the results from three different association rules mining algorithms on two datasets, Car and TicTacToe obtained from LUCS/KDD from [27], for which minimum confidence and support values must be specified. The minimum support value for all of our test is calculated from 1/(No. of rows) per dataset, with minimum confidences of 0.0, 0.2, 0.4, 0.6, 0.8, respectively. Apriori, Charm, and FP-Growth algorithms were chosen for this study, all of which are implemented in SPMF tool (an open source data mining library) [28].

5.4.1 Association Rules Mining using Symbolic Methods

From our preliminary tests, we have found that Apriori and FP-Growth algorithms gave the same number of association rules, confidence average, support average, and rules length average. However, FP-Growth algorithm outperformed Apriori algorithm only in running time on both datasets, see Table 5.1. Thus, we have decided to continue our comparison with FP-Growth and Closed Itemset algorithms only.

Dataset	Algorithm	Support	Confidence	Results
Car	FP-Growth	0.0008	0.0	**Time:** 3656ms **Average Rules Length:** 5.736 **Average Support:** 0.0000043 **Average Confidence:** 0.113 **Numbers of Rules:** 732252
	Apriori	0.0008	0.0	**Time:** 3982ms **Average Rules Length:** 5.736 **Average Support:** 0.0000043 **Average Confidence:** 0.113 **Numbers of Rules:** 732252
TicTac	FP-Growth	0.001	0.8	**Time**: 5782ms **Average Rules Length:** 8.043 **Average Support:** 0.0000011 **Average Confidence:** 0.996 **Numbers of Rules:** 1220682
	Apriori	0.001	0.8	**Time:** 136875ms **Average Rules Length:** 8.043 **Average Support:** 0.0000011 **Average Confidence:** 0.996 **Numbers of Rules:** 1220682

Table 5.1 FP-Growth and Apriori algorithm

As depicted in Table 5.2, we can see that FP-Growth algorithm always performs faster than Closed Itemset with relatively small difference in average rules length for both datasets. As the average confidence differences remained close, we can see a huge variant in the average support values from some tests in particular for the reason that the number of association rules is variant by each algorithm. Furthermore, Closed itemset generates relatively less number of association rules, which means Closed algorithm is more accurate, and also the contrary, the FP-Growth algorithm generates more redundant and not useful rules.

Algorithm	*min. conf*	Time	Average Support	Average Confidence	Average Rules Length	Rules No.
FP-Growth	> 0.0	3656ms	0.0000043	0.113	5.736	732252
	>= 0.2	563ms	0.000037	0.388	5.556	136193
	>= 0.4	204ms	0.000096	0.632	5.788	41506
	>= 0.6	125ms	0.00030	0.899	5.815	15316
	>= 0.8	94ms	0.00036	0.994	6.041	10905
Closed	>= 0.0	4844ms	0.0000091	0.129	5.906	435618
	>= 0.2	1406ms	0.000063	0.397	5.670	94247
	>= 0.4	640ms	0.00015	0.645	5.893	30071
	>= 0.6	375ms	0.00044	0.890	5.872	12363
	>= 0.8	328ms	0.00055	0.992	6.116	8484

(a)

Algorithm	*Min. conf*	Time	Average Support	Average Confidence	Average Rules Length	Rules No.
FP-Growth	> 0.0	142953ms	0.000000061	0.148	8.470	29902240
	>= 0.2	29609ms	0.00000031	0.468	7.637	6715054
	>= 0.4	10562ms	0.00000070	0.717	7.744	2928630
	>= 0.6	6610ms	0.0000012	0.939	7.791	1483778
	>= 0.8	5782ms	0.0000011	0.996	8.043	1220682
Closed	>= 0.0	229657ms	0.00000078	0.185	7.397	5145156
	>= 0.2	158812ms	0.0000059	0.453	7.546	1536384
	>= 0.4	107266ms	0.0000061	0.671	7.751	691990
	>= 0.6	67374ms	0.000010	0.863	7.948	347384
	>= 0.8	66187ms	0.000012	0.985	8.665	210826

(b)

Table 5.2 Applying FP-Growth and Closed itemset algorithms: (a) Car (b) TicTacToe

The second step in our evaluation is to filter the obtained association rules from the above tests. A simple application was written for this task where all rules from each test are compared and validated according to a condition and upon match, the rule is kept. For car dataset, the condition is: for all attributes numbers from the left side of the rule (i.e. left view) should be within the range of 1 to 15, and from 16 to 25 for the right side of that rule (i.e. right view). And since the association rules are unidirectional, we also considered the reverse order of that condition (1 to 15 and 16 to 25, but from right to left). As for the TicTacToe dataset, the left side of the rule should be within the range of 1 to 15, connected to the right side which should be from 16 to 39, and vice versa.

