Sudan University of Science and Technology

COLLEGE OF GRADUATE STUDIES

A PROPOSED FRAMEWORK FOR ASSOCIATION RULES MINING

مقترح إطار للتنقيب عن قواعد الارتباط

A thesis submitted in fulfillment of the requirements for the degree of
Doctor of Philosophy of Science

By
Wafaa Faisal Mukhtar

Supervised by
Prof. Eltayeb Salih Abuelyaman

September 2017

الآيـة

رب هب لي حكما وألحقني بالصالحين، واجعل لي لسان صدق في الآخرين، واجعلني من ورثة جنة النعيم.

الشعراء 83-85

Dedication

To my

Beloved ones

Who really cares

Acknowledgment

I owe my success and the work of this thesis to my supervisor, Prof. Eltayeb Salih Abuelyaman, who have been patience and kept putting faith in me, encouraging me whatever happens.

My gratitude also are undeniable to Prof. Izzeldin Mohammed Osman, whom I respect much and regard him as my Godfather.

Thanks extended to the honest teacher, Dr. Mohammed Elhafiz Mustafa Musa for his appreciatable leadership of the Ph.D. program.

I would like to show appreciation to all my colleagues at the College of Computer Science and Information Technology, starting from its Dean, Dr. Talaat Mohieldin Wahbi to every worker, for giving me space and time to complete my research. I would like to thank my students especially Mudathir, Mohammed, Mashair, Nuha and Mariam.

My appreciation goes to my family who helped me as no others could. My husband for acceptance my absence from home and taking care of my kids. My love goes to my kids who believed in their mom.

Abstract

Association rule mining as one of the descriptive tasks of data mining aims to extract interesting correlations, frequent patterns or associations among sets of items. Association rules mining became the primary tool for improving decisions in all aspects of life. This dissertation proposes an Intelligent Association Rule Mining Framework (IARMF) that guides inexperienced users through the process of selecting the best technique for their needs, which produces interesting rules.

Four mining tools namely Weka, Orange, Tangara and Knime were explored, regarding their association rule mining algorithms used and the employed interestingness measures. The researcher performed a new method to rank the suitability of measures to the type of the dataset. Experiments were carried out on three data sets from different domains. The experiments implemented the Apriori algorithm on Weka and Orange. Setting different values for the tool's parameters, the researcher got different results. The method then selects the measure that gives the most consistent rankings than the previous ranking.

Users with different knowledge and expertise tend to extract frequent patterns for their own uses. However users need to acquire the knowledge of using the available tools and mining algorithms. IARMF is a menu-driven, user-friendly framework that can be used by inexperienced users and researchers who wish to fine-tune parameters of their choice. The Sudanese Transplanted Kidney dataset was constructed from the records at the Sudanese Kidney Transplantation Society and Ahmed Gasim Hospital. Experimental evaluation of the proposed framework on the constructed dataset reveals performance that is significantly better than the traditional approaches. Preliminary results from a prototype for the proposed framework show quite useful outcomes and opened up a wide range of interesting future research opportunities.

المستخلص

التنقيب عن قواعد الارتباط هو احد المهام الوصفية لتنقيب البيانات والذي يهدف لاستخلاص العلاقات او الانماط، مما جعلها وسيلة هامة لدعم القرار في شتى اوجه الحياة. تقدم هذه الاطروحة نمذج ذكي للتنقيب عن قواعد الارتباط والذي يقوم بتوجيه المستخدمين، غير المتمرسين على برامج الحاسوب، من إختيار التقنية الانسيب لاحتياجاتهم.

تمت دراسة اربعة ادوات وهي Weka, Orange, Tangara and Knime، فيما يختص بالخوارزميات الخاصة بالتنقيب عن قواعد الارتباط وقياسات اهمية القواعد المستخدمة. اتبع الباحث منهجية جديدة لترتيب مدى موائمة القياسات لكل نوع من البيانات. تم اجراء التجارب على ثلاثة مجموعات من البيانات تخص مجالات متنوعة. تم تطبيق خوارزمية Apriori على كل من اداتي Wekaو Orange . باعتماد قيم مختلفة للمتغيرات تم الحصول على نتائج متغايرة. تقوم المنهجية المتبعة باختيار افضل القياسات التي تقدم ترتيب افضل من طريقة الترتيب السابقة.

المستخدمين على اختلاف المعرفة والخبرة يحاولون ايجاد انماط متكررة تناسب تطبيقاتهم، ولكن تنقصهم المعرفة بكيفية عمل الادوات المتوفرة في المجال وكيفية استخراج الخوارزميات لهذه الانماط. هذا النموذج الذكي يحتوي على قوائم سهلة التعامل حتى بالنسبة لمستخدمي الحاسوب من الباحثين وغيرهم والذين يرغبون في إعادة ضبط قيم المتغيرات لتوائم احتياجاتهم الخاصة. تم اختبار النموذج المقترح وتطبيقه على قاعدة البيانات الخاصة والتي جمعت من جمعية زارعي الكلى السودانية ومستشفى احمد قاسم التعليمي والذى وضح فعالية واداء افضل من الطرق التقليدية. النتائج الاولية للنموذج اوضحت مخرجات مفيدة وفتحت افاقا أوسع للابحاث في المستقبل.

Table of Contents

List of figures

List of tables

List of Symbols/Abbreviations

AR	Association rules
ARM	Association Rules Mining
ARMiner	Association Rules Miner
ARTool	Association Rules mining tool
CAR	Classified Association Rules
DIC	Dynamic Itemset
DM	Data Mining
FIM	Frequent Itemset Mining
KDD	Knowledge Discovery in Databases
minconf	minimum confidence
minsup	minimum support

CHAPTER ONE

Introduction

1.1 Preface

An increasing number of organizations are creating ultra large databases of business data, such as consumer data, transaction histories, sales records, etc. The growth in the amount of available information collected and generated far exceeds the growth of corresponding knowledge which creates both a need and an opportunity for extracting knowledge from databases.

The traditional database systems cannot functionally extract knowledge from these data. Statistical and machine learning methods long ago have been used to mine such knowledge, but they perform poorly when it comes to the huge amount of data. Statistics were used for survey analysis, generating reports and representing charts, comparative studies. Due to the fast evolution of computation and computers, machine learning became the new solution. A major subfield of artificial intelligence based on statistical foundation, machine learning is a multidisciplinary field involving information theory, philosophy, neurobiology and other fields.

Due to the variety of data, which holds images, forms, databases, and surveys, analysis techniques that deal mainly with large amount of data is adopted to extract knowledge from databases. Machine learning algorithms and data mining techniques aid and support the process of decision making.

Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Given a set of facts (data) F, a language L, and some measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset FS of F with a certainty C, such that S is simpler than the record of all facts in FS. A pattern that is interesting and certain enough is called knowledge(Frawley and Piatetsky-Shapiro 1992).

Coined in the mid-1990s, the term data mining has today become a synonym for 'Knowledge Discovery in Databases', Data mining is sorting through data to identify patterns and establish relationships. Figure 1.1 describes where data mining techniques fit in the KDD framework. Data mining problems are often solved by using different approaches drawn from computer science, including multidimensional databases, machine learning, soft computing and data visualization, and from statistics, including hypothesis testing, clustering, classification, and regression techniques.

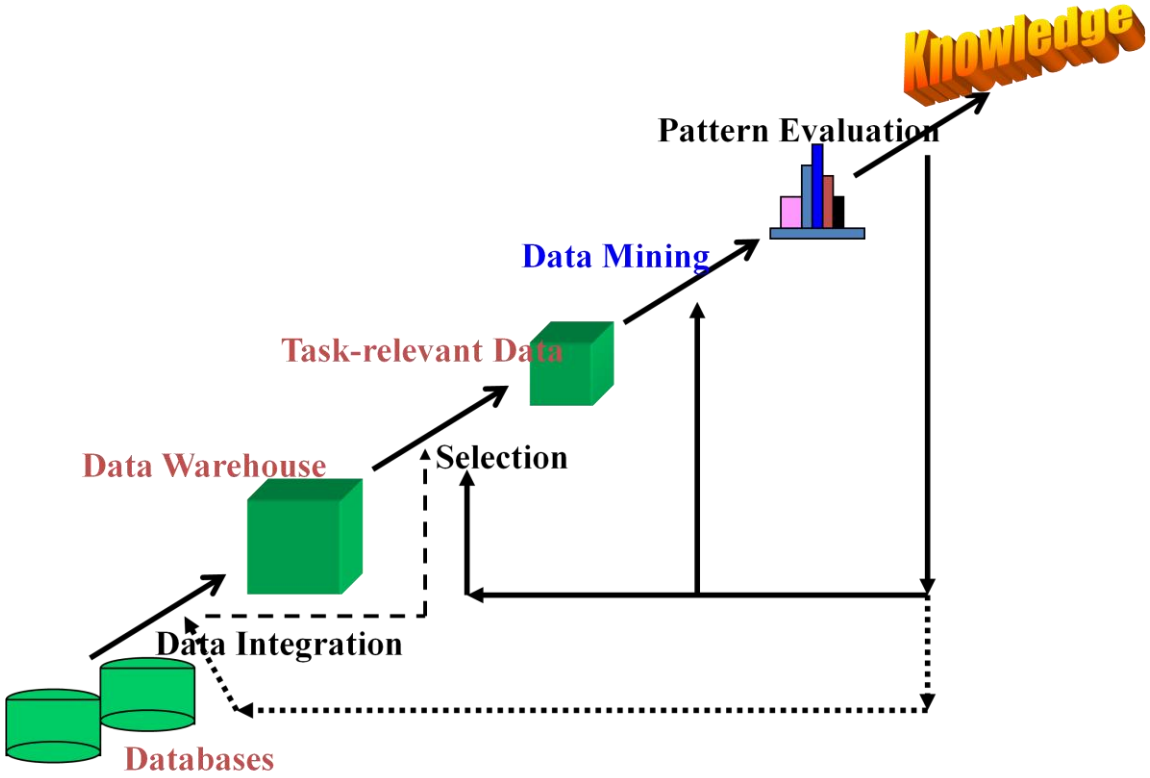


Figure 0-1: Data Mining and KDD process

There are two major tasks in data mining, descriptive and predictive as illustrated in figure 1.2. Descriptive mining is to summarize or characterize general properties of data in a data repository, while predictive mining is to perform inference on current data, to make predictions based on the historical data.

There are various types of data mining techniques such as classifications, clustering, association rule mining, regression, deviation detection and sequential pattern discovery.

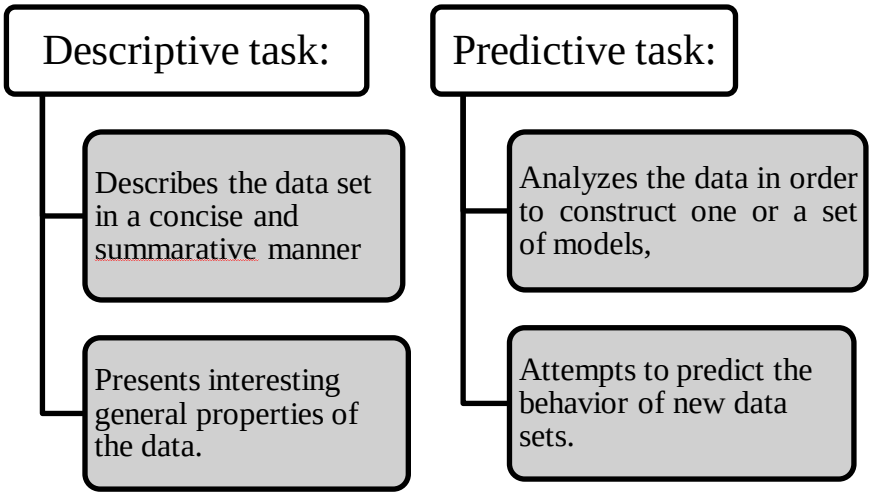


Figure 0-2: Data Mining tasks

Supervised learning algorithms automatically discover a knowledge model as a result of the learning process that provides a description of the data explored. Support Vector Machine (SVM), Decision Tree (DT), Artificial Neural Networks (ANNs), Regression, K-nearest neighborhood (KNN), Bayesian Networks (BN) and many more algorithms were considered for machine learning to classify and organize big data easily. The aim of the knowledge model is to predict the value of the target attribute for new unseen patterns. Another type of learning is unsupervised learning, which analyzes the information or data with unknown target variable, building the model which solely describes the data analyzed. The goal of the process is to create a model that describes interesting regularities in the data.

Association rule mining is one of the important and well-researched techniques of data mining, which identify relationships among a set of items in a database. As first introduced in 1993 for market basket analysis, to identify from a given database, consisting of itemsets (e.g. Shopping baskets), whether the occurrence of specific items, implies also the occurrence of other items with a relatively high probability. It aims to

extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transactional, relational databases or other information repositories.

Mining of association rules from a database consists of finding all rules that meet the user-specified threshold support and confidence, which can be decomposed into two subproblems:

1. Find all sets of items which occur with a frequency that is greater than or equal to the user-specified threshold support, s.
2. Generate the desired rules using the large itemsets, which have user-specified threshold confidence, α.

There are dozens of algorithms used to mine frequent itemsets, which use different strategies and data structures. Apriori algorithm(Agrawal, Imieliński, et al. 1993) is the most famous for its vast applications and can be considered as a state of the art algorithm.

1.2 Motivation

Association rule mining became the primary tool for improving decisions in all aspects of life. Arts, sciences, engineering, medical fields, use association rule mining in a way or another, knowingly or unknowingly. Users with different knowledge and expertise tend to extract frequent patterns for their own uses, but they need to tackle with the computer jargon regarding the terminology used in the different algorithms or tools.

1.3 Research problem

- A wide range of Apriori-based algorithms with different variations each suited for a specific application.
- The decision upon the interestingness measurements and their minimum threshold value.

Solution:

Design an intelligent framework that best fit the appropriate algorithm with the appropriate settings that produce interesting association rules.

1.4 The significance of the problem

. The framework, which recommends the best algorithm and parameter settings for different users of association rule mining technique, will help many researchers and scientist in many ways, e.g.

- Bioinformaticians to
  - Detect frequent patterns
  - Find similar sequences, structures, functions
- Physicians in finding
  - Similar symptoms
  - New relations between symptoms and disease
- Others
  - Extract information from the interpretation of the discovered association rules regarding understandability and interpretability of rules.

1.5 Objectives

1. Review the most used Association rules mining algorithms
   - Inspect the design issues, modifications and improvement made to the classic algorithm
2. Identify the Apriori-based association rules mining algorithms uses
   - Identify the appropriate algorithms used for each area of research
   - Track the literature for the parameter setting per application and algorithms

- Decide upon the interestingness measures to be used for each application
- Adjust the parameter setting for each application using available tools
- Compare the results with the best results achieved in the literature
- Describe what other parameters can influence the algorithm execution

3. Design a framework
    - Encode the algorithms for each group
    - Map the ARM's terms to each application
    - Design a user-friendly interface for the IARMF
4. Evaluate the framework performance

1.6 Methodology

The researcher will follow a qualitative analysis through the following steps;

- Literature survey
- Data analysis
    - Conduct experiments on selected (different) datasets
    - For each application adjust the confidence threshold with the minimum support;
        - Adjust the length of frequent itemset
    - Generate association rules based on the frequent item set
    - Specify another interesting measure for the application.
- System settings
- Simulate the application and test the performance of the proposed approach

1.7 Thesis organization

The theory behind Association rule mining techniques and the interestingness measures will be discussed in chapter two. Chapter three surveys the different types of

algorithms, especially Apriori-based algorithms and their application in other fields rather than market analysis. The outlined methodology of the research is explained in chapter four explaining five of the well known free data mining tools. In chapter five, experiments were conducted on three different datasets along with an explanation of the experimental results, followed by the framework design. A conclusion, recommendations and future work are stated in chapter 6.

CHAPTER TWO

Association Rules Mining

2.1. Introduction

Several organizations create very large databases, such as consumer data and transaction histories of sales records, patient records, images at hospitals, fingerprints and DNA samples at crime scenes, and data acquired via satellites or surveillance cameras. It is apparent that there is a considerable gap between extracting hidden knowledge. The amount of information in the world is terribly increasing, conversely, is the size and number of databases increase much faster. Computers used statistical techniques long ago to analyze this flood of raw data, but they were ineffective when it comes to an understanding. Knowledge Discovery in Databases(KDD) or Data Mining(DM) seems to be the best choice that visualizes, analyze, summarize and unearth comprehensible knowledge. Several tasks can be carried out in Data Mining and will be clarified in this chapter.

The theory behind the descriptive task of the Association rules mining technique is the state of this chapter along with a brief explanation of the common algorithms. The algorithmic aspects of the classic algorithm will be discussed in conjunction with the different adjustment made to it.

2.2. Data growth

The automation of business activities produces an ever-increasing stream of data because even simple transactions, such as a telephone call, the use of a credit card, or a medical test, are typically recorded in a computer. Scientific and governmental databases are also rapidly growing.

If it is understood at all, it will have to be analyzed by computers. Although simple statistical technique for data analysis were developed long ago, advanced

techniques for intelligent data analysis are not yet mature. As a result, there is a growing gap between data generation and data understanding. At the same time, there is a growing realization and expectation that data, intelligently analyzed and presented, will be a valuable resource to be used for a competitive advantage.(Frawley & Piatetsky-Shapiro 1992)

The computer science community is responding to both the scientific and practical challenges presented by the need to find the knowledge adrift in the flood of data. In assessing the potential of AI technologies, Michie (1990), a leading European expert on machine learning, predicted that "the next area that is going to explode is the use of machine learning tools as a component of large-scale data analysis." A recent National Science Foundation workshop on the future of database research ranked data mining among the most promising research topics. A combination of business and research interests has produced increasing demands for, as well as an increased activity to provide tools and techniques for discovery in databases.

As the gap between produced data and its understanding is widening, it is high time to look for potentially useful information which lies hidden in all these hypes of data or information. Data mining or Knowledge Discovery from Databases (KDD) helps to extract such patterns (Witten, Frank, & Hall, 2011). Data mining has been used earlier in market analysis, financial data analysis, business management, space exploration and proved to be the best solution for various domains especially medical data analysis (Tomar & Agarwal, 2013).

2.3 Knowledge Discovery in Databases (KDD)

As firstly defined at 1992 by the AI Magazine;

“Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data. Given a set of facts (data) F, a language L, and some measure of certainty C, we define a pattern as a statement S in L that describes relationships among a subset F_S of F with a certainty c, such that S is simpler (in some sense) than the enumeration of all facts in F_S. A pattern that is

interesting (according to a user-imposed interest measure) and certain enough (again according to the user's criteria) is called knowledge. The output of a program that monitors the set of facts in a database and produces patterns in this sense is discovered knowledge." (Frawley & Piatetsky-Shapiro 1992)

Data Mining is an essential step in the KDD process, where intelligent methods are applied in order extract data patterns evaluated afterwards as illustrated in fig 2.1. Data mining is a misnomer for knowledge mining from data, whereas many used the two terms as synonyms. Although, other terms carry a similar or slightly different meaning, such as data analysis, data dredging, knowledge extraction, data archeology or business intelligence.

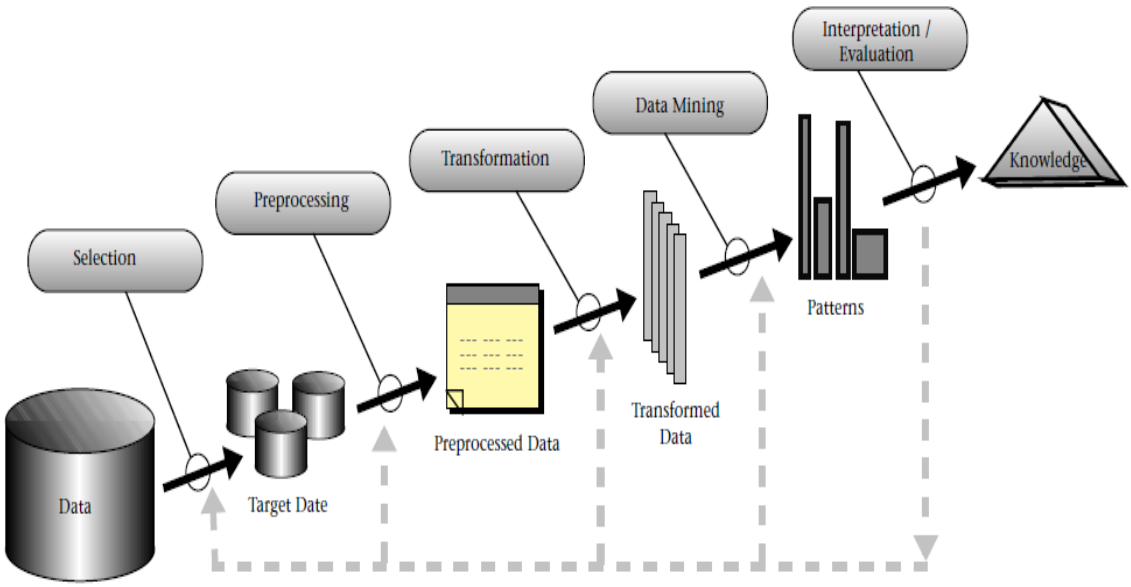


Figure 0-1: An Overview of the Steps That Compose KDD Process.(Fayyad, U. et al. 1996)

Database mining problems can be uniformly viewed as requiring discovery of rules embedded in massive data. Data mining could be applied to many kinds of repository as well to transient database. Description of the dataset in a brief and summarative manner and the interesting characteristics of the data is called descriptive

tasks of data mining. The predictive task analyzes the data in order to construct one or a set of models, to attempt to predict the behavior of new data sets. A summary of data mining tasks is shown in Figure 2.2.

Three classes of database problems were identified in (Agrawal, Imielinski, et al. 1993), as classifications, associations, and sequences. Classification problem involves finding rules that partition the given data into disjoint groups. Association is finding a relation between sets of items with some specific threshold. Sequences' fetching for patterns that happen in sequence over a specified time.

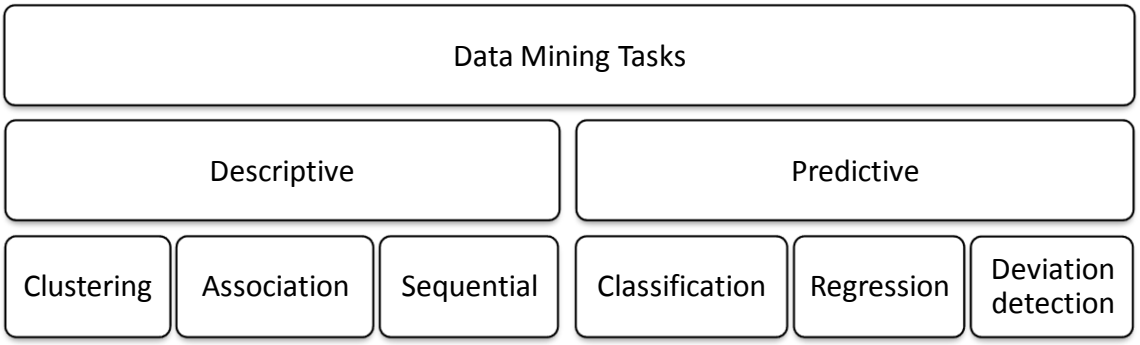


Figure 0-2: Data mining tasks

2.4 Mining association rules

Analyzing past transactions in retail business or what is known as supermarket basket analysis to derive associations between products purchased together was the start point for Association rule mining. Business decisions about what to put on sale, how to design coupons, how to place products on shelves in order to maximize the profit and improve the quality of such decisions. The qualitative sequential algorithms which are based on support and confidence will be discussed in this section.

2.4.1 Formal Model

A formal statement of the problem *(Agrawal 1993)*:

- Let I = 1_1, I_2,..., I_m be a set of binary attributes, called items.

- Let ***T*** be a database of transactions. Each transaction *t* is represented as a binary vector, with *t[k]* = 1 if *t* bought the item I_k, and *t[k]* = O otherwise. There is one tuple in the database for each transaction. Let ***X*** be a set of some items in I. We say that a transaction *t* satisfies *X* if for all items 1_k in *X*, *t[k]* =1.
- $X \Rightarrow Y$ has *support* s in the transaction set *D* if s% of transactions in *D* contain $X \cup Y$. $P(X) = \frac{Count(X)}{|T|}$, e.*g, X is frequent if P(X)* $\geq s$
- An association rule is an implication of the form $X \Rightarrow Y$, where $X \subset I$, $Y \subset I$, and $X \cap Y = \emptyset$.
- $X \Rightarrow Y$ holds with *confidence* c if c% of transactions in *D* that contain *X* also contain *Y*. $P(Y|X) = \frac{P(X,Y)}{P(X)}$.

Given the set of transactions *T*, one is interested in generating all rules that satisfy certain additional constraints of two different forms:

1. Syntactic Constraints: These constraints involve restrictions on items that can appear in a rule. Sometimes, interested only in rules that have a specific item I_x appearing in the consequent, or rules that have a specific item I_y appearing in the antecedent. Combinations of the above constraints are also possible, may request all rules that have items from some predefine itemset *X* appearing in the consequent, and items from some other itemset *Y* appearing in the antecedent.
2. Support Constraints: These constraints concern the number of transactions in *T* that support a rule. The support for a rule is defined to be the fraction of transactions in *T* that satisfy the union of items in the consequent and antecedent of the rule.

   Support should not be confused with confidence. Confidence is a measure of the rule's strength, whereas support corresponds to statistical significance.

2.4.2 Agrawal, Imieliński, and Swami (AIS) algorithm

Agrawal, Imieliński, and Swami, were the first to discuss qualitative association rules generation that can satisfy confidence and support constraints(Agrawal 1993). Their work was done in the context of the Quest project at IBM Almaden Research Center for retail data. They formulated the problem of mining association rules as a decomposition of two subproblems:

1. Generate all combinations of items that have fractional transaction support above a certain threshold, called minsupport(*s*). Call those combinations large itemsets, and all other combinations that do not meet the threshold small itemsets.
2. For a given large itemset $Y = I_I I_2 \ldots I_k$, $k \geq 2$, generate all rules (at the most k rules) that use items from the set $1_1,1_2,\ldots, I_k$. The antecedent of each of these rules will be a subset X of Y such that X has $k - 1$ items, and the consequent will be the item $Y - X$. To generate a rule $X \rightarrow I_j$ / c, where $X = I_I I_2 \ldots I_j\text{-}_1 l_{j+l} \ldots Ik$, take the support of Y and divide it by the support of X. If the ratio is greater than c then the rule is satisfied with the confidence factor c; otherwise it is not.

These two steps guarantee that all the itemsets generated are all large or interesting itemsets. All rules in the second step derived from Y must satisfy the support constraint because Y satisfies the support constraint and Y is the union of items in the consequent and antecedent of every such rule.

The second problem was set the least important as generating the association rules would be straightforward after deciding the large itemsets, which will consume all efforts. The algorithm in (fig 2.2) makes multiple passes over the database, and in each pass it scan the entire database for itemsets that satisfy the support constraint, known as candidate itemsets, to finally find the large itemsets.

Initially, the frontier set consists of only one element, which is an empty set. At the end of a pass, the support for a candidate itemset is compared with minsupport to determine if it is a large itemset. At the same time, it is determined if this itemset

should be added to the frontier set for the next pass, The algorithm terminates when the frontier set becomes empty. The support count for the itemset is preserved when an itemset is added to the large/frontier set.

Although the algorithm did not specify any data structure to store the database in it, it also generates a huge number of candidate itemsets that turns out to be small at the end. The algorithms did not discuss plainly the process of rule generation other than its limitation of discovering only one item in the consequent of the rule. Many algorithms have been developed to mine the association rules since then.

Algorithm 2.1: AIS template algorithm

```
Procedure LargeItemsets
begin
    let Large set L = 0;
    let Frontier set F = {0};
    (passes over the database)
    while F ≠ 0 do begin
        let Candidate set C = 0;
        forall database tuples t do
            forall all itemsets  in F do
                if t contains f then begin
                    let Cf = candidate itemsets that are extensions of f
                and contained in t;
                    forall itemsets cf in Cf do
                        if cf ∈ c then
                            Cf.count = cf.count + 1;
                              else begin
                            cf count = 0;
                            C=c+cf;
                                end
                    end
(join procedure)
                        let F = 0;
                        forall itemsets c in C do begin
                    if count(c)/ dbsize > rninsupport then
                        L= L+c;
                    if c should be used as a frontier in the next pass then
                        F= F+c;
                    end
                end
```

2.4.3 SETM algorithm

The SET-Oriented mining algorithm proposed in (Houtsma and Swami 1995) was drawn up to mine association rules for retail business stored in a relational databases system, motivated by the desire to use SQL to calculate large itemsets. It first saved a copy of the candidate itemsets sorted on itemsets then it generated candidate itemsets by using relational merge-join operation, but after pruning of small itemsets, another sorting is needed. Its disadvantage is due to the number of candidate sets that could not fit in memory with no buffer management strategy.

In both the AIS and SETM algorithms, candidate itemsets are generated on the fly during the pass as data is being read. Specifically, after reading a transaction, it is determined which of the itemsets found large in the previous pass are present in the transaction. New candidate itemsets are generated by extending these large itemsets with other items in the transaction. However, the disadvantage is that this results in unnecessarily generating and counting too many candidate itemsets that turn out to be small.

2.4.4 Apriori algorithm

The Apriori algorithm developed by (Agrawal R 1994) was a great achievement in the history of mining association rules. It is by far the most well-known association rule mining algorithm.

The Apriori algorithms generate the candidate itemsets to be counted in a pass by using only the itemsets found large in the previous pass, without considering the transactions in the database as described in fig 2.3. The basic intuition is that any subset of a large itemset must be large. Therefore, the candidate itemsets having *k* items can be generated by joining large itemsets having *k -1* items, and deleting those that contain any subset that is not large. The procedure resulted in generation of a much smaller number of candidate itemsets.

This technique does not use the database for counting the support of candidate sets but uses an encoding of (Tid, item) and uses the property that any subset of a large itemset must be a large itemset. Fig 2.4 describe the pruning step that checks all the subsets and remove candidate which are not large (small itemsets).

Apriori always outperforms its successor algorithms, AIS and SETM, by reducing the number of large itemsets and rules that have more than one item as a consequent, but it shares the same problem of scanning the entire database whenever it needs to determine the support value.

Algorithm 2.2: Apriori algorithm

$L_1 = \{large\ 1\text{-}itemsets\}$
For $(k = 2;\ L_{k-1} \neq \phi;\ k++)$ do begin
    C_k = apriori- gen (L_{k-1});
    forall transactions $t \in D$ do begin
        C_t = subset (C_k, t)
        forall candidates $c \in C_t$ do
            $c.count++$;
        end
    end
    $L_k = \{ c \in C_k | c.count \geq minsup \}$
end
$Answer = \bigcup_k L_k$;

Algorithm 2.3: Prune algorithm

for all $itemsets\ c \in C_k$ do
    for all $(k\text{-}1)\text{-}subsets\ s\ of\ c$ do
        if $(s \notin L_{k-1})$ then
            delete $c\ from\ C_k$

AprioriTid (Agrawal R 1994) solved this problem by reading the whole database and store all itemset alongside with their transaction in an array, which will be scanned instead of the entire database. The large number of candidate sets which can fit in memory is also a shortcoming of AprioriTid. Thus, led to the design of a hybrid algorithm called AprioriHybrid(Agrawal R 1994) that uses Apriori in the initial passes and switches to AprioriTid when it expects that the candidate set can fit in memory. However, the implementation of AprioriHybrid is more complex than Apriori. Hence, the worse performance may be an acceptable tradeoff in situations such as the cost of switching without realizing the benefits, especially if it occurs in the last pass.

Algorithm 2.4: AprioriTid algorithm

$L_1 = \{large\ 1\text{-}itemsets\}$
$C_1^{\wedge} = database\ D;$
For $(k = 2;\ L_{k-1} \neq \phi\ ;\ k++)$ do begin
    C_k = apriori- gen (L_{k-1});
    $C_k^{\wedge} = \phi\ ;$
    forall *entries* $t \in C_{k-1}^{\wedge}$ do begin
        $C_t = \{c \in C_k | (c - c[k] \in t.set - of - items$
            $\wedge (c - c[k-1]) \in t.set - of - items\};$
        forall candidates $c \in C_t$ do
            $c.count++;$
            if $(C_t \neq \varphi)\ then\ C_k^{\wedge} += < t.TID, C_t >\ ;$
        end
    end
    $L_k = \{\ c \in C_k | c.count \geq minsup\}$
end
$Answer = \bigcup_k L_k;$

The larger the candidate set, the higher the processing cost for discovering large itemsets and the generation of the 2-large itemsets is the key to improving the performance of the algorithm, so a raise to the research of increasing the efficiency and

performance by modifying the Apriori algorithm to eliminate rules by specifying a minimum efficiency improvement.

2.4.4.1 Apriori variations

Other algorithms adapted Apriori as a basic strategy, and made modifications in order to produce faster and more sophisticated algorithms. A hash-based technique (Park and Chen 1995) was designed to reduce the number of itemsets to be explored in the candidate set Ck in initial iteration by collecting information about Ck+1 in advance which reduced the corresponding processing cost to determine large itemsets. Furthermore all itemsets after pruning are hashed to a hash table which reduced effectively the database size. DHP outperforms Apriori in execution time after the second pass with various minimum support but the execution time increases as DB size increase

Recall that the reason the database needs to be scanned multiple numbers of times is because the number of possible itemsets to be tested for support is exponentially large if it must be done in a single scan of the database. The Partition algorithm has been designed especially for very large databases (Savasere, Omiecinski, & Navathe 1995). It assumes that the DB resides in secondary storage dividing it into a number of non-overlapping partitions designed to fit in main memory. The algorithm scans the DB twice to generate a set of all potentially large itemsets with local support as a fraction of transactions containing that itemset in a partition, set counters for these itemsets and their actual (global) support is measured thus reducing the disk I/O. Nevertheless local large itemset may or may not be large in the context of the entire database and moreover, determining the number of partitions given the available memory. A number of algorithms are available for association rules mining centralised databases but a few for distributed data mining, but now after a decade, databases are no longer centralized. These algorithms which are available for the centralised database mining can't be used directly for Distributed Data Mining (DDM). DDM requires local processing at all the sites to find the local frequent itemsets, communication between the sites and finally finding frequent global itemsets. This requires huge storage space,

communication, synchronization and processing capabilities and trade-off between them.

(Toivonen 1996) put forward a new algorithm for sampling large databases for association rules mining via one full pass over DB and two at the worst case which reduced significantly disk I/O. It picks a random small sample enough to be handled totally in main memory, apply Apriori for the sample with lowered minimum support. As a tradeoff between accuracy against efficiency, a risk of losing valid association because their frequency in the sample is below the threshold, raising the questions of discovering exact association rules in one pass and the gain by sampling.

Addressing performance, a new algorithm was developed (Brin et al. 1997), faster than Apriori by reducing the number of passes made over the data while keeping the number of itemsets which are counted in any pass relatively low and better than sampling. It addressed functionality by using other interestingness measures from a semantic perspective by developing conviction, an alternative to confidence and interest, introducing implication rules as an alternative to association rules. The intuition behind DIC(Brin et al. 1997) is that it works like a train running over the data with stops at intervals M transactions apart, with added flexibility of having the ability to add and delete counted itemset on the fly. It approximately takes half the time spent by Apriori for the same k-itemsets and varying the size of the interval will achieve higher speed up. When compared with Apriori, in high support Apriori outperforms while DIC outperform 30% faster in low support. DIC is sensitive to how homogeneous is the data, i.e. if the data is very correlated, the itemset will not be realized until counting it in most of the DB.

A comparison between four algorithms namely, Apriori, DIC, Partition, and Eclat, was conducted in (Hipp 2000). They used two generated datasets and two real-world applications, basket market and car equipment. Their main question was concerning; the runtime of the algorithms, the strategies used to traverse the search space and to determine the support. Two strategies were used so far, counting occurrences and intersecting Tid-list combined with breadth-first and depth-first search

techniques. The comparison brought up similar results in run time and the strategies were balanced for the basket data. Concluding using DFS combined with counting occurrences.

A hybrid approach was developed afterward in(Hipp, Güntzer et al. 2000), by counting occurrences in BFS manner whenever determining the support values of small candidates and switch to tid-set intersections for the remaining candidates using rightmost DFS. The new algorithm performs best in nearly all cases, but it suffered when setting a small average size of frequent itemsets.

2.4.4.2 Improving Apriori

Rivaling the above mentioned Apriori and its variants, it has been found that Apriori algorithm needs several database scans. Apriori needs n+1 scans, where n is the length of the longest pattern. Almost after a decade, (Han et al. 2000) proposed a new data structure, frequent pattern tree (FP-tree). It reduces the number of scans of the entire database using only two scans of database when mining all frequent itemsets for effectual mining. As the itemset in any transaction is always encoded in the corresponding path of the FP-trees consequently this method assured that it under no circumstances generates any combinations of new candidate sets which are absent in the database.

Since the FP-growth method is faster than the Apriori, it is found that few lately frequent pattern mining methods being effectual and scalable for mining long and short frequent patterns.

2.4.4.3 Rare items

In the classical Apriori algorithm, there is only one threshold value, which implies that all items in data sets have the same property which is far from the condition in real life. In the retailing business, customers buy some items very frequently, but other items very rarely.

Algorithm 2.5: (FP-tree construction)

1. *Input: A transaction database DB and a minimum support threshold _.*
2. *Output: Its frequent pattern tree, FP-tree*
3. *Method: The FP-tree is constructed in the following steps.*
4. *Scan the transaction database DB once.*
5. *Collect the set of frequent items F and their supports.*
   3. *Sort F in support descending order as L, the list of frequent items.*
   4. *Create the root of an FP-tree, T, and label it as \null".*
   5. *For each transaction Trans in DB do the following.*

   5.1 Select and sort the frequent items in Trans according to the order of L.

   5.2 Let the sorted frequent item list in Trans be [pjP], where p is the first element and P is the remaining list.

   5.3 Call insert tree([pjP]; T).

   The function insert tree([pjP]; T) is performed as follows.