As we can see from Table 5.3 below, which represents the filtering results after applying the condition to each test from table 5.2, that there is a notable reduction in the number of rules. For example, using FP-Growth algorithm generated 732252 rule when minimum confidence was set to zero, but only 36322 remained after applying the filtering condition.

Algorithm	Time	Average Support	Average Confidence	Average Rules Length	Rules No.
FP-Growth	1849ms	0.00019	0.120	6.048	36322
	720ms	0.0020	0.344	5.525	7385
	295ms	0.0087	0.548	5.651	1522
	213ms	0.0660	0.815	5.213	403
	189ms	0.1393	0.984	5.350	191
Closed	1395ms	0.00046	0.140	6.029	20444
	586ms	0.0034	0.357	5.488	5090
	275ms	0.0144	0.578	5.632	1134
	211ms	0.0754	0.827	5.197	374
	164ms	0.1421	0.984	5.343	189

(a)

Algorithm	Time	Average Support	Average Confidence	Average Rules Length	Rules No.
FP-Growth	55469ms	0.0000074	0.1082	8.384	498488
	13079ms	0.00010	0.3760	7.078	72238
	6217ms	0.00043	0.6279	6.998	21351
	3295ms	0.00080	0.8453	7.145	8547
	3117ms	0.00073	0.9798	7.731	4721
Closed	10632ms	0.000081	0.1569	6.702	116490
	3827ms	0.00050	0.3649	6.370	29437
	1892ms	0.0019	0.5748	6.210	9041
	1270ms	0.00050	0.8453	6.371	3102
	846ms	0.0118	0.9187	6.826	950

(b)

Table 5.3 Results after applying filtering condition: (a) Car, (b) TicTacToe

5.4.2 Association Rules Mining using MCARIC approach

Using MultiSOM method to represent two viewpoints of the same dataset requires choosing the right number of clusters per view for each dataset. Therefore, we tested different numbers of clusters, namely 2, 3, 4, 9, 16, and 25 with 20000 iterations for each test. Thereafter, calculated the time, recall, and precision values per test. Table 5.4 below summarizes the results for car left view, car right, TicTacToe left, and TicTacToe right, respectively.

Number of Clusters	Number of Iterations	Clustering Time	Recall	Precision
2	20000	3m2s	0.8523	0.6429
3	20000	3m27s	0.6342	0.5817
4 (3 non-empty)	20000	3m28s	0.6342	0.5817
9 (6 non-empty)	20000	4m58s	0.4217	0.6526
16	20000	6m11s	0.3079	0.9286
25 (17 non-empty)	20000	8m58s	0.2023	0.6513

(a)

number of clusters	number of iterations	clustering time	Recall	Precision
2	20000	2m12s	0.8041	0.4447
3	20000	2m16s	0.7384	0.6753
4	20000	2m10s	0.6445	0.6936
9	20000	3m48s	0.401	0.7889
16	20000	5m1s	0.4049	0.9286
25 (17 non-empty)	20000	7m11s	0.3259	0.8386

(b)

Number of Clusters	Number of Iterations	Clustering Time	Recall	Precision
2	20000	2m11s	0.646	0.4379
3	20000	3m10s	0.6852	0.6819
4	20000	3m9s	0.4722	0.6396
9	20000	3m14s	0.3074	0.7551
16	20000	4m10s	0.2116	0.8487
25	20000	5m22s	0.1568	0.8594

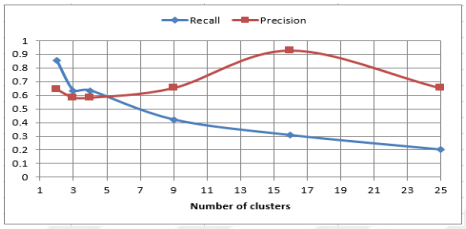
(c)