1 *If T has a child N such that N.item-name = p.item-name, then*

   2 *increment N's count by 1;*

6. *else create a new node N, and let its count be 1,its parent link be linked to T, and its node-link be linked to the nodes with the same item-name via the node-link structure.*
7. *If P is nonempty,*
8. *Call insert tree(P;N) recursively.*

Usually, the necessities, consumables and low-price products are bought frequently, while the luxury goods, electric appliance, and high-price products infrequently. The same difficulty may occur when we are about to mine medical records. Mining medical records is a very important issue in the real-life application and it can reveal which symptoms are related to which disease. However, many important symptoms and diseases are infrequent in medical records. For example, flu occurs much more frequent than severe acute respiratory syndrome (SARS), and both have symptoms of a fever and persistent cough. In such situations, if the minsup is set too high, all the

discovered patterns are concerned with those low-price products or critical symptoms, which only contribute a small portion of the profit to the business or defining a disease. On the other hand, if the minsup is set too low, many meaningless frequent patterns will be generated too, which will overload the decision makers, who may find it difficult to understand the patterns generated by data mining algorithms.

The dilemma faced by the two applications above is called the rare item problem. In view of this, researchers either split the data into a few blocks according to the frequencies of the items and then mine association rules in each block with a different minsup or group a number of related rare items together into an abstract item so that this abstract item is more frequent. But they proved the inconvenience of these solutions. (Liu et al. 1999) have extended the existing association rule model to allow the user to specify multiple minimum supports to reflect different natures and frequencies of items. Specifically, the user can specify a different minimum item support for each item. Thus, different rules may need to satisfy different minimum supports depending on what items are in the rules. The new model enabled the users to produce rare item rules without causing frequent items to generate too many meaningless rules. However, the proposed algorithm in Liu et al. named the MSapriori algorithm, adopts an Apriori-like candidate set generation-and-test approach and it is always costly and time-consuming, especially when there exist long patterns(Liu et al. 1999).

The algorithm with multiple-support is proposed for solving “rare item” problem, which means the common items and rare items cannot be satisfied within a single support. The problem with the existing algorithm is that it considers every category with same preference, which is not what want in a supermarket.

A semantic measure named utility based measure was defined in (Geng & Hamilton 2006), takes into consideration not only the statistical aspects of the raw data, but also the utility of the mined patterns.

Algorithm 2.6: MSapriori

1 *M= sort(I, MS);/* according to MIS(i)'s stored in MS*/*
2 *F= init-pass(M,T); /* make the first pass over T*/*
3 *L_1= {<f> | f∈F, f.count ≥MIS(f)};*
4 *for (k= 2; L_{k-1}≠Ø; k++) do*
5 *if k= 2 then C_2= level2-candidate-gen(F)*
6 *else C_k= candidate-gen(L_{k-1})*
7 *end*
8 *for each transaction t∈Tdo*
9 *C_t= subset(C_k, t);*
10 *for each candidate c∈$C_t$do c.count++;*
11 *end*
12 *L_k= {c∈C_k| c.count ≥MIS(c[1])}*
13 *end*
14 *Answer = ∪$_k L_k$;*

The simplest method to incorporate utility is called weighted association rule mining, which assigns to each item a weight representing its importance. They can represent the price or profit of a commodity.

The weighted support = $(\sum_{ij \in AB} wj)$*Support(A→ B),

where i_j denotes an item appearing in rule A→ B and wj denotes its corresponding weight.

The first factor of the measure has a bias towards rules with more items. When the number of items is large, even if all the weights are small, the total weight may be large. The second measure, normalized weighted support, is proposed to reduce the bias and is defined as $\frac{1}{k}(\sum_{ij \in AB} wj)$ * Support(A→ B), *where k is the number of items in the rule.*The traditional support measure is a special case of normalized weighted support because when all the weights for items are equal to 1, the normalized weighted support (Cai et al. 1998).

2.5 Interestingness measurements

Interestingness of patterns is one of the features that determine the uses of Data Mining techniques. Thus, defining interestingness measures is a crucial part of the mining process. There are mainly two types of interestingness measures; Subjective measures which depends on specific user's needs and preferences, Objective measures where interestingness of patterns is measured in terms of structure and the underlying data used. Data mining and statistical measurement techniques can be combined to arrive at a more reliable and interesting set of rules.

Assessing rules with interestingness measures is the pillar of successful application of association rules discovery. However, association rules discovered are normally large in number, some of which are not considered as interesting or significant for the application at hand. In the last two decades, interestingness measures, each of which estimates the degree of interestingness of a discovered pattern, have been actively studied(Lenca et al. 2008)(Geng & Hamilton 2006)(Vo & Le 2011)(Ohsaki et al. 2007)(Railean et al. 2013).

Probability-based objective measures evaluate the generality and reliability of association rules have been thoroughly studied by many researchers. They are usually functions of a 2 × 2 contingency table. A contingency table stores the frequency counts that satisfy given conditions. Table 2.1 lists some of the common objective interestingness measures for association rules as in (Geng & Hamilton 2006).

Given an association rule A→ B, the two main interestingness criteria for this rule are generality and reliability. Support P(AB)or coverage P(A) is used to represent the generality of the rule. The support of A→ B is the percentage of transactions that contain A and B. Confidence P(B|A) or a correlation factor such as the added value P(B|A)− P(B)or lift P(B/A)/P(B) is used to represent the reliability of the rule. The confidence of A→B is the ratio of the number of transactions that contain A and B against the number of transactions that contain A.

Table 2.1. Probability Based Objective Interestingness Measures for Rules

Measure	***Formula***
Support	$P(AB)$
Confidence/Precision	$P(B\|A)$
Coverage	$P(A)$
Prevalence	$P(B)$
Recall	$P(A\|B)$
Specificity	$P(\neg B\|\neg A)$
Accuracy	$P(AB)+ P(\neg A\neg B)$
Lift/Interest	$P(B\|A)/P(B)$ $P(AB)/P(A)P(B)$
Leverage	$P(B\|A)- P(A)P(B)$
Certainty Factor	$(P(B\|A)- P(B))/(1- P(B))$,
Conviction	$P(A)P(\neg B)$ $P(A\neg B)$

Association rules should be filtered and sorted according to given goals, thus subjective measures were not used because they relates to individual's satisfaction and rely on user's domain knowledge which is difficult to obtain. Objective measures are preferably used because they do not depend on specific's user knowledge, but depends on the structure of data and patterns extracted from it (Sahar 2010).

The search for the best rules among a vast set of rules generated by a KDD procedure is directly linked to the search and the use of a good interestingness measure. From the user's point of view, the problem can then be resumed as a search for finding the best measure(s) according to the context(Lenca et al. 2008). This context is defined by many parameters such as:

- the nature of the data (what is their type, do they suffer from noise, how imbalanced is the distribution of each attribute?),
- the type of rule extraction algorithm (what are its biases?),
- the goals,
- and the preferences of the user.

2.6 Summary:

Finding association and association rules is one of the descriptive tasks of data mining. It is more than two decades, and the research for finding fast and efficient mining algorithms of association rules is not off. Very different algorithms in construction and prospect have richen this area. Frequent itemset mining has been the core problem in many data mining tasks, and varied approaches to the problem appear in numerous papers across all data mining researches. While the problem was introduced in the context of market basket analysis, the scope of the problem is much broader. Generally speaking, the problem involves the identification of items, products, symptoms, characteristics, and so forth, that often occur together in a given dataset. As a fundamental operation in data mining, algorithms for frequent itemset mining or association rules mining can be used as a building block for other, more sophisticated data mining processes. Finding interesting associations and deciding what and how to measure, is the vital operation for many applications that will be discussed in the next chapter.

CHAPTER THREE

Applications of Association Rules Mining

3.1 Introduction

As larger and larger datasets become available, data mining techniques can be applied to identify patterns of interest in the data. Association rule mining has been the most used data mining technique in marketing, industry, Bioinformatics, and Medicine as will be revealed through this chapter. The researcher reviewed the uses and applications of association rules mining using Apriori, which is considered the base algorithm for the research in all the previously mentioned fields.

3.2 Associations in biological data

There are numerous sources of biological data that provides challenging opportunities for data mining. For example, the structural genomics initiative aims to catalog the structure-function information for proteins. Advances in technology such as microarrays have launched the subfield of genomics and proteomics to study the genes, proteins, and the regulatory gene expression circuitry inside the cell(Zaki 2004).

DNA and protein sequences are essential biological data that exist in huge volumes. Data mining approaches seem ideally suited for Bioinformatics. It is essential to develop effective methods to compare and align biology sequences and discover biosequence patterns. The mass of genomic and proteomic is analyzed with the intention of predicting protein structure/function/interaction (Oyama et al. 2002) (Luksza 2005) (Tang et al. 2005), gene regulation through microarray technology (Tuzhilin & Adomavicius 2002) ((Georgii et al. 2005).

Bioinformatics is a promising young field that applies computer science and technology in the molecular biology and develops algorithms and methods to manage

and analyze biological data, and extraction of useful information from these data. The second problem is one of the leading challenges in computational biology, which requires the development of tools and methods capable of transforming all these heterogeneous data into biological knowledge about the underlying mechanism (Larrañaga et al. 2006).

3.2.1 Gene sequence:

It is possible that important hidden relationships and correlations exist in the data. Association rule mining is an effective method and algorithm to apply to bioinformatics. Association rules(or Genetic Association) in gene, biology sequences, and biosequence patterns are mainly implemented (Wang et al. 2010). Association rule mining can be applied to compare and align biology sequences, finding biosequence patterns, discovery of disease-causing gene connections and in exploring gene-drug interactions.

Metabolic pathways have a relatively long history compared with other biological networks. They characterize the process of chemical reactions that, together, perform a particular metabolic function. With the recent progress in the application of computational methods to cell biology, there have been successful attempts at modeling, synthesizing and organizing metabolic pathways into public databases.(Koyutürk et al. 2004)

Gene regulatory networks, also referred to as genetic networks, represent regulatory interactions between pairs of genes and are generally inferred from gene expression data through microarray experiments (Akutsu et al., 1998). A simple and common mathematical model for gene regulatory networks is a Boolean network model. In this model, nodes correspond to genes and a directed edge from one gene to the other represents the regulatory effect of the first gene on the second. The edge is labeled by either a ‘+’or‘−’ sign to represent the direction of regulation, namely up- or down-regulation, respectively. More sophisticated computational models that capture the

degree of regulation through weighted graphs and/or differential equations have also been proposed.(Koyutürk et al. 2004)

3.2.2 Mining Gene Expression:

As gene expression data sets become larger and larger, spreadsheets will become less and less of an adequate tool for doing analysis (as a single worksheet in Excel can hold no more than 256 columns), and data mining techniques using large databases should find more and more use in analyzing expression data. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression.

Global gene expression profiling, both at the transcript level and at the protein level, can be a valuable tool in the understanding of genes, biological networks, and cellular states. As larger and larger gene expression data sets become available, data mining techniques can be applied to identify patterns of interest in the data. Association rules can reveal biologically relevant associations between different genes or between environmental effects and gene expression. Items in gene expression data can include genes that are highly expressed, or repressed, as well as relevant facts describing the cellular environment of the genes (e.g. the diagnosis of a tumor sample from which a profile was obtained).

One goal in analyzing expression data is to try to determine how the expression of any particular gene might affect the expression of other genes. Another goal of expression data analysis is to try to determine what genes are expressed as a result of certain cellular conditions, e.g. what genes are expressed in diseased cells that are not expressed in healthy cells. (Creighton 2003) developed a database application, a Microsoft Access Database Project (ADP), that implement a version of the Apriori. using the data set from Hughes et al. (Hughes et al. 2000) of 300 expression profiles for yeast. The application accepts an expression data set in the format of one or more spreadsheets as input, with items organized by row, and experiments organized by column. Then it mines the database for frequent itemsets that exist within the data. The

application proceeds iteratively until all frequent itemsets have been found. Additional criteria was specified such as requiring selected itemsets to form at least one rule where the LHS set has a single item. numerous rules were found in the data but with a cursory analysis of some of these rules reveals numerous associations between certain genes, many of which make sense biologically, others suggesting new hypotheses that may warrant further investigation. 40 rules were generated having the number of itemsets which exceeds the seven itemsets and the LHS of the rules contains only 1 item, with minimum support to be 10% and the minimum confidence to be 80%. (Creighton 2003)

Association rules are used to involve multiple motifs and to predict expression in multiple cell types. Association rules are enhanced with information about the distances among the motifs, or items, which are present in the rule. Rules of interest are those whose set of motifs deviates properly. (Member et al. 2010)

Clustering and biclustering techniques are one of the most used computational strategies for analyzing microarrays, but Gene association analysis (GAA) provide associations which do not appear adjacent to each other in a one shot clustering strategy. Though using of frequent itemset is related to frequent subset of genes using Apriori verified by the correlation coefficient. (Alves et al. 2010) recommend that rules have to be evaluated for verifying their biological significance.

Association rule analysis methods are important techniques applied to gene expression data for finding expression relationships between genes. However, previous methods implicitly assume that all genes have similar importance, or they ignore the individual importance of each gene. The relation intensity between any two items has never been taken into consideration. Therefore, we proposed a technique named REMMAR (RElational-based Multiple Minimum supports Association Rules) algorithm to tackle this problem. This method adjusts the minimum relation support (MRS) for each gene pair depending on the regulatory relation intensity to discover more important association rules with stronger biological meaning.(Liu et al. 2011)

3.2.3 Protein structures

There are some biological problems in which experts can specify only input/output pairs, but not the relationships between inputs and outputs, such as the prediction of protein structure and structural and functional sequences. This limitation can be addressed by machine learning methods.(Prompramote et al. 2005)

Analysis of protein sequence and structure databases usually reveal frequent patterns (FP) associated with the biological function. (Chen & Bahar 2004) introduced the discovery of FPs in the protein families as part of the analysis process in order to detect functional patterns in rapidly growing structure databases and providing insights into the relationship among protein structure, dynamics, and function. A set of proteins belonging to a given family is selected as the training dataset. Features are extracted from all the amino acids in the dataset which consists of 780 PDB of serine proteases and 122 entries cysteine proteases. Each amino acid corresponds to one entry, represented by the amino acid sequence index and the extracted features. These entries are organized into 20 groups by amino acid type. An implementation of the Apriori algorithm(Borgelt & Kruse 2002) is applied to each group to find FPs that correspond to different amino acids. The decision on minimum support threshold depends upon the dataset. Without any prior knowledge of functional motifs, the method discovers the frequent protein structures for each type of amino acid and identifies the conserved residues in three protease subfamilies; chymotrypsin and subtilisin subfamilies. A gene interaction network was produced and showed that the catalytic concurrence residues are distinguished by their strong spatial coupling (high interconnectivity) to other conserved residues.

A methodology was implemented by (Stelle et al. 2011) to extract rules that associate hydrophobicity patterns/profiles to specific secondary structures of proteins. Firstly, a local database (DB) with 20,000 proteins extracted from the Protein Database (PDB) was implemented using the database management system (DBMS) PostgreSQL. The database was composed of proteins from four folding classes (a-helix, b-sheet, turn and loop). The primary structure, the size and the class of each protein were stored in the

database together with the physic-chemistry properties of each residue. The amino acid sequences were divided and grouped in windows with different sizes (7, 11, 15 and 21) in the database. The residues were stored with their associated secondary structure. Secondly, an implementation of the Apriori algorithm with a predefined minimum support of 2% was used to identify hydrophobicity patterns/profiles that achieve a specific secondary structure in different proteins. Then rules were extracted with a predefined minimum confidence of 10%. Lastly, selecting the secondary structure formation rules with values equal or greater than 33% for support, values equal or greater than 15% for confidence and values equal or greater than 0.9 for lift and selected the class sequences that were involved in the chosen rules. The results were analyzed by comparing it with the Apriori implemented in the Weka (Waikato Environment for Knowledge Analysis) which yields the same results except for greater database. The class sequences were analyzed and indicated that the technique can be efficient to investigate hydrophobicity patterns/profiles that reaches particular secondary structure and that it can help to identify candidates to structure motifs.

3.2.3 Predict Protein-Protein Interactions

Protein-Protein Interactions (PPIs) play a key role in many essential biological processes in cells, including signal transduction, transport, cellular motion and gene regulation. The comprehensive analysis of these biological interactions has been regarded as very significant for the understanding of underlying mechanisms involved in cellular processes. The prediction of protein interaction sites has gained much attention in recent years. Using classification techniques for analysis and prediction, have shown that the interfaces of interaction sites share common properties that distinguish them from the rest of the protein. But the classification techniques generate prediction models which do not provide users with explicit rules and thus result in low interpretability of the results and poor knowledge extraction capability. So, discovering patterns, in the form of association rules that characterize interaction sites in different PPI types will be more useful (Park et al. 2009).

Protein interaction networks comprise groups of interacting proteins that are observed experimentally. They provide the experimental basis for the understanding of the modular organization of the cells as well as useful information for predicting the biological function of individual proteins.(Koyutürk et al. 2004). Graph mining is a powerful tool for finding motifs and commonly occurring patterns in datasets that contain interactions. With the progression of molecular biology from sequences to biological networks, motif and pattern discovery become interesting and useful for such networks as for sequences. Graph mining algorithms are generally based on frequent itemset mining where graphs (pathways) correspond to transactions and connected edge sets correspond to itemsets. (Koyutürk et al. 2004) the goal was mining metabolic pathways to discover common motifs of enzyme interactions that are related to each other. Mining the pathways for different support thresholds allows evaluation of frequent sub-pathways in a multi-level fashion. For relatively high support values pathway collections were mined to obtain meaningful results in terms of the size of the discovered frequent sub-pathways. For lower values of support, many sub-pathways turn out to be frequent and the size of the frequent pathways also grows significantly.

A dataset containing records of interactions between a set of HIV proteins and a set of human proteins has been analyzed using association rule mining has been published (Mukhopadhyay et al. 2010). The main objective is to identify a set of association rules among the human proteins with high confidence. The interaction data set handled in the research is composed of; group-1 interactions representing direct physical interactions, group-2 interactions representing indirect interactions, with a total of 1288 group-1 and group-2 interactions between 17 HIV-1 proteins and 773 human proteins, a binary matrix of size 17 × 773 was constructed. The rows represent the viral proteins(transactions) and the columns represent the human proteins(items). An entry of 1 in the matrix denotes the presence of interaction between the corresponding pair of HIV-1 and human proteins, and an entry of 0 represents the absence of any information regarding the interaction of the corresponding viral and human proteins. Initially, it is treated as non-interaction. The resulting binary matrix is treated as the input to the ARM algorithm. The Apriori algorithm was applied on the transactions to find frequent itemsets and from these frequent itemsets, highly confident Association Rules were

extracted concentrating only on the rules with only one item in the consequent. The human proteins represented the antecedent and the consequent parts of the rule. New viral-host interactions were predicted based on the discovered association rules that have high confidence (Mukhopadhyay et al. 2010).Park et al (2009) describe a computational approach for the prediction of PPI types employing association rule based classification (ARBC), which includes association rule generation and posterior classification based on the discovered rules. Their aim was to discover patterns in form of association rules in order to be able to characterize interaction sites in different PPI types. 147 protein complexes were selected from the PDB. ARBC comprises three main steps. Association rules were generated using 10 g Oracle Data Miner (ODM), which implements the classical Apriori algorithm with minimum support and confidence of 3% and 25% respectively. Association rules were pruned by removing redundant information. The last step is classification based on the pruned set of association rules, they generate a rule profile consisting of an m × n matrix, where m is the number of examples and n is the number of different association rules obtained after the pruning step. The rule profile matrix takes values of 1 or 0 depending on whether the different rules are dependent or not. Over 354 known PPI domains using 14 properties yielded a total of 1.168 rules, but only 157 were selected after the pruning process. ARBC performed competitively with other methods, and building the prediction model using association rules is interpretable and straightforward and simple for a biologist to work with (Park et al. 2009).

Christian Borglet's Apriori implementation in Java was applied by (Becerra & Vanegas 2009) and (Besemann et al. 2004) to build a biological sequence feature classification and find differences between items(amino acids) belonging to different interacting nodes or different protein interaction networks. The result in the first study is passed into predictor (SVM + ANN) in order to generate the classifier. While results of the second study were able to confirm expected biological knowledge as well as identifying as yet unknown associations that were successfully supported by further inspection of the data. While (Hung & Chiu, in 2007) used the graphical interface of the algorithm ARView to inspect the relationship of protein functional regions in PPI. They used a dataset from the Database of Interacting Proteins (DIP) and Universal Protein Resource (UniProt). Redundant data was filtered and stored in a MySQL database

Gupta, Mangal, & Tiwari (2006) tried to decipher the nature of associations between different amino acids that are present in a protein. by trying to predict if there are any co-occurrence patterns among the 20 amino-acids by means of quantitative association rules. They applied the Apriori algorithm based on the partitioning approach. 12 association rules were resulted by 30-50% for minsup and minconf respectively. The resulted rules discover rules based not only the presence of amino acids but also on absence, thus acknowledging the fact that absence of a particular amino acid can also be important to the structure and/or function a protein (Gupta et al. 2006). AR can also be used to represent the relations between the features of a single protein. (Oyama et al. 2002) decided to regard an interaction itself as a transaction. Interaction data was represented as a pair of two proteins that directly binds to each other, where each transaction represents an interaction, has the features of both left 'LSP' and right 'RSP'-hand side protein of the interaction protein pair.

3.3 Associations in Medical Data

There is a growing need in the health sector to store, organize and analyze medical data, assist the health care professionals in decision making, and develop data mining methodologies to mine hidden patterns and discover new knowledge from clinical.

Analysis of electronic medical records assist the health care professionals in decision making and develop methodologies and techniques to generate knowledge from the huge data warehouses in hospitals as known as clinical data. Some studies regarding finding pattern of interest (Meyfroidt et al. 2009), diagnosing of some diseases such as cancer (Li et al. 2004) and more specifically for breast cancer survivability(Delen et al. 2005)and search for useful associations in such large databases(Almodaifer et al. 2011).

Medical diagnosis researches as in (Kumar et al. 2011)(Soni et al. 2011)(Karabatak & Ince 2009)(Kharya 2012)(M.-J. Huang et al. 2007)(Ha 2011) have proved great success, because the data about the disease and the patient under

examination is always available, in fact the medical diagnostic knowledge can be automatically derived from the description of cases solved in the past.

Ordonez et al's experiments focus on discovering association rules on a real data set to predict absence or existence of heart disease (Ordonez et al. 2005). Association rules are constrained to reduce the number of discovered patterns and they are summarized to get a concise set of rules. The significance of association rules is evaluated using support, confidence and lift. The proposed constraints include maximum association size, an attribute grouping constraint and an antecedent/consequent rule filtering constraint. Association rules are summarized using rule covers in order to summarize rules having the same consequent. Two main measurements to quantify the quality of medical findings are sensitivity which refers to the probability of correctly identifying sick patients and specificity which is the probability of correctly identifying healthy individuals trusted by a medical doctor.

The relationship of the symptoms and disorders in the medical databases were mined to find frequent illnesses and generates association rules using Apriori algorithm. constructing association rules for disorders frequently seen in the patient and determining the correlation of the health disorders and symptoms observed simultaneously. (Zuhtuogullari & Allahverdi 2011) mined the relationships of the symptoms observed together using C#.net programming language. The itemset generation process can be stopped by the user according to the itemset number.

3.4 Associations and recommendation system

Recommender systems help users find and evaluate items of interest. They connect users with items to "consume" (purchase, view, listen to, etc.) by associating the content of recommended items or the opinions of other individuals with the consuming user's actions or opinions. Recommender systems that incorporate data mining techniques make their recommendations using knowledge learned from the actions and attributes of users. One of the best-known examples of data mining in recommender systems is the discovery of association rules or item-to-item correlations

which identify items frequently found in “association” with items in which a user has expressed interest(Schafer 2009).

The research on the recommendation system for tourism is boosting. The recommendation system for tourism is to provide information or suggestions on items concerning tourism, such as travel agencies, hotels, tourism route, and tourism attractions; it can simulate a travel consultant to help the users to realize the recommendation services process like an intelligent online assistant. The functions cover converting browsers into real tourists, increasing the cross-sell of the related products of this website and building users’ loyalty to the website.(Zhang et al. 2009).

(Hu et al. 2008) Considers an Apriori-Based Personal Recommendation Algorithm to get the data of those access web pages. They constructed a matrix model having relatively high purchasing power about customer behavior, in order to get the similar access behavior over the all or partial property space with high efficiency. Thus, will help the customer find out the products he wishes to buy, through mining of the similar pattern character between latent buyer and high buyer and consequently promote customer satisfaction and truly promote the sale achievements for the enterprise.

Association rules recommendation is a realization of an elementary level recommendation technology Amazon, CDNOW, eBay. Association rules analyze the customer’s need and preference and thus recommend customized product and service. Recommender system has won great success in e-commerce because it increase new users, promotes sales, enhances the satisfaction of users. Recommender system focuses on the user who shares the similar preference with object user referring to his preferred product to predict the object user’s favorable product and hence figure out the recommendation. In (Zhang et al. 2009) an evaluation record of new or old tourist destinations or scenic spot by the visitor can be found on the internet and then recommend a tour.

Web usage mining focuses on techniques that could predict user behavior while the user interacts with the web, it deals with the extraction of interesting knowledge from logging information produced by Web servers. Apriori algorithm was applied to the log

files of a website (primarily, an e-commerce or an e-learning site) to help the users to the selection of the best user-tailored links (Lazcorreta et al. 2008).Each user access to websites has their own purpose where a conversation was obtained from their browsing behavior. Conversation is a page sequence when user access to the web. Association rules applied to such matters purpose is to find a site's content pages relationship or so-called frequent access paths (Jingfang & Busheng 2011)(Kosala & Blockeel 2000)(Facca & Lanzi 2005)(Mobasher et al. 2002).

Customer Relationship Management (CRM) System applies data mining techniques to quickly and effectively analyse the massive data and information about customer profitability, customer segmentation, cross- selling analysis, customer access and maintaining. To find out regulations and patterns, acquire necessary knowledge, and help the enterprise to have a better decision-making and a high rate of return.(Gong et al. 2007). CRM System applies Data Mining Association Rules into processing Cross-selling Analysis. Cross-selling is a process that the company provides the current customers with the new product and the new service. The reasonable sales match acts as a key of Cross-selling. The companies need to know which products should be put together to sell, to make a more successful enterprise's Cross-selling. The decision which products should be put together to sale and point at the possibility of what customers to buy will be made to realize the reasonable sales match.

3.5 Summary

Association rule mining has been widely used in many different areas and applied almost to various data types. Association rules handled transaction data in market basket analysis, biological and medical data and many more areas. AR has also been a preliminary step for other data mining techniques such as clustering, classification and recommendation systems. It is apparent that using specifically the Apriori algorithm or Apriori based plays a great role in the research area of Association rules mining, as described in this chapter. Many researchers act solely in solving their

individual problems, although there is no unique platform that can handle all types of data and generate meaningful association rules for different perspectives.

CHAPTER FOUR

Methodology

4.1 Introduction

This chapter will describe the structured methodology adopted by the researcher to design an intelligent association rule mining framework. To achieve the objective of this research, the methodology was carried on three phases. The review of the most used algorithms, inspecting their design issues, modification and improvement were implemented in chapter two. Although there are a large number of association rule mining algorithms, the Apriori algorithm was the most researched. The researcher found that many algorithms adopted the Apriori as a basic strategy and tended to adapt the whole set of procedures and data structure as well. Chapter three explored the wide use of association rule mining and many areas to serve many applications. The Apriori algorithm was dominating the research spectrum.

The researcher attempts to solve the following questions for a non-specialist:

- Which algorithm to choose according to a specific area?
- Which interestingness measurements would yield to the correct association rules?
- How should the parameters be set?

Figure 4.1 identify the main steps followed to answer the questions mentioned theoretically.

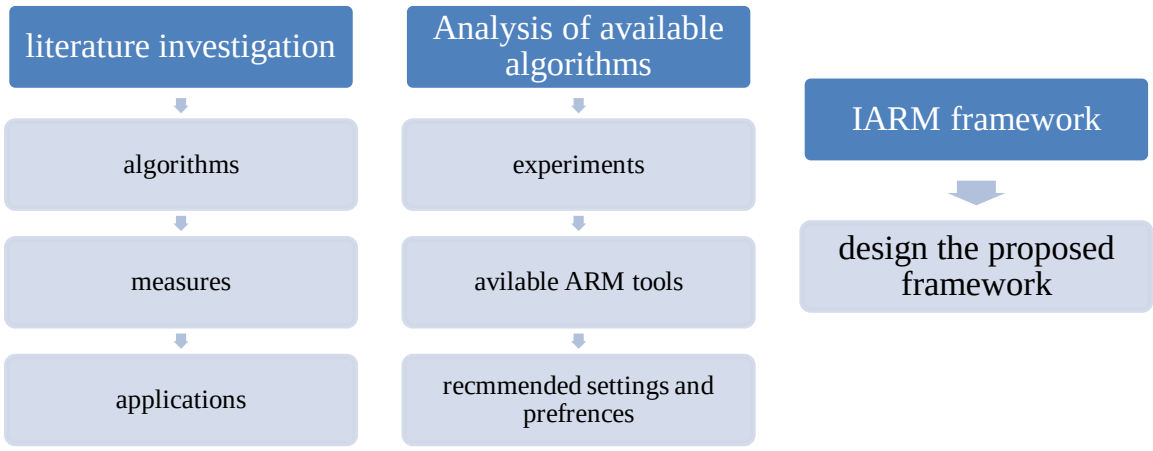


Figure 0-1: Research Methodology

4.2 Literature investigations

In 1993, Agrawal, Imielinski, and Swami introduced the algorithm which mines large collections of market basket databases in order to extract useful rules that might help in business decision. And since then, a lot of algorithms have been applied some algorithms that are based on the classical Apriori algorithm, finding association rules for many other applications rather than market basket as discussed in chapter three.

The literature survey observed that Apriori algorithm is the well-known association rules algorithm. Apriori has its unique advantages in mining frequent itemsets and has been used for most commercial products. The performance of the algorithm used, how the parameters were set and the choice of measures applied heavily depends on the nature of datasets. For some applications, only frequent itemsets were needed and for others, specific rules were mined. Table 4.1. summarizes the choice of other measures rather than the support and the confidence along with the specific algorithm used are according to to the literature.

In the retail business, it is obvious that the dominant algorithm used is the classical Apriori algorithm as described in (Agrawal & Srikant 1994) and programmed in different programming languages such as Matlab, Java or C. Some researchers

implemented Cristian Borgelt's version of the algorithm, which is also available through *(http://www.borgelt.net/apriori.html)*. Market basket analyists searches for frequent items sold and association rules between items which are sold together.

Table 4 -1: Summary of mining objectives, measures, and algorithm

Application	Output	Algorithm	Measures	Others
Market Basket	– Frequent patterns – Association rules	– Apriori – MSApriori – DIC	– Support, confidence	– Lift
Gene Expression	– Gene Associations – Genes Bindings to motif – Frequent sub genes	– Apriori – AprioriC – Apriori SMP	– Support, confidence	– Chi-square value, – correlation coefficient
Proteomic	– Frequent Structures – Protein interactions – Binding sites	– Apriori	– Support, confidence	– lift
Medical Diagnosis	– Symptoms observed together – Disease diagnoses	– Apriori – AprioriC – MSApriori – CAR	– Support, confidence	– Lift, Sensitivity and Specifity

Transactional databases are frequently incrementing and growing, thus using the Dynamic Itemset Counting (DIC) algorithm described in (Brin et al. 1997) is preferred especially for the increasing volumes of data. But, for decision making or if the business analyst may want to investigate the association of selling a specific good with other goods, using multiple support(MS) for some identified itemset will be helpful(Hu & Chen 2006).

In many bioinformatics problems, biologists are interested in comparing different sets of items. They apply association rule mining to compare and align biology sequences, find biosequence patterns, the discovery of disease-causing gene connections and gene-drug interactions. Association rules were used to build a classifier after identifying the patterns which turn to be features for the classifier for many medical researches. Association rules can reveal patterns that might not have been revealed using clustering techniques in bioinformatics and industry. Association rules can reveal

patterns that might not have been revealed using clustering. Classified association rules(CAR or AprioriC) are both used in medical databases and genetic databases when the scientists need to bind some symptoms or genes with some disease.

During the literature investigation, the researcher noticed that the term's naming is different when working on different areas. Table 4.2 maps this naming for each application to support ease of use. In the context of market basket analysis, a gene expression profile can be thought of as a single transaction, and each transcript or protein can be thought of as an item. However, while in market basket analysis any particular item is either purchased or not purchased in a transaction, in an expression profile each transcript or protein is assigned a real value that specifies the relative abundance of that transcript or protein in the profiled sample. In applying association rules to gene expression data, one technique would be to first bin each measured value as being up (i.e. highly expressed), down (i.e. highly repressed), or neither up nor down.

Table 4-2 Naming mapping

Application area	Terms used
Protein Protein Interaction	– items as features
Protein sequences	– Items as amino acids – Transactions as protein sequences
Microarray(gene expression)	– Items as genes transcription: expressed or repressed – Transactions as gene expression profile
Recommendation systems	– Items as user prefrences – Transactions as opinions – Asscociation rules as recommendations

4.3 Exploration of algorithms and applications

Repositories of frequent item set mining algorithms are found in (Goethals 2003) which includes the implementations and data sets of frequent item set mining algorithms. Arules(Hahsler et al. 2005) can provides the infrastructure for representing,

manipulating and analyzing transaction data and patterns (frequent itemsets and association rules). Arules also provides interfaces to C implementations of the association mining algorithms Apriori and Eclat by C. Borgelt. The Association Rules Miner, ARMiner has been written in Java and it is distributed under the GNU General Public License(Cristofor 2006). ARMiner is a client-server data mining application specialized in finding association rules. ARtool represents a collection of algorithms and tools for the mining of association rules in binary databases.

Plenty of tools are available for data mining tasks using artificial intelligence, machine learning and other techniques to extract data. The following section describes five of the most frequently, powerful open source data mining tools available as rated in(Imanuel 2014). The researcher discusses their formats, how they deal with dataset, data mining techniques handled and in particular the generation of association rules.

4.3.1 Orange

Orange is an open source data visualization and analysis tool for novice and experts developed at the Bioinformatics Laboratory at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. Orange is a component based data mining and machine learning software suite written in the python scripting language. Orange's graphical user interface is illustrated in figure 4.2.

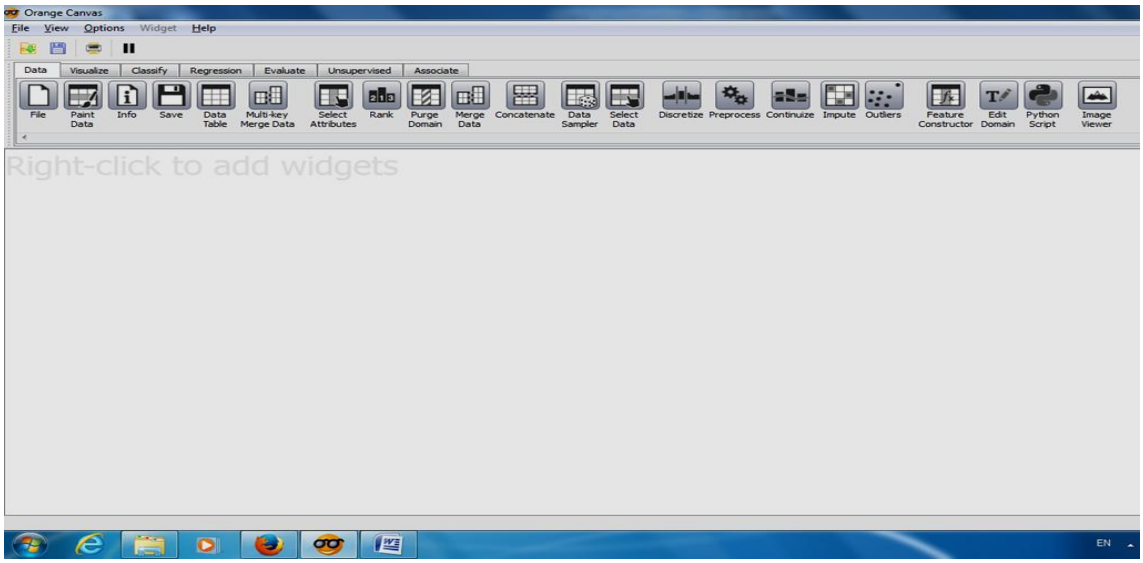


Figure 0-2: Orange Canvas

Data mining can be done through visual programming or Python scripting. It has a component for machine learning methods for classification, or supervised data mining as displayed more clearly in figure 4.3. There are add ons for Bioinformatics and text mining. It is also packed with features for data analytics, different visualizations, from scatter plots, bar charts, trees, to dendrograms, networks and heat maps.

Association Rules Sparse Inducer induces frequent itemsets and association rules from sparse data sets. These can be either provided in the basket format or in an attribute-value format where any entry in the data table is considered as presence of a feature in the transaction (an item), and any unknown (empty) entry signifies its absence. Association Rules Inducer works feature-value data, where an item is a combination of feature and its value (e.g., *astigmatic=yes*). Firstly, the user must click on the file icon illustrated in figure 4.3 and apply filters, preprocess, or merge the data file which might be in many formats. Orange can read files in native tab-delimited format, or can load data from any of the major standard spreadsheet file type, like CSV and Excel as described in figure 4.4.