Number of Clusters	Number of Iterations	Clustering Time	Recall	Precision
2	20000	2m13s	0.655	0.4807
3	20000	2m20s	0.5767	0.6911
4	20000	1m54s	0.4918	0.7148
9	20000	2m20s	0.342	0.8856
16	20000	3m11s	0.2264	0.9502
25	20000	4m28s	0.1611	0.9207

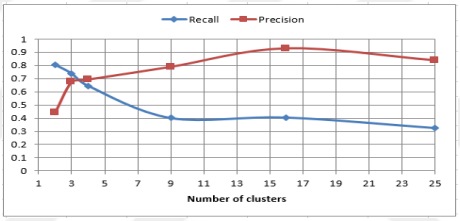
(d)

Table 5.4 Clustering per view test results: (a) Car left, (b) Car right, (c) TicTacToe Left (d) TicTacToe right

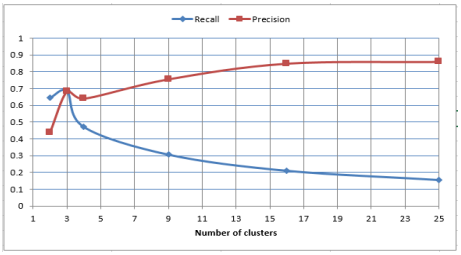
However, and as we discussed in the literature survey (see Section 3.6.1), the optimum number of clusters can be concluded from the graph-intersection point between recall and precision measurements. Therefore, and as shown in Figure 5.6, we chose three clusters for the left view and three clusters for the right view regarding car dataset, and three, three clusters respectively for TicTacToe left and right view.



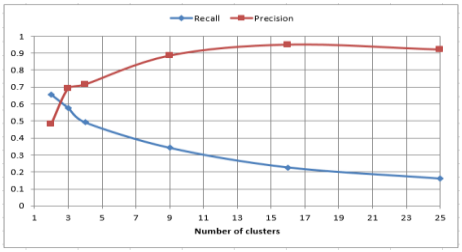
(a)



(b)



(c)



(d)

Figure 5.6 Recall and precision measurements per view: (a) Car left, (b) Car right, (c) TicTacToe Left (d) TicTacToe right

5.5 Discussion

A simple comparison of the results shown in the table below, clearly, the number of association rules extracted by MCARIC approach (Mining Complex Association Rules from Inter-cluster) method are notably less than to what we obtained from applying Closed and FP-Growth mining methods on the same datasets. The values for both Closed and FP-Growth algorithm depicted in the table below are the average values calculated based on Table 5.3. However, the time values represent the average of total time for each algorithm (i.e. time for association rules mining + filtering time).

We can see that the number of rules extracted using MCARIC approach is notably less than the rules extracted using Closed and FP-Growth algorithms on the same data sets, which proves that complex association rules extraction based on two viewpoints of the same dataset can produce useful and easy to understand rules. And more importantly, excludes redundant not useful rules. Applying our approach on car dataset generates nine rules compared to 5446, and 9164. Where for TicTacToe dataset, it generates nineteen rule only as oppose to 31804, and 121069.

Time in data mining algorithms generally and to a certain extent is not considered as a very important factor, since extracting less number of association rules (i.e. only useful rules) or summarizing huge amount of information into useful information is the sole purpose of these algorithms. With that being said, we can also notice that MCARIC approach outperformed the symbolic methods in running time. It was 1.3 seconds for Car dataset and 1.2 second in TicTacToe, compared to average time of 2s, 1.5s, ~129s, and 58s respectively.

Algorithm	Time	Support	Confidence	Rule Length	No. of Rules
Closed	2044.8ms	0.0472	0.577	5.53	5446
FP-Growth	1581.6ms	0.0432	0.562	5.55	9164
Simple Association Rules	-	-	-	-	0
MCARIC	1362.9ms	0.1659	0.4593	2.25	9

(a)

Algorithm	Time	Support	Confidence	Rule Length	No. of Rules
Closed	129552.6ms	0.003	0.572	6.49	31804
FP-Growth	57954.4ms	0.0004	0.587	7.46	121069
Simple Association Rules	-	-	-	-	0
MCARIC	1258.8ms	0.1474	0.4588	2.15	19