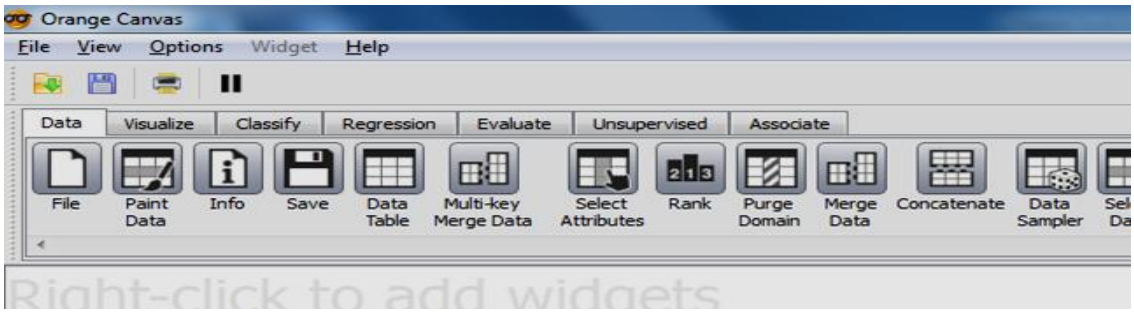


Figure 0-3 Data mining tasks as presented in Orange Canvas

Orange provides two algorithms for induction of association rules, a standard Apriori algorithm [Agrawal, Srikant1994] for sparse (basket) data analysis and a variant of Apriori for attribute-value data sets. Both algorithms also support mining of frequent itemsets. Figure 4.5 display the associate tab, but lacks the description of the underlying algorithms.

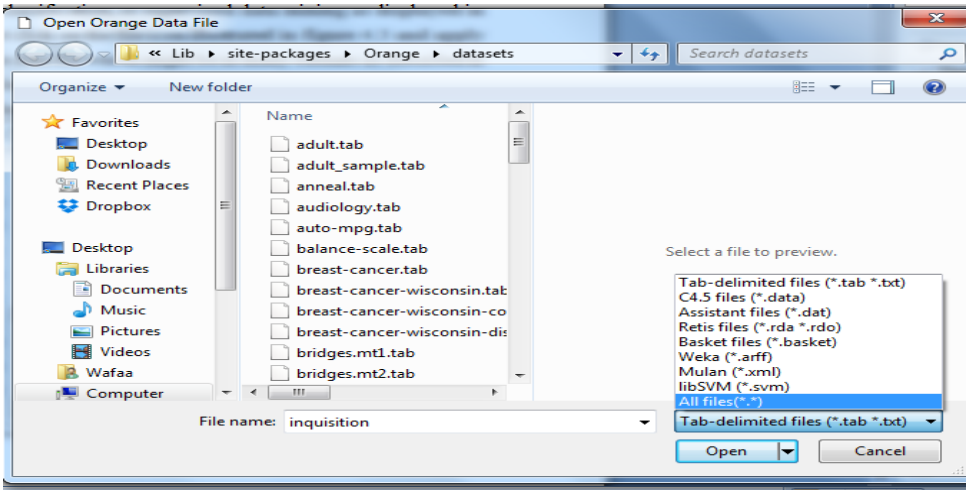


Figure 0-4: Data formats

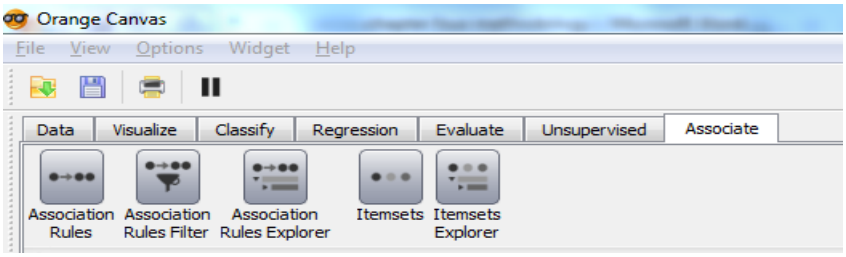


Figure 0-5 Associate tab

There are a lot of tasks which can be performed in Associate tab. Working on any tasking is performed by dragging widgets as described in figure 4.6.

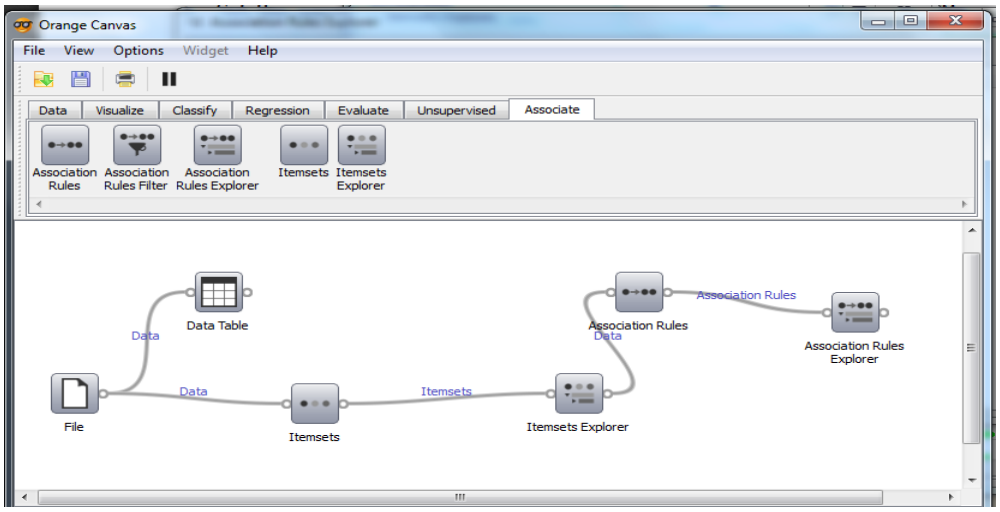


Figure 0-6: dragging widgets on the canvas

If the association rules icon is selected, a window appear which specify the selection of the algorithm and defining the support and confidence threshold value along with maximal number of rules as in figure 4-7(b). The user of the Orange tool, may want to find only the frequent itemsets, and this can be achieved after defining the support threshold and the maximal number of rule as in figure 4-7(a). Depending on the data set, the minimal support value should be set to sufficiently high value to avoid running out of working memory (default: 4.6). Orange will stop with inference of frequent itemsets once this number of itemsets is reached (default: 1000). The Association Rules Explorer window shown on figure 4.8, list the number of discovered rules along with other preferences of options selected. a report of all tasks can be invoked at any time during working with different tasks. Figure 4.9 illustrate a summary report for the whole process.

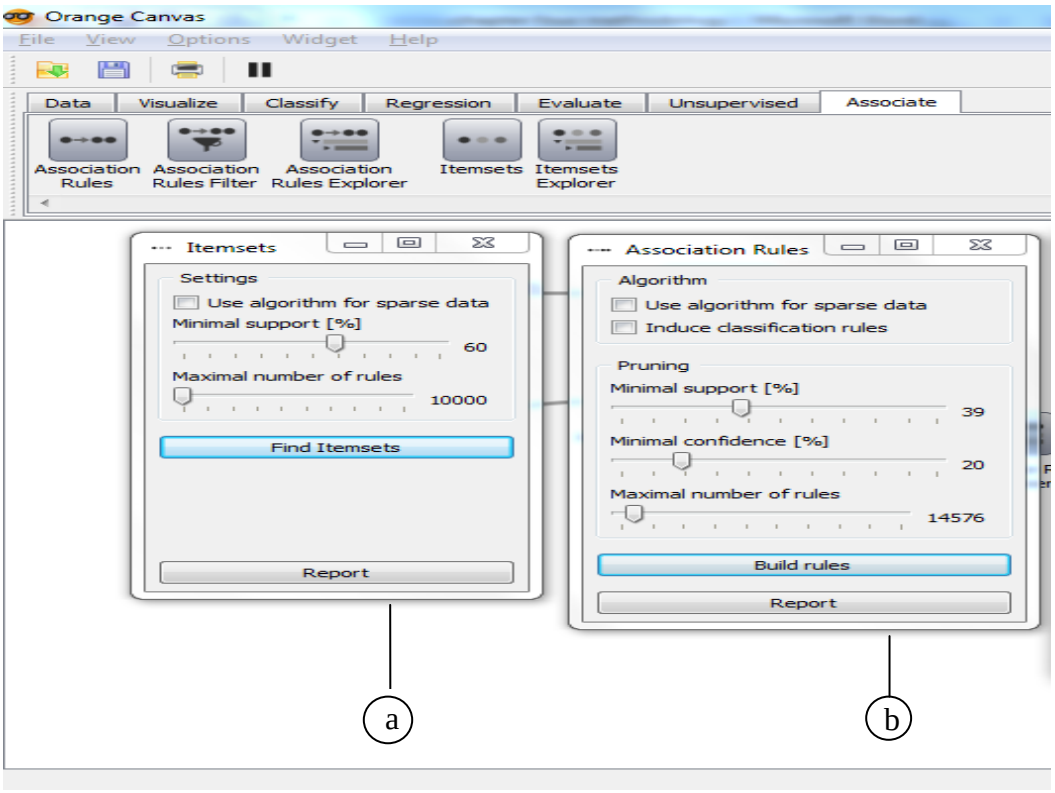


Figure 0-7: Mining frequent patterns as in (a) and Association rules (b)

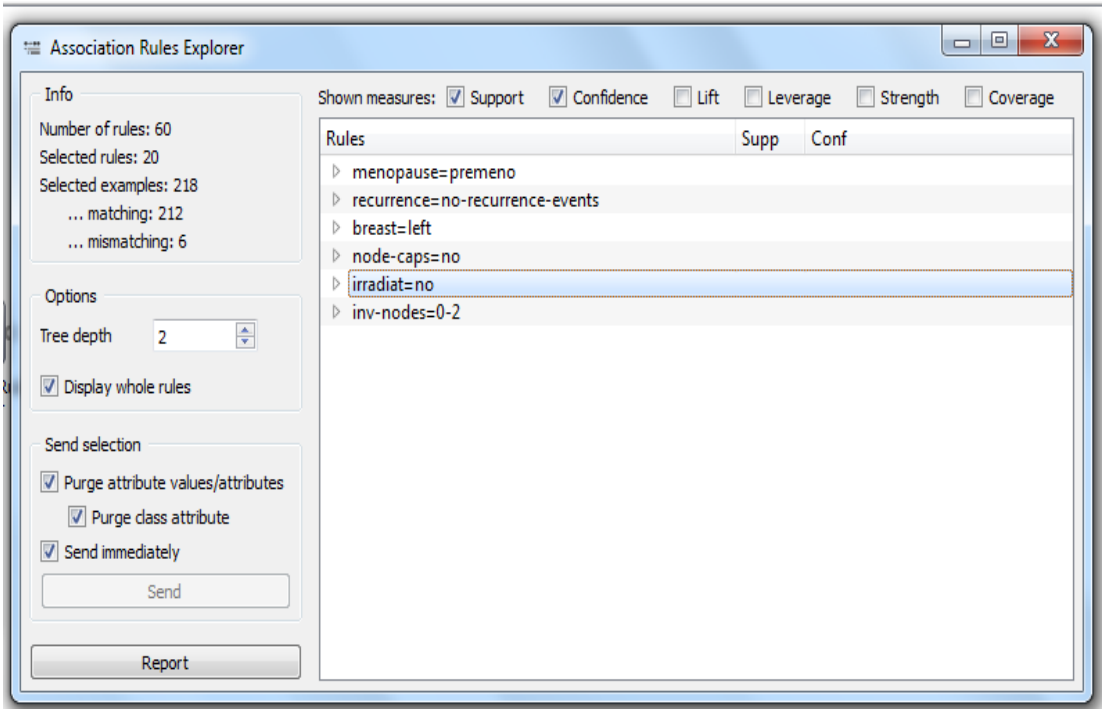


Figure 0-8: Association Rules Explorer widget

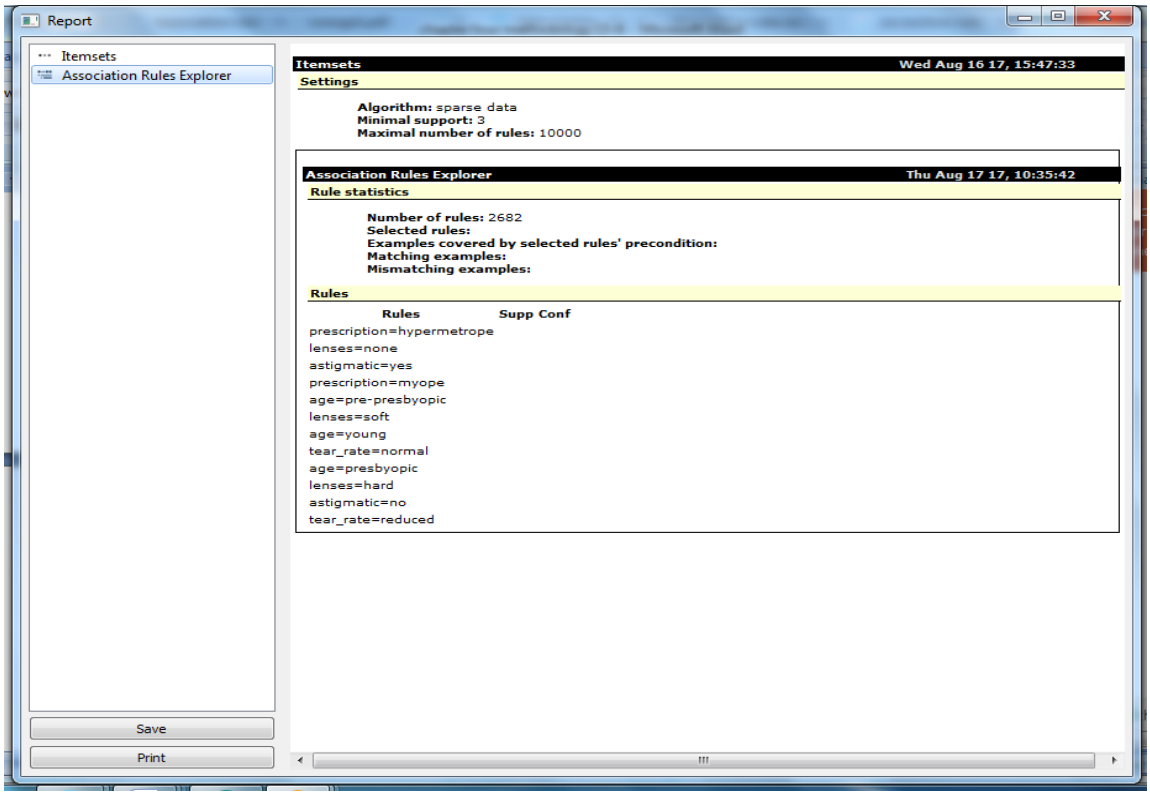


Figure 0-9: Summary report

Association rule induction from non-sparse data sets. An item is a feature-value combination. Unknown values in the data table are ignored. The algorithm can also be used to search only for classification rules where the feature on the right-hand side is the class variable.

Orange remembers the choices, and suggest most frequently used combinations, and intelligently chooses which communication channels between widgets to use. By combining the various widgets the design of data analytics framework can be done. There is over 100 widgets with coverage of most standard data analysis tasks and specialized add-ons for Bioinformatics. Orange can read files in native and other data formats, it comes with multiple classification and regression algorithms.

4.3.2 Weka

Weka is an open source software issued under the GNU General Public License. Weka is a suite of machine learning software applications written in the Java

programming language. Weka is Waikato Environment for Knowledge Analysis, developed at the University of Waikato, New Zealand. Weka contains many paradigms; explorer, experimenter, knowledge flow and simple command line interpreter, as noticed in figure 4.10. It is a collection of machine learning algorithms for data mining tasks such as classification, clustering, and association rules, as shown in figure 4.11. Weka also contains tools for data pre-processing, and visualization.

Figure 0-10: Explorer's widget

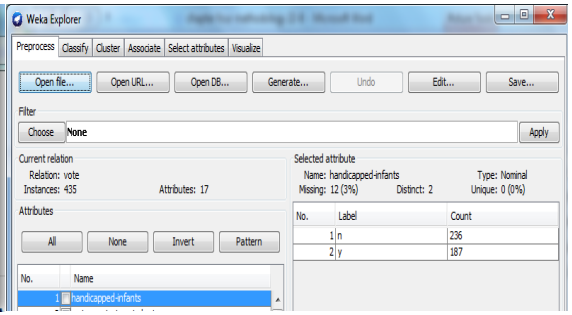


Figure 0-11: Weka paradigms

Data can be imported from a file in various formats such as ARFF, CSV, binary, from a URL or from an SQL database using JDBC. A synthesized dataset can also be generated as visible in figure 4.11. Weka provide access to SQL databases using Java Database Connectivity and can process the result returned by a database query. Pre-processing tools are called filters, and there are filters available for discretization normalization, resampling, attribute selection, transforming and combining attributes.

The Associate tab on Weka enables the user to choose from algorithms which are perceived in figure 4.12. Apriori algorithm expects data that is purely nominal: If present, numeric attributes must be discretized first. Weka runs an Apriori - type algorithm to find association rules, but this algorithm is based on Christian Borglet's Apriori. The FilteredAssociator runs also the Apriori algorithm but apply a filter to the data as described in figure 4.13.

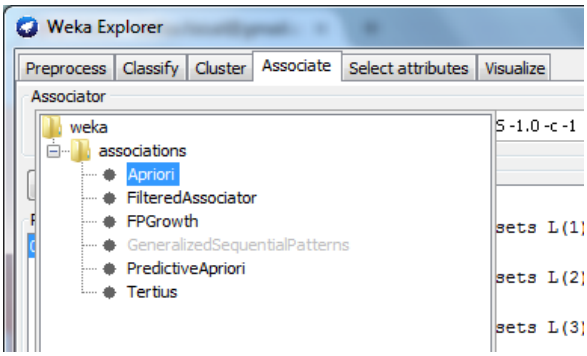


Figure 0-12: Associator Algorithms in Weka Figure

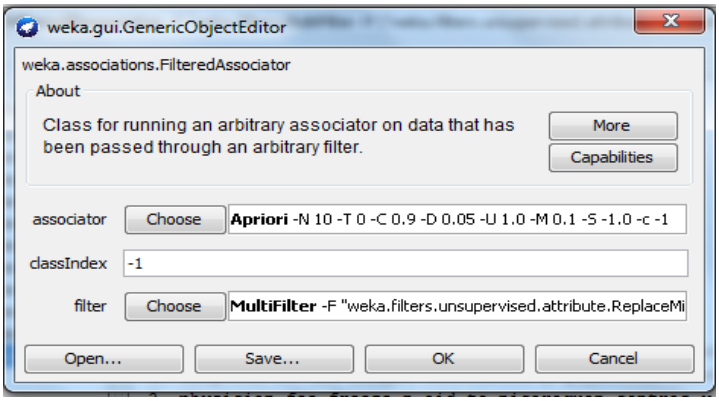


Figure 0-13: FilteredAssociator Algorithms

The Apriori algorithm which is used is the default algorithm selected. However, in order to change the parameters for this run (e.g., support, confidence, etc.), the user has to click on the text box immediately to the right of the "Choose" button. Note that this box, at any given time, shows the specific command line arguments that are to be used for the algorithm. The dialog box for changing the parameters and other options like the number of rules to be displayed and itemets is represented in Figure 4.14.