(b)

Table 5.5 MCARIC approach results: (a) Car, (b) TicTacToe

Generating association rules using symbolic methods requires *min. conf* value, which defines the strength of the strength of the association rule. Such that, all rules generated with confidence smaller the specified value will be excluded automatically. From our experiments, we have found choosing low value for *min. conf* has three drawbacks. First, its time-consuming process especially for large datasets, second, it generates huge amount of redundant and not useful rules which make using these rules in any decision-making process really difficult, and third, high resource consumption. For example, when we used Apriori algorithm on TicTacToe dataset with *min. conf* of > 0.0, the process of generating association rules stopped after it consumed all our available memory resources (8GB of memory). Such drawback makes symbolic methods completely not feasible to apply especially when working on huge datasets or with average size but containing higher number of attributes. But it can be solved through providing more temporary storage resources.

Clustering any dataset and generating viewpoints (i.e. subspaces) of the same data set solely guarantees that all data with similarities (i.e. highly related to each other) will be grouped together, leaving and eliminating all weak relationships among variables in different groups, also, considers all variables within each group to have the same

importance. And therefore, MCARIC approach doesn't rely on *min. conf* in it association rules generation which makes it more feasible to be implemented regardless of dataset size.

Another recognized contribution of MCARIC is that it works on binarized and non-binarized datasets as appose to conventional symbolic methods which only works on binarized datasets. Therefore, binarization of datasets were required before applying Closed and FP-Growth algorithms (discussed in Section 123). It is also important to mention that Simple Association Rules algorithm failed to extract any complex association rules due to the reasons we discussed in Section 5.2. Therefore, our proposed algorithm solved the problem related to the later mentioned algorithm. Also, it improves it by extracting association rules from any dataset types and by extracting complex association rules.

5.6 Application

We applied MCARIC approach in a study conducted on one hundred students from several faculties at the Türk Hava Kurumu Üniversitesi. The aim of the study was to find the relationship between the social life of students and its relation to their academic performance at these faculties. Table 5.6 shows the results of testing different clusters numbers to conclude the optimum number of clusters on social and academic datasets, whereas Figure 5.7 depicts the recall and precision measures for each both datasets. From the results shown in Table 5.7, MCARIC successfully extracted eleven important rules from the source dataset (social) to the target dataset (academic) and vice versa at only ~1.5 second. Table 5.8 lists the extracted association rules and their confidence values. To simplify, let us consider two random rules: (first rule in the table) if the student is married, studying type is dispatching, and wife is not working, then student's language performance is around medium. Also (fourth rule in the table) and regarding those students who find that the curriculum difficulty is medium are found to be males, married, no children, study type is dispatching, with enough salary, and last, no working wife.

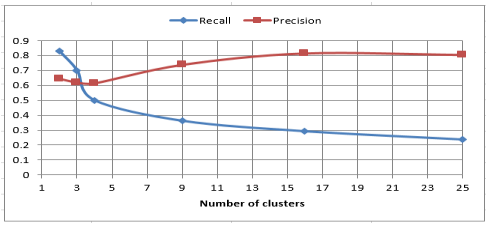
Number of Clusters	Number of Iterations	Clustering Time	Recall	Precision
2	20000	2m1s	0.8274	0.6441
3	20000	1m36s	0.7003	0.6193
4	20000	1m30s	0.4996	0.6136
9	20000	1m20s	0.3624	0.738
16	20000	1m42s	0.2925	0.8154
25	20000	1m52s	0.2372	0.8053

(a)

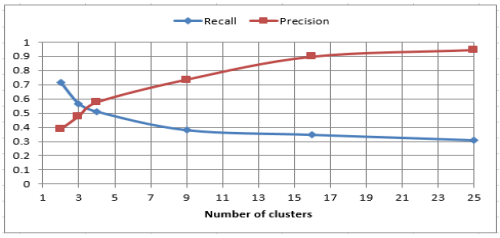
Number of Clusters	Number of Iterations	Clustering Time	Recall	Precision
2	20000	1m17s	0.7176	0.3847
3	20000	1m15s	0.5669	0.4768
4	20000	1m15s	0.5112	0.5747
9	20000	1m20s	0.38	0.7347
16	20000	1m29s	0.3466	0.897
25	20000	2m6s	0.3086	0.944

(b)

Table 5.6 Clustering test results: (a) Social, (b) Academic



(a)



(b)

Figure 5.7 Recall and precisions measurements: (a) Social, (b) Academic

Numbers of Rules	Time	Confidence average	Length Average	Support Average
11	1.5375s	0.6766	3.8182	0.2273

Table 5.7 Results from applying the MCARIC approach on social-academic study

Source	Target	Conf.	Rule
Social	Academic	0.5000	'married ', 'study_type_dispatching ', 'Accompany_person_not_work'➔'Language_difficulty_medium'
Social	Academic	0.8947	'married ' , 'child_0' , 'study_type_dispatching ' , 'Accompany_person_not_work'➔'Orientation_mixed'
Social	Academic	0.2222	'Gender_m', 'married ', 'child_0' ➔ 'Language_difficulty_medium', 'Curriculum_difficuly_medium', 'Avg_conservation_material_v_good'
Social	Academic	0.8889	'Gender_m', 'married ', 'child_0', 'study_type_dispatching ', 'salary_enough_yes', 'Accompany_person_not_work'➔'Curriculum_difficuly_medium'
Academic	Social	0.6538	'Curriculum_difficuly_little' ➔ 'single'
Academic	Social	0.7907	'Avg_conservation_material_v_good' ➔'single'
Social	Academic	0.8272	'single' ➔ 'Accomplish_duties_yes'
Social	Academic	0.4688	'single', 'sons_primary' ➔ 'Avg_conservation_material_medium'
Social	Academic	0.8077	'children_mt_2', 'sons_primary' ➔'Orientation_mixed'
Social	Academic	0.8171	'Gender_m' ➔'Accomplish_duties_yes'
Academic	Social	0.5714	'Time_study_yes', 'Accomplish_duties_yes', 'Language_difficulty_little' 'Curriculum_difficuly_medium' ➔ 'single' , 'children_mt_2'

Table 5.8 Association rules extracted from social-academic study

5.7 Conclusion

At this chapter, we explained the algorithms of how simple association rules and complex association rules are extracted based on two-viewpoints of the same dataset. Then, we presented a compete comparison of two symbolic methods used for association rules mining, FP-Growth and Closed itemset, applied on two datasets, Car and TicTacToe. From the obtained results, we saw FP-Growth performance to be faster than Closed itemset algorithm, however, Closed algorithm generates much less number of rules which is considered to be better in data mining field. Then after, our approach, MCARIC, was applied on both datasets and the results show that our approach outperformed both algorithms in both time and quality of association rules. A case study

presented at the end of this chapter where we applied MCARIC on a study at the Türk Hava Kurumu Üniversitesi, which aims at finding the relationships between the social and the academic performance of one hundred student. MCARIC approach successfully extracted eleven important association rules based on the aforementioned two views.

CHAPTER SIX

CONCLUSION

6.1 Conclusion

The notable rapid growth of the data electronically stored in databases, for example, or any information system is the derive force behind data mining. It required us and still to think of developing techniques/approaches so for such large amount of information to be processed into small and summarized useful information. The extracted knowledge will be used later in decision making process, or to increase our understanding about a research-related subject. Symbolic methods are the most commonly used methods in knowledge extraction from databases, but these methods come with drawbacks, mainly, association rules generation which is an expensive process and produces a huge number of rules for the simple reason of it consider data and its characteristics to be equal in terms of importance, thus it fails to isolate the important and strong relationships from the weak ones.

In this thesis, we proposed and implemented a novel approach, called MCARIC, for mining complex association rules from datasets. The idea basically relies on finding relationships between different subspaces (viewpoints) instead of using weights (which is the case of symbolic methods) to extract association rules. These subspaces describe the same set of data and represent it in multi-dimensional space, where items of similarities are grouped together. Such approach, which is achieved through the use of clustering techniques, helps in gathering items of strong relationships into groups which eliminates by default all the weaker ones. As a consequence, all weak and not useful association rules will be excluded.

6.2 Findings

When we implement the proposed mining approach and compared the obtained results with two symbolic methods, the following summarizes our findings:

1 - The number of association rules extracted by MCARIC approach are notably less than to what we obtained from applying Closed and FP-Growth mining methods on the same datasets. Which proves that complex association rules extraction based on two viewpoints of the same dataset can produce useful and easy to understand rules. And more importantly, excludes redundant not useful rules.