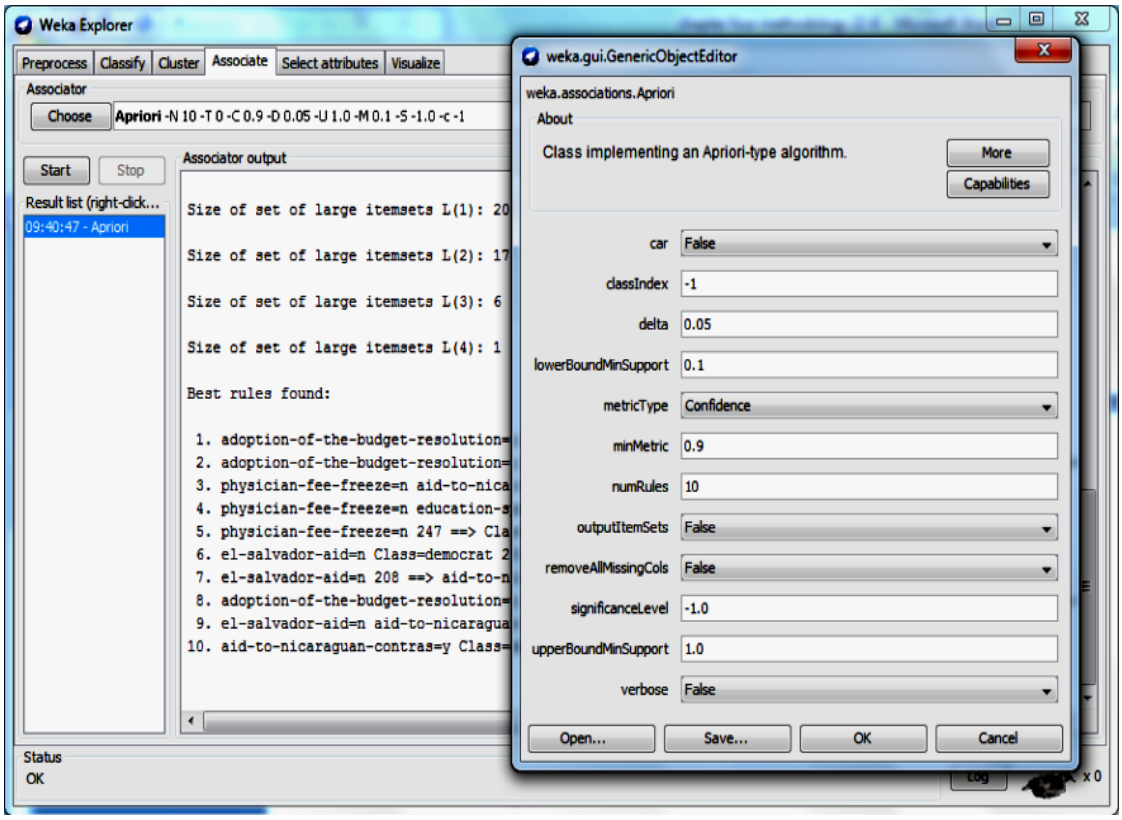


Figure 0-14: Setting parameters

The min. support is not fixed. This algorithm starts with min. support as upperBoundMinSupport(default 1.0 = 100%), iteratively decrease it by delta(default 0.05 = 5%). Note that upperBoundMinSupport is decreased by delta before the basic Apriori algorithm is run for the first time. The algorithm stops when lowerBoundMinSupport (default 0.1=10%) is reached, or required number of rules – numRules(default value 10) have been generated. all frequent itemsets found will not be displayed in the result pane unless the choice of outputItemSets is set as True.

The rules generated are ranked by metricType(default Confidence). Only rules with score higher than minMetric(default 0.9 for Confidence) are considered and delivered as the output. Other metrics are also available such as lift, leverage and conviction. Figure 4.15 illustrate the output of the assciator, where the parameter setting appears first the number of frequent itemset are shown and displayed(if set to true) and finally the rules will be displayed. The first 10 rules will be displayed by default unless altered by the user.

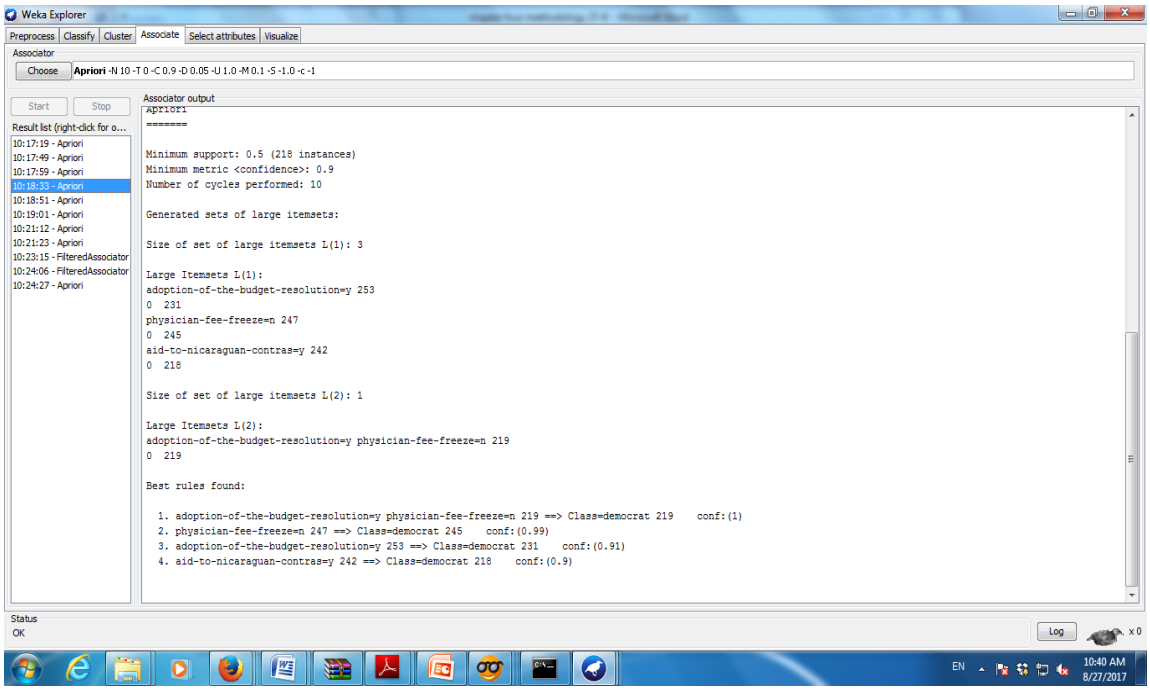


Figure 0-15: Associator output pane

4.3.3 Knime

The Konstanz Information Miner, KNIME is an open source data analytics, reporting and integration platform written in Java and based on Eclipse.It makes use of its extension mechanism to add plugins providing additional functionality. KNIME integrates various components for machine learning and data mining through its modular data pipelining concept and provides a GUI that allows assembly of nodes for data preprocessing and visualization as described in Figures 4.16 and 4.17. It contains over 1000 data analytics routines, either natively or through R and Weka. KNIME analytic workflows can be run through interactive user interface and in batch execution mode, enabling the data analysis process to be easily integrated into local job management and executed on a periodic basis.

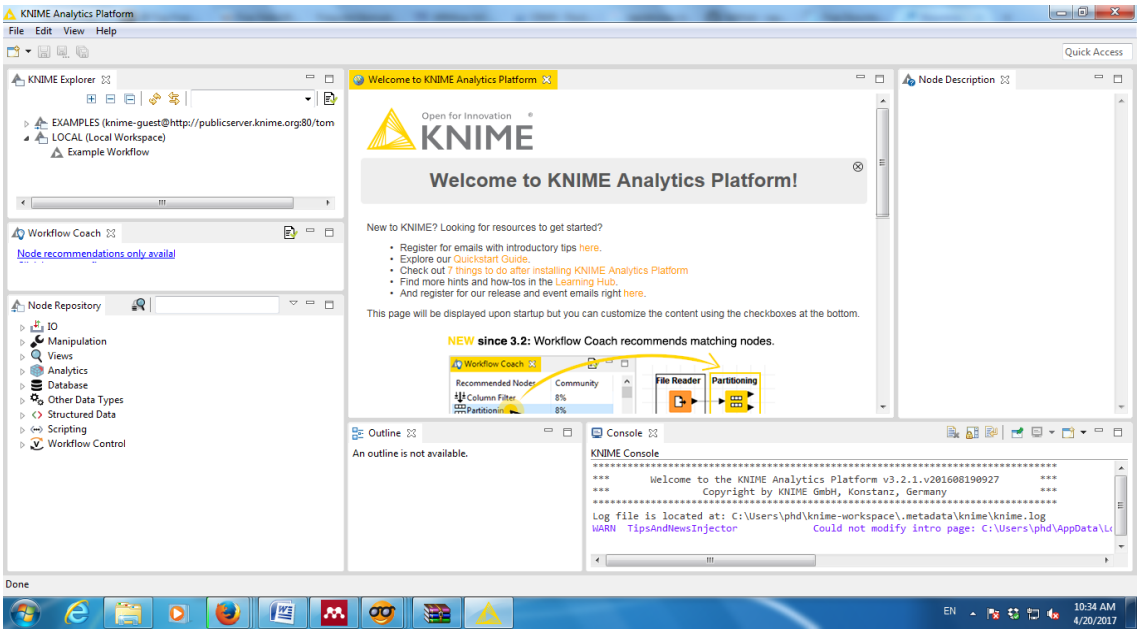


Figure 0-16: Knime GUI

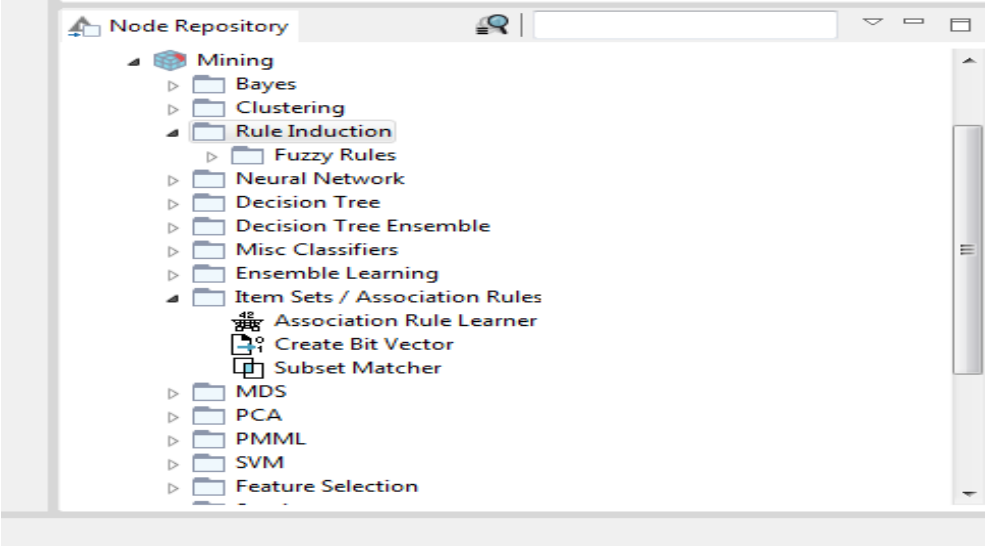


Figure 0-17: Data Mining Tasks on Knime

The user works on the Knime workplace, where he/she can select the desired object which will be displayed in it. One should only have to click from the menu or drop the icon to be displayed in workplace. The first step is to open or read file from the existing data sets. Knime can analyze and mine data files in CSV and ARFF format as shown in figure 4.18.

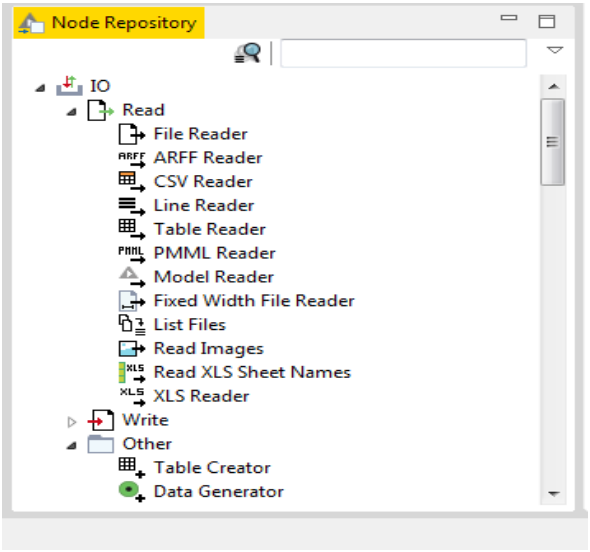


Figure 0-18: Read node of data set types

Item Sets and Association rules is a selection from other mining tasks as noticeable in figure 4.16. Knime, Searches for frequent itemsets meeting the user-defined minimum support criterion and, optionally, creates association rules from them. Figure 4.19, shows the contents of the Knime Associator project workplace.

The column containing the transactions (BitVectors or Collections) has to be selected. Figure 4.18 show how to perform Association rule mining in the Knime workplace. The minimum support as an absolute number must be provided (therefore a check must be performed on the number of transactions to obtain a sensible criterion). If the frequent itemsets should be free (unconstrained) or closed or maximal has also be defined. Closed itemsets are frequent itemsets, which have no superset with the same support, thus providing all the information from free itemsets in a compressed form. Maximal itemsets are sets which have no frequent superset at all. The maximal itemset length must also be defined.

If association rules are generated, a confidence value has to be provided. The confidence is a value to define how often the rule is right. Association rules generated here are in the form to have only one item in the consequence. The underlying data structure used by the algorithm can be either an ARRAY or a TIDList. Choose the

former when there are many transactions and less items, and the latter if the structure of the input data is vice versa. These different settings are clearly clarified in figure 4.20.

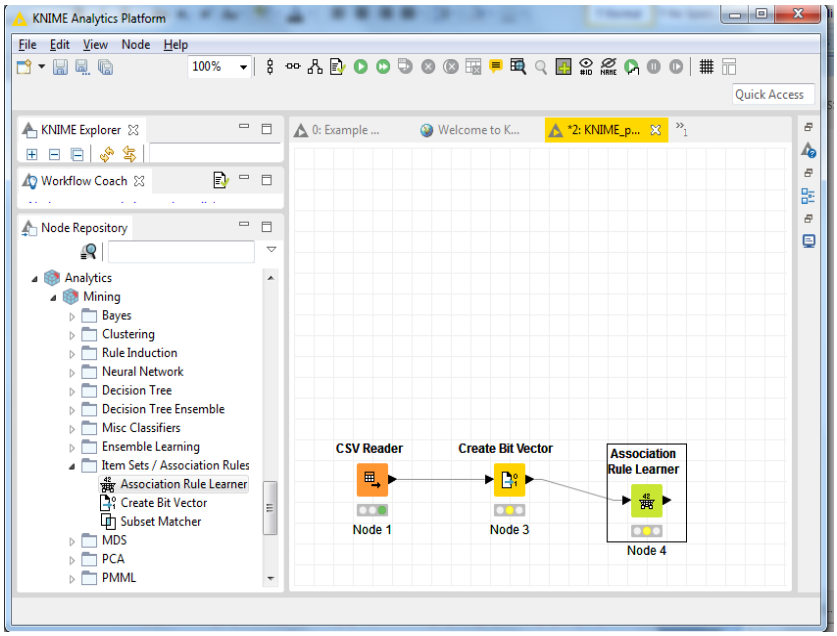


Figure 0-19: Associator project

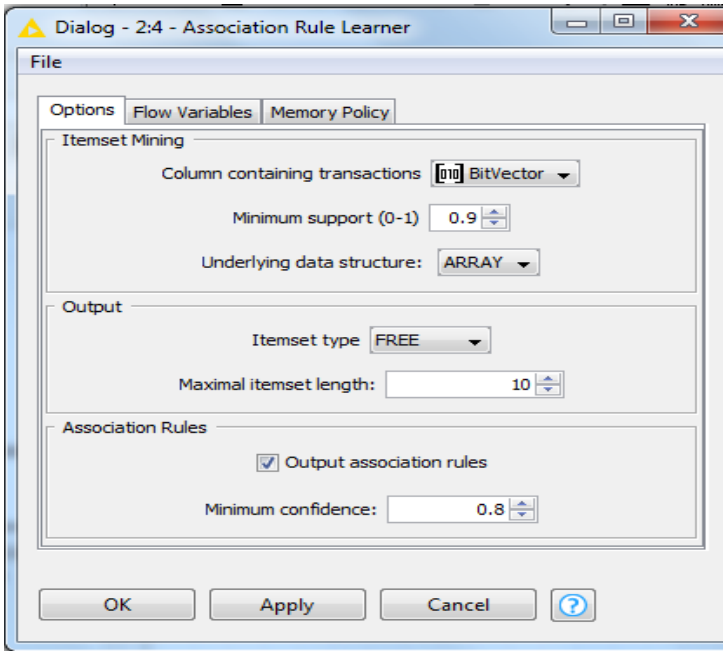


Figure 0-20: Association rule learner

4.3.4 Tangara

TANAGRA is a free DATA MINING software for academic and research purposes. It was developed in France and released in 2004. It is the successor of SIPINA, a classification program. It proposes several data mining methods from exploratory data analysis, statistical learning, machine learning and databases area. TANAGRA is more powerful, it contains some supervised learning but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection and construction algorithms...

The main purpose of Tanagra project is to give researchers and students an easy-to-use data mining software, conforming to the present norms of the software development in the domain (especially in the design of its GUI and the way to use it), and allowing to analyse either real or synthetic data. Another purpose, is to propose to researchers an architecture allowing them to easily add their own data mining methods, to compare their performances. TANAGRA acts more as an experimental platform in order to let them go to the essence of their work, dispensing them to deal with the unpleasant part in the programming of this kind of tools.

TANAGRA has three windows: data mining diagram, components and output as shown in figure 4.21. It has a 'drag-and-drop' type interface, where the user can drag icons (from the components window) and drop them into a nested diagram that represents a set of processes. The diagrams can be saved.

TANAGRA's has many categories of components as outlined by the round edges box of figure 2.21. the Associate category consists of the following as described in figure 2.22. Tangara implements, as well as the other tools, the Borgelt's "apriori.exe" program as its default algorithm for finding frequent itemsets and association rules.

To extract the frequent itemsets and association rules, the component was added into the diagram. A right click on a component in the diagram, brings up a small menu which is shown in figure 4.23. The user can click on the PARAMETERS contextual menu to specify the settings of the analysis illustrated in figure 4.24 and figure 4.25 for

frequent itemsets and rules respectively. The user had also the opportunity to select where he/she wants to save his output file. One of the options in that menu is 'Execute' which runs each component from the start of the diagram, down the hierarchy to the selected component.