2 –MCARIC approach outperforms the two symbolic methods in running time.

3 – Extracting association rules using symbolic methods requires *min. conf* value, which sets the strength of the extracted association rule. From our experiments, we have found choosing low value for *min. conf* has three drawbacks, namely, its time-consuming process especially for large datasets, second, it generates huge amount of redundant and not useful rules, and third, high resource consumption. Such drawback makes symbolic methods completely not feasible to apply especially when working on huge datasets or with average size but containing higher number of attributes. MCARIC approach doesn't rely on *min. conf* in it association rules generation which makes it more feasible to be implemented regardless of dataset size.

4 – Finally, MCARIC approach works on binarized and non-binarized datasets as appose to conventional symbolic methods which only works on binarized datasets. Therefore, binarization of dataset is required.

REFERENCES

[1] Al Shehabi S., Lamirel J., "*Multi-Topographic Neural Network Communication and Generalization for Multi-Viewpoint Analysis*". International Joint Conference on Neural Networks, pp.1564–1569, 2005.

[2] Lamirel J., Al Shehabi S., "*Efficient Knowledge Extraction using Unsupervised Neural Network Models*". 5th Workshop on Self-Organizing Maps – WSOM, Paris/France, 2005.

[3] Lamirel J., Francois C., Al Shehabi S., Hoffmann M., "*New Classification Quality Estimators for Analysis of Documentary Information: Application to Patent Analysis and Web Mapping",* Scientometrics, Springer Verlag, Vol. 3, pp.445-462, 2004.

[4] Lamirel J., Al Shehabi S., Hoffmann M., François C., "*Intelligent Patent Analysis Through the Use of a Neural Network: Experiment of Multi-Viewpoint Analysis with The Multisom Model"*, Proceedings of the ACL workshop on Patent corpus processing, Vol. 20, pp. 7-23, 2003.

[5] Leeuwen M., Galbrun E., "*Association Discovery in Two-View Data*", IEEE Transactions on Knowledge and Data Engineering, Vol. 27, Issue: 12, pp. 3190 – 3202, 2015.

[6] Manning C., Raghavan P., Schütze H., "*An Introduction to Information Retrieval*", Cambridge University Press, pp. 109-135, 2009.

[7] SALTON G., "*Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer",* Addison Wesley, Amsterdam, North Holland, pp. 115–123, 1989.

[8] Aggarwal C. C., "*Data Mining: The Textbook*", IBM T. J. Watson Research Center, pp. 28-29, 2015.

[9] Tan P., Steinbach M., Kumar V., "*Introduction to Data Mining",* Pearson Addison Wesley, pp. 57-63, 2006.

[10] BERRY M. J. A., LINOFF G. S., "*Data Mining Techniques for Marketing, Sales, and Customer Relationship Management",* 2nd edition Wiley-IEEE Press, USA, New Jersey, pp. 8-12, 2004.

[11] Tan P., Steinbach M., Kumar V., "*Introduction to Data Mining",* Pearson Addison Wesley, pp. 44-65, 2006.

[12] Williams G. J., Huang Z., "*A Case Study in Knowledge Acquisition for Insurance Risk Assessment using a KDD Methodology*", The Pacific Rim Knowledge Acquisition Workshop, pp. 117-129, 1996.

[13] Hand D. J., Mannila H., Smyth P., "Principles of Data Mining", MIT Press, pp. 28, 2001.

[14] Tan P., Steinbach M., Kumar V., "*Introduction to Data Mining",* Pearson Addison Wesley, pp. 487-493, 2006.

[15] Bradley P.S., Fayyad U.M., *"Refining Initial Points for K-Means Clustering",* in Proceed of the 1st malison, Kaufmann, Morgan, pp. 91-99, 1998.

[16] Tan P., Steinbach M., Kumar V., "*Introduction to Data Mining",* Pearson Addison Wesley, pp. 569-647, 2006.

[17] Kohonen T., "*The Self-Organizing Maps*", Proceedings of the IEEE, Vol. 78, Issue: 9, 1464 - 1480 1990.

[18] François C., Polanco X., *"Information visualization and Analysis for Knowledge Discovery: using a Multi Self-Organizing Mapping",* 4th European Conference of Principles and practice of Knowledge Discovery in Databases (PKDD), Lyon, France, pp.12-16, 2000.

[19] LAMIREL J. C., TOUSSAINT Y., FRANCOIS C., POLANCO X., "*Using a MultiSOM Approach for Mapping of Science and Technology",* "In ISSI, Australia, vol. 1, pp. 339-351, 2001.

[20] Cichosz P., "*Data Mining Algorithms: Explained Using R*", John Wiley & Sons, Ltd, pp. 109-232, 2015.

[21] Agrawal R., Imielinski T., Swami A., "*Mining Association Rules between sets of items in large database*", Proceedings of the ACM SIGMOD international conference on Management of data, pp. 207-216, 1993.

[22] Tan P., Steinbach M., Kumar V., "*Introduction to Data Mining",* Pearson Addison Wesley, pp. 327-404, 2006.

[23] Aggarwal C. C., "*Data Mining: The Textbook*", IBM T. J. Watson Research Center, pp. 87-112, 2015.

[24] Han J., Pei J., Yin Y., Mao R., "*Mining Frequent Patterns without Candidate Generation: A Frequent-Pattern Tree Approach*", Data Mining and Knowledge Discovery, Vol. 8, Issue: 1, pp. 53-87, 2004.

[25] Pasquier N., Bastide Y., Taonil R., Lakhal L., "*Efficient mining of association rules using closed itemset lattices* ", INFORMATION SYSTEMS journal, Vol. 24, pp. 25-46, 1999.

[26] Zaki M. J., Hsiao C., "*CHARM: An Efficient Algorithm for Closed Itemset Mining*". SDM, pp. 457-473, 2002.

[27] The University of Liverpool, Selection of Discretized datasets, <http://cgi.csc.liv.ac.uk/~frans/KDD/Software/LUCS-KDD-DN/DataSets/dataSets.html>, Accessed at 5/5/2016.

[28] SPMF: An open source data mining library, <http://www.philippe-fournier-viger.com/spmf/>.

APPENDIX A

Questionnaire of the Impact of Social Status on the Academic Status

— **Gender**	☐ **Male** ☐ **Female**
— **Age**	☐ **Less than 30** ☐ **30-35** ☐ **36-40** ☐ **More than 40**
— **Qualification**	☐ **Lisans degree** ☐ **Bachelor degree** ☐ **Higher Diploma**
— **Marital status**	☐ **Married** ☐ **Single**
— **Do you have children?**	☐ **No** ☐ **One Child** ☐ **Two children** ☐ **Three or more children**
— **Type of study**	☐ **Dispatching** ☐ **On your account**
— **What are the study stages that sons belong?**	☐ **Without study** ☐ **Primary** ☐ **Preparotry** ☐ **Secondary** ☐ **A university**
— **If you are a dispatch from your government, does the salary enough for you and your family living?**	☐ **Yes** ☐ **No** ☐ **Almost**
— **Accompanying person**	☐ **Works** ☐ **Doesnot work**
—**Do you find a plenty of time in the house to study?**	☐ **Yes** ☐ **No**

— Do you accomplish your duties of study at home?	☐ Yes ☐ No
— Does your family consume some of your study time?	☐ Yes ☐ No
—Academic Orientation	☐ Conservation ☐ Undrstand ☐ Mixed / two together
— How difficult to understand because of the language	☐ Little ☐ Medium ☐ Much
—How difficult to understand because of the curriculum	☐ Little ☐ Medium ☐ Much
— The average in the descriptive material that rely on conservation	☐ Excellent ☐ Very good ☐ Medium ☐ Weak
— The average in the descriptive material that rely on understanding	☐ Excellent ☐ Very good ☐ Medium ☐ Weak

CURRICULUM VITAE

PERSONAL INFORMATION

Name, Surname: Iman A. F. Hasse

Date and Place of Birth: 29.09.1980 / Libya- Derna

Marital Status: Married.

Phone: 00905530713222

Email: Iman_mzm@yahoo.com.

EDUCATION

High School: Asma Secondary School.

Undergraduate: High Polytechnic Institute / Department of Computer / Programming and system analysis 2005. Derna.