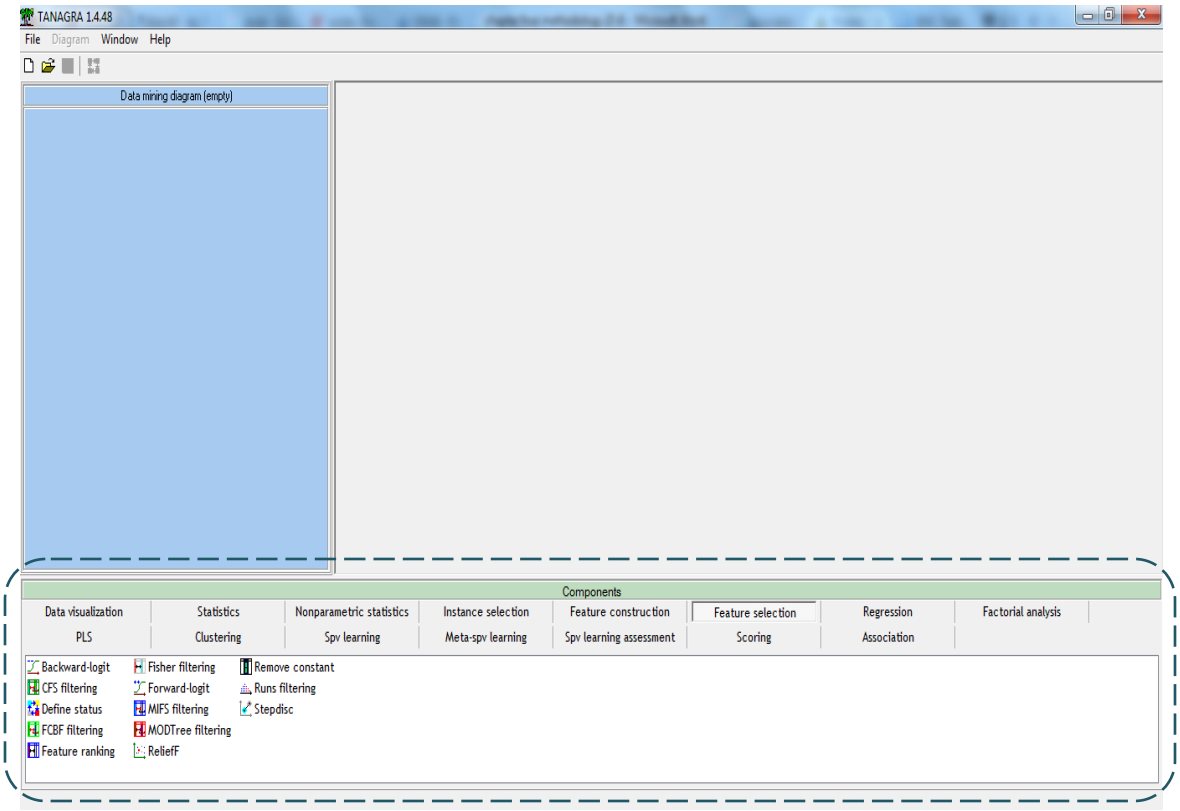


Figure 0-21: Tangara GUI

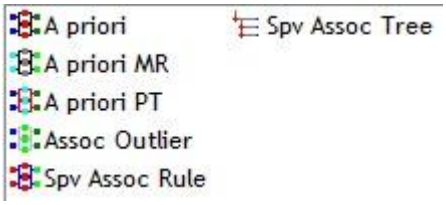


Figure 0-22: Associate category

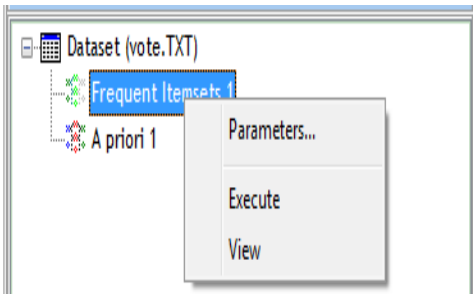


Figure 0-23: Frequent Itemset sub menu

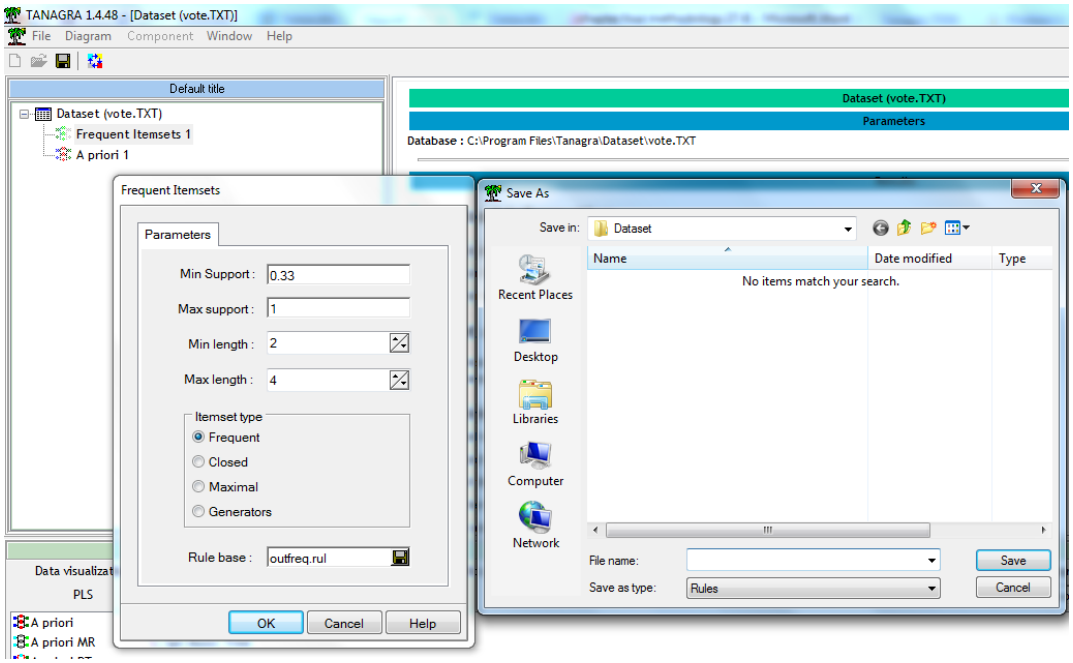


Figure 0-24: Frequent itemsets parameter settings

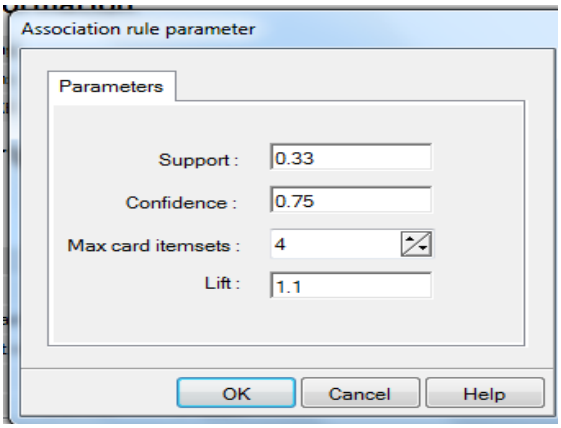


Figure 0-25: Association rule parameter settings

4.3.5 RapidMiner

RapidMiner provides an integrated environment for machine learning, data mining, text mining, predictive analysis and business analytics. It is used for business and industrial applications as well as for research, education, training, rapid prototyping and application development. RapidMiner is written in the Java programming language and provides a GUI to design and execute analytical framework. It supports all steps of data mining process, including results visualization, validation and optimization. RapidMiner provides schemes and models and algorithms from Weka and R scripts that can be used through extensions.

RapidMiner only provide support for FP-Growth algorithm for association rule mining. If the user wants to make use of Apriori, he/she has to add WEKA extension into RapidMiner. Weka provides an implementation of Apriori as shown in Figure 4.26.

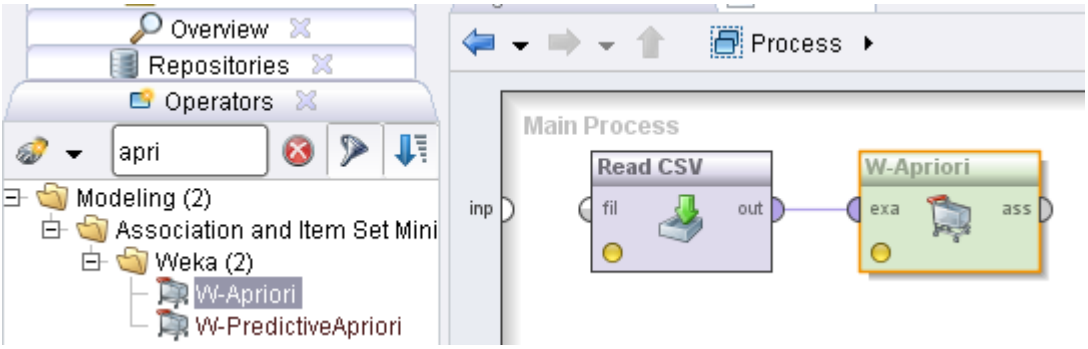


Figure 0-26: Weka-Apriori extension on RapidMiner

A summary of the existing platforms is briefly clarified in Table 4-3. Most of the investigated platforms adopt the Apriori algorithm and its variations. The open sourced tools are written in Java except Orange which is programmed in Python. The available tools can handle efficiently large databases since the speed of the processors is not a topic of concern anymore.

Table 0-3: Summary of association rules mining tools and platforms

Name	Description	AR Algorithms adopted	Comments	Comments	language
Arules	toolbox	Apriori, Eclat			C
Artool	toolbox for binary databases	Apriori, Eclat, FP Growth			Java
Weka	Data analytics platform	Apriori, Filtered associator, FP-Growth	Lift, leverage, conviction	Implement CAR	Java
Rapidminer	Data analytics platform	FP Growth!!	Weka and R plugin		Java
Orange	Data analytics platform	Standard Apriori			Python
KNIME	Data analytics platform	Apriori (CB)			Java
Tangara	Data analytics platform	Apriori	External save	Spv apriori	Java

4.4 Summary

Patterns are interesting when they are novel, useful, and non- trivial to compute. Knowledge is useful when it can help achieve a goal of the system or the user. Patterns completely unrelated to current goals are of little use and do not constitute knowledge within the given situation.

The tools presented in this chapter, shows diversity in results presentation, parameter's default value. The usability of analytical widgets vary from one to another. The way of representing the rules or frequent item sets is not easily understood by a novice or a non-expert user. The choice of the mining algorithm is a matter of decision made by the user, although the previously used algorithm will be the default for the next usage of the tool. Meanwhile the normal user might be a scientific researcher, who does not understand the difference between the methods and techniques of the underlying

algorithms, is not a programmer or computer scientist, he might get incorrect results. An Intelligent Framework might be helpful for scientific researcher on diversity of fields. The next chapter will describe the design issues of an Intelligent Association Rule Miner Framework (IARM), which provide acceptable and reliable results with both ease of use and accuracy.

CHAPTER FIVE

5 Intelligent Framework Design

5.1 Introduction

Different implementations of the same algorithm could behave completely different for different datasets. The behavior of several tools running the same algorithms on the same dataset may completely vary. Objective interestingness measures indicate the support and degree of the correlation of a pattern for a given dataset. However, they do not take into account the knowledge of the user who uses the data. In applications where the user has background knowledge, patterns ranked highly by objective measures may not be interesting.

A subjective interestingness measure takes into account both the data and the user's knowledge. Such a measure is appropriate when the background knowledge of users varies, the interests of the users vary, and the background knowledge of users evolve. Unlike the objective measures considered in chapter two, subjective measures may not be representable by simple mathematical formulas because the user's knowledge may be represented in various forms. Instead, they are usually incorporated into the mining process.

Association rules have been used with success in many domains. Most currently existing mining tools uses association rule algorithms with market basket analysis in mind. Such algorithms are inefficient because they mine many rules that are not relevant to a given user. Also, it is necessary to specify the minimum support of the mined rules in advance, often leading to either too many or too few rules which impacts the performance of the overall knowledge discovery process.

The main goal of this chapter is to find out the main implementation aspects of some of the data mining tools discussed in the previous chapter. A design of the proposed framework will be explained and its behavior with respect to the same types of

datasets. Testing of the new framework will be conducted on the Sudanese Kidney Transplantation Dataset.

5.2 Analysis of available tools (recommended settings and preferences)

Due to the overwhelming number of interestingness measures, the means of selecting an appropriate measure for a given application is an important issue. Machine learning typically can be divided into three phases, as described in figure 5-1:

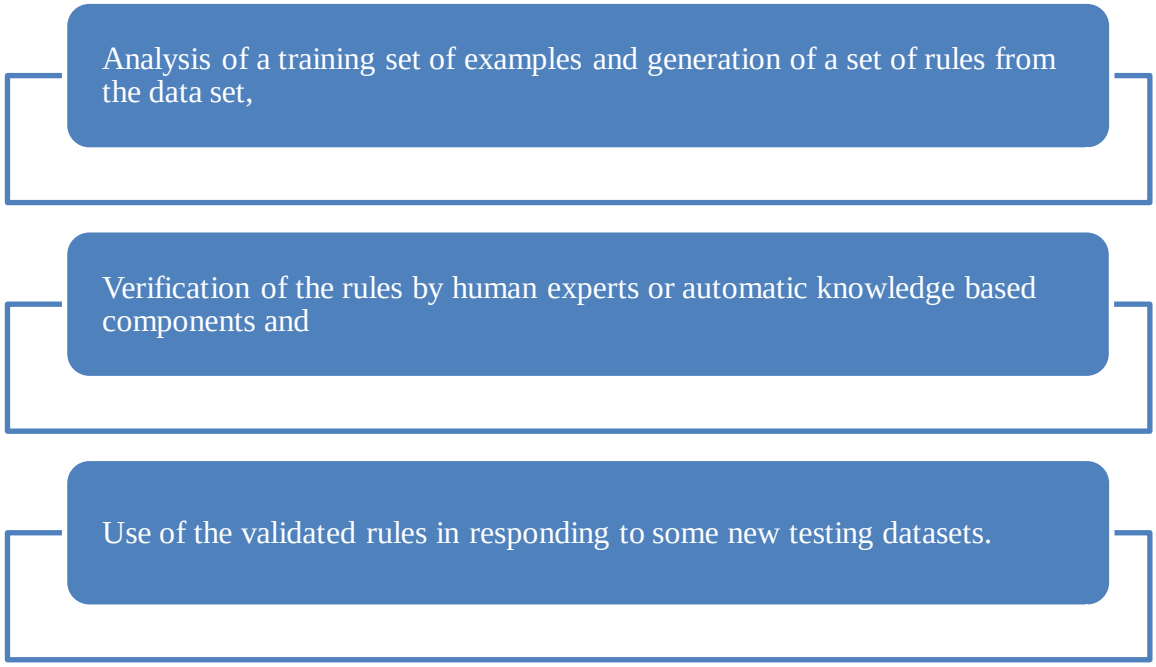


Figure 5-1: Machine learning phases

Following the research methodology adopted by the researcher, and implementing the three phases described above, the researcher carried a method to rank measures based on a specific dataset. In this method, the researcher has required first to rank a set of mined patterns, and the measure that has the most similar ranking results for these patterns is selected for further use. Experiments were performed on three data sets from different domains. The experiments implemented the Apriori algorithm on Weka and Orange. The researcher applied the default values in Weka for the minsup and

minconf which are 10% and 90% respectively. At the first, the minconf value was set fixed, setting a different value for the minsup, the researcher got different values for the size of the largest itemset size, the frequency of largest itemset and the number of association rules generated. The method then selects the measure that gives rankings most consistent with the manual ranking. This method is based on the three dataset as will be described in the next section.

5.2.1 First dataset (Breast Tissue dataset)

The Breast Tissue Dataset is obtained from the UCI repository. Dataset with electrical impedance measurements of freshly excised tissue samples from the breast(Anon 2010). A screenshot of the excel sheet containing the data is shown in the appendix. The dataset consists of 106 instances of patients suffering from breast cancer. It has 11 attributes, one is a class attribute. The data is numeric; it underwent a preprocessing step before applying the association rule mining algorithm.

The Apriori algorithm was run in the Weka Explorer, using only 10 attributes, excluding the class attribute. Tables 5-1 through 5-3 describe the behavior of the algorithm on different parameter settings.

According to the literature, the longer rules, having more number of itemsets, are usually more important than the rules that are shorter in size in the real applications, while rules with low support were regarded as false or trivial. Accordingly, the suitable settings are highlighted and will be used for the second iteration of the experiment. In Tables 5-2 (a) and 5-2(b), the fitting setting of the minconf was reset to 99% or 100%. With reference to the literature, the suitable setting for the minsup is ranging between 30%-50%, and the minconf is above 80%.

Table 5-1: Frequent patterns and association rules(fixed minconf=90%)

Min-conf=90%			
min-sup	largest itemset size	# largest itemset	#of rules generated
0.1	7	2	>1000
0.15	6	3	513
0.17 – 0.3	6	1	>200
0.4	5	1	97
0.42	4	3	57
0.43	4	1	39
0.44	**4**	**1**	**39**
0.45	**3**	**4**	**18**
0.5	2	3	3

Table 5-2(a):Frequent patterns and association rules(fixed minsup=44%)

Min-sup= 44%			
min-conf	largest itemset size	# largest itemset	#of rules generated
0.8	4	1	60
0.9	4	1	39
0.91	4	1	36
0.92	4	1	33
0.93	4	1	21
0.95-0.98	4	1	18
0.99-1.0	**4**	**1**	**15**

Table 5-2(b): Frequent patterns and association rules(fixed minsup=45%)

Min-sup= 45%			
min-conf	largest itemset size	# largest itemset	#of rules generated
0.8	3	4	33
0.85	3	4	27
0.9, 0.91	3	4	18
0.92	3	4	15
0.93-0.98	3	4	12
0.99-1.0	**3**	**4**	**9**

The same parameter setting was applied in the Orange Associate. The resulted rules was the same as the Weka's as illustrated in Figures 5-2 and 5-3.

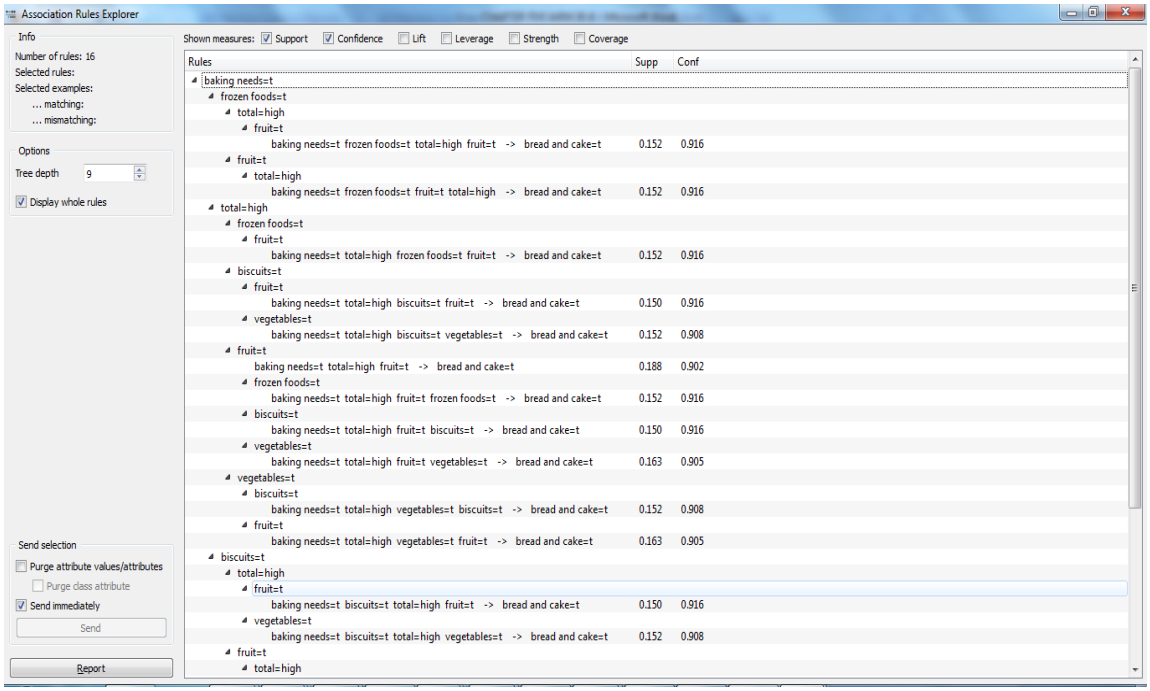


Figure 5-2: Generated rules in Orange

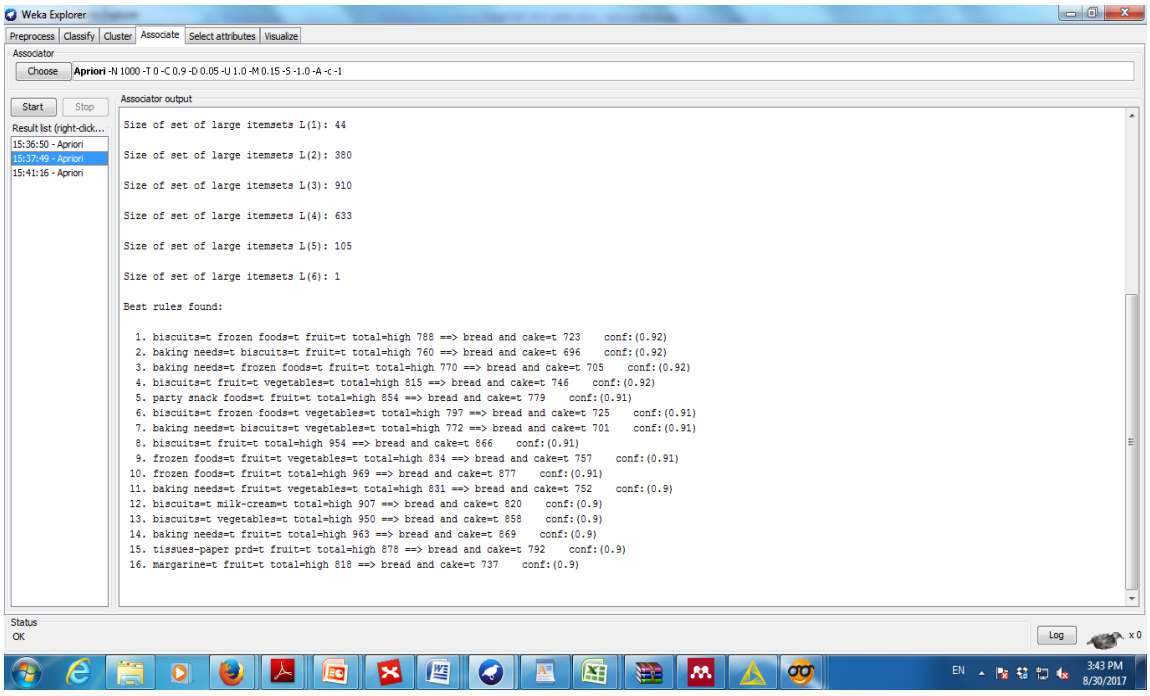


Figure 5-3: Generated rules in Weka

When setting classified association rules(CAR) as true on the whole set of the attributes, i,e, 11 attributes, no results were obtained for all values of min support until reaching 1% and less, while the confidence value dropped to 10% as shown in Table 5-3. The number of rules generated when proportioned to the number of instances is questionable.

Table 5-3: Classified association rules mining

Inducing classification association rules (CAR)				
min-sup	min-conf	largest itemset size	# largest itemset	#of rules generated
0.01-0.18	0.1	9	1	444
>0.019	0.1	-	-	0
0.018	0.2	9	1	256
0.018	0.3	9	1	64
0.018	0.4	9	1	56
0.018	0.5	9	1	0

5.2.2 Second dataset(Supermarket)

Supermarket dataset is a dataset from the Weka repository of datasets consisting of 4627 transactions and 217 items. The same procedure applied to the first dataset will be followed. Starting at first with default values and the good results are highlighted in Tables 5-4 and 5-5. The same results were obtained when using Orange or Weka tools as shown in Figures 5-5 and 5-6. Table 5-6 shows the results after inducing the classification association rules.

Table 5-4: FP and AR for Supermarket data (fixed mincon= 90%)

Min-conf=90%			
min-sup	largest itemset size	# largest itemset size	#of rules generated
0.1	7	20	>1000
0.13	7	20	>1000
0.14	6	4	37
0.15	6	1	16
0.16	5	49	10
0.17	5	20	6
0.2	5	2	0

Table 5-5: FP and AR of Supermarket dataset (fixed minsup= 15%)

Min-sup= 15%			
min-conf	largest itemset size	# largest itemset size	#of rules generated
0.85	6	1	>1000
0.87	6	1	>1000
0.89	6	1	42
0.9	6	1	16
0.91	6	1	5
> 0.92	6	1	0

Table 5-6: FP and AR of Supermarket dataset with different minsup and minconf

min-sup	min-conf	largest itemset size	# largest itemset	#of rules generated
0.2	0.85	5	2	42
0.3	0.85	3	20	0
0.25	0.85	4	6	3
0.25	0.8	4	6	53

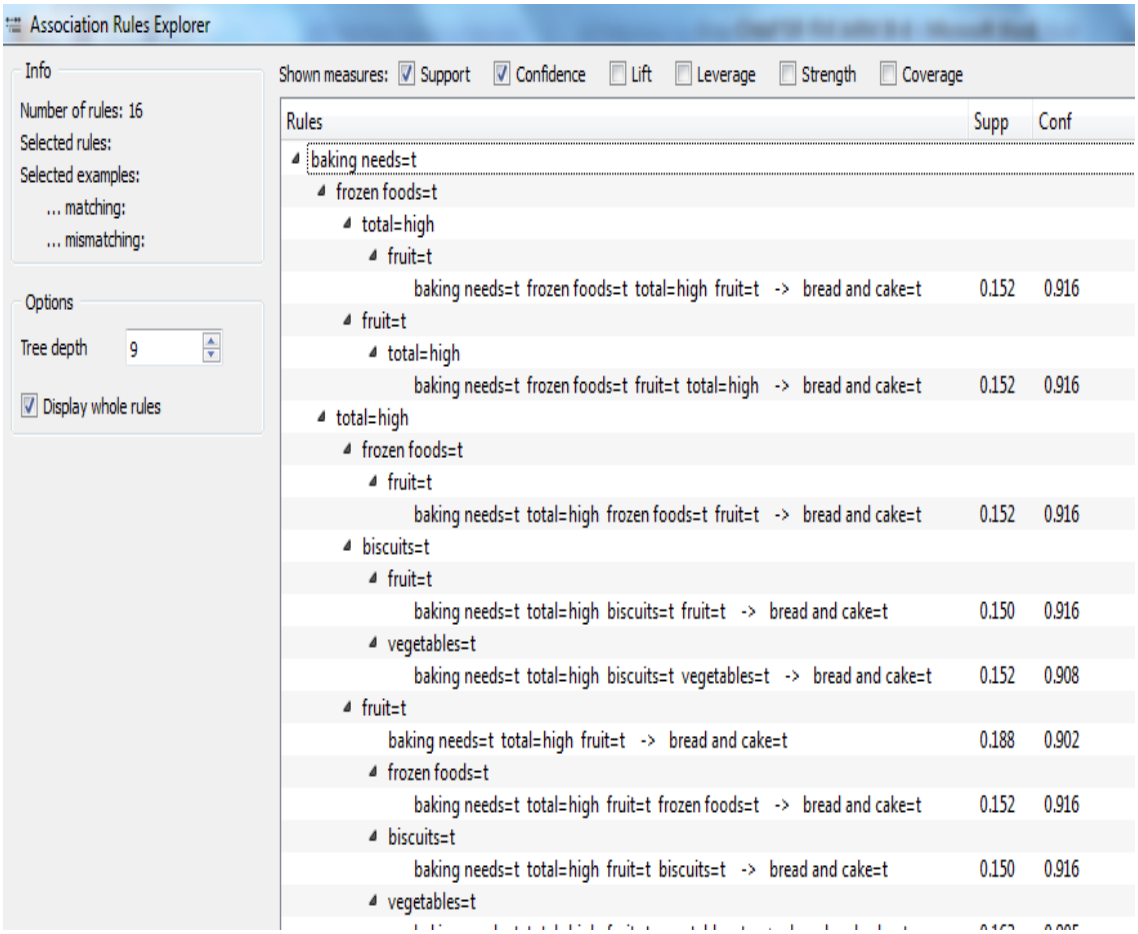


Figure 5-4: Orange's Association rules explorer

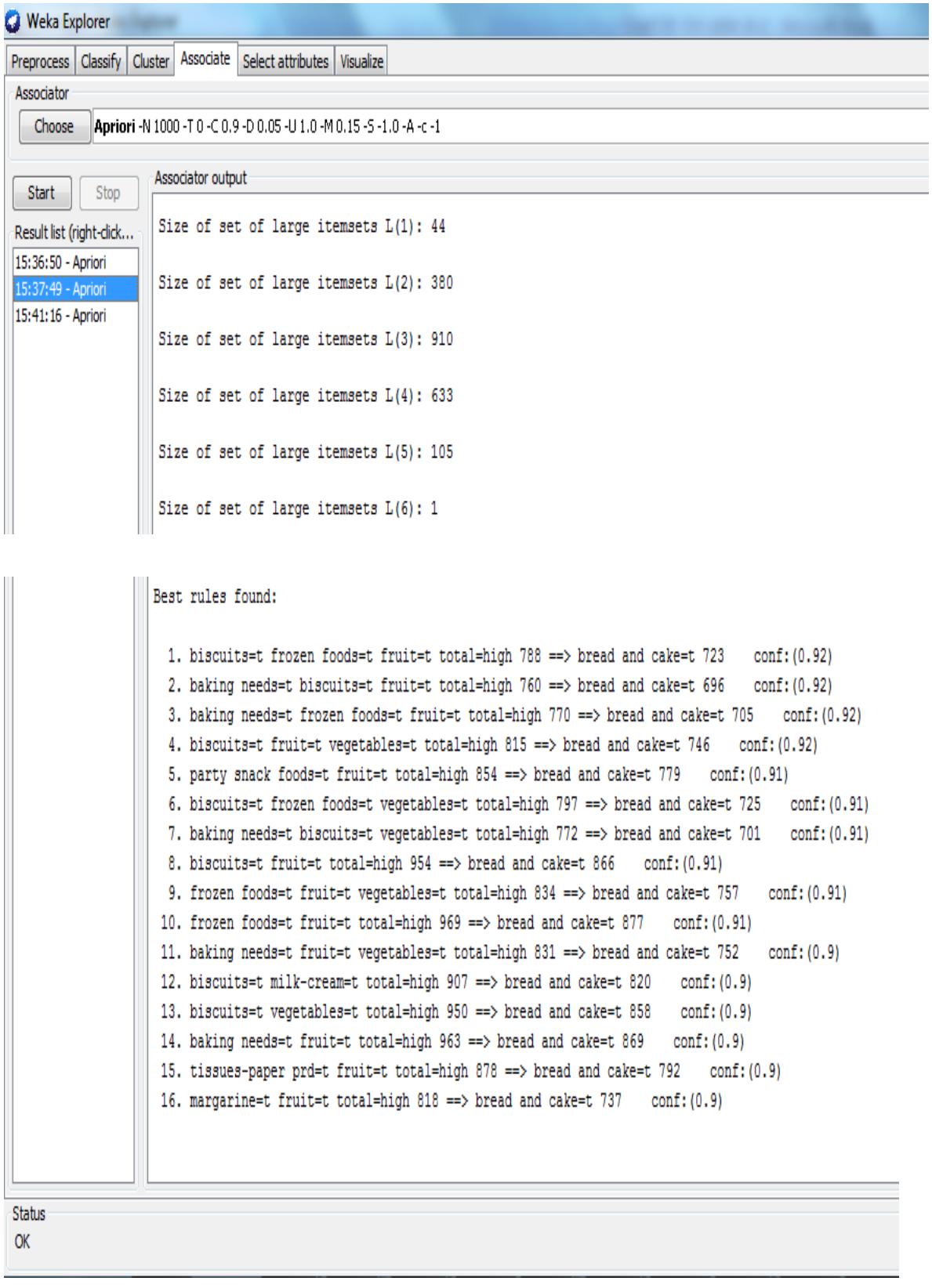


Figure 5-5: Weka's Association rules results window

Table 5-7: FP and AR of Supermarket dataset (CAR)

Inducing classification association rules (CAR)				
min-sup	min-conf	largest itemset size	# largest itemset	#of rules generated
0.25	0.85	2	4	0
0.15	0.9	4	11	0
0.13	0.7	5	3	22
0.25	0.5	2	4	15
0.15	0.6	4	11	88

5.2.3 Third dataset(Facebook Metrics)

The data is related to posts' published during the year of 2014 on the Facebook's page of a renowned cosmetics brand(Moro et al. 2016). The dataset is in excel sheet as illustrated in the figure A-2 in the appendix. It consists of 500 pages summary with 19 attributes. Applying the same method resulted in the suitable parameter setting which is clarified in table 5-6. The setting of the minimum support was surprising and contradict the settings of the previous two dataset. No rules were generated when using classification rules.

Table 5-8: FPs and ARs of Facebook Metrics(fixed minconf=90%)

Min-conf=90%			
min-sup	largest itemset size	# largest itemset size	#of rules generated
0.1 -0.9	4	2	10
0.93	3	2	10
0.94	2	3	6
0.95	2	2	4
0.96	**2**	**1**	**2**
0.97	-	-	-

Table 5-9: FPs and ARs of Facebook Metrics(fixed minsup=90%)

min-conf	largest itemset size	# largest itemset size	#of rules generated
0.1-0.9	**4**	**2**	**10**

Table 5-10: FPs and ARs of Facebook Metrics using CAR

Inducing classification association rules (CAR)				
min-sup	min-conf	largest itemset size	# largest itemset	#of rules generated
0.01	0.1-0.9	13	2	0

5.3 Analysis of the experimental results

What constitute the discovered knowledge are its form, its representation, and its degree of certainty. In the context of association rules mining, the user may need simple and certain rules which are represented in a form appropriate for the intended user.

Representing and conveying the degree of certainty is essential to determining how much faith the system or user should put into a discovery. As examined, certainty involves several factors, including the integrity of the data; the size of the sample on which the discovery was performed; and, possibly, the degree of support from available domain knowledge. Without sufficient certainty, patterns become unjustified and, thus, fail to be knowledge.

The following observations were drawn after performing the experiment above:

- The effect of parameter setting on the number of association rules generated
- Decision upon the minimum threshold values is the key issue
- Relying on the mining tool, and using its default parameter settings may result in either false rules or too many rules.

- Every domain has its own features and requires special settings
- The number of mined rules should be of reasonable length with regard to the type of datset.
- In some domains, the researcher may be interested in rules of preferred length or a special item to be in the consequent or the antecedent of the rule.

A discovery system must be capable of deciding which calculations to perform and whether the results are interesting enough to constitute knowledge in the current context. Another way of viewing this notion of non triviality is that a discovery system must possess some degree of freedom in processing the data and evaluating its results.

The use of domain knowledge taken to the extreme will produce a specialized algorithm that will outperform any general method in its domain but will not be useful outside it. A desirable compromise is to develop a framework for augmenting the general method with the specific domain knowledge. Efficient algorithms will be crucial. Incremental methods are needed to efficiently keep pace with changes in data. Interactive systems will provide, perhaps, the best opportunity for discovery: Use human judgment but rely on the machine to do search.

5.4 Intelligent Association Rules Mining Framework

Knowledge discovery in databases exhibits four main characteristics as described in (Frawley & Piatetsky-Shapiro 1992):

- Accurately expressed by the fittest measures of certainty.
- Results are interesting according to user-defined biases.
- Running times for large-sized databases are predictable and acceptable.
- Expressed in high-level language that is understandable by any users

The interesting rules were often found in an area of intermediate support sizes. Sometimes low support generate noisy, frequent itemsets and trivial (uninteresting) while in some domains and depending on the size of the data base, rules could not be

found until reaching very low support threshold. Well-known rules have very high value of support and confidence, but this also rely mainly on the user preference and needs some domain knowledge.

As hundreds of thousands of frequent itemsets may exist in a sizeable data set, the user can have the application limit the search space of candidate rules to those that could be generated from itemsets of a specified size (e.g. all itemsets with more than seven items) or which include at least one item within a specified set of items. To further limit the search space of candidate rules, the application looks only for rules where either the LHS or the RHS sets of the rule LHS⇒RHS contain only one item as may be favoured in genetic data.

An intelligent interactive framework is designed to mine association rules and frequent patterns based on Apriori algorithm. Many algorithms will be applied while the user has no intention. In the proposed framework, the user does not require the minimum support to be specified in advance. Rather, a target range is given for the number of rules, and the algorithm adjusts the minimum support for each user in order to obtain a ruleset whose size is in the desired range. Rules are mined for a specific target user, reducing the time required for the mining process.

The general structure of the developed software depends on finding frequent itemsets using candidate generation (Apriori) algorithm. The input data can be read from text or comma delimited files and can be extendable for large data bases. The minimum threshold should suitably be set to meet the need of different users. After the conducted experiments and domain knowledge searched out in previous usage of the algorithm in the literature, an array is constructed containing the suitable parameter settings for each domain of knowledge explored so far. The support count can be also determined as different values for having the feasibility to be adaptable to different data bases. Subjective measures are included depending on prior domain knowledge or further investigation. A mapping of terms is also available to make the interface more friendly and accepted by many users who do not need to know computer jargons.

At the core of the system is the discovery method, which computes and evaluates patterns. Three algorithms are considered; the classical Apriori, a variation which deals with rare items, and another algorithms which deals with biological data. Figures 5-6 to 5-9 represent the general framework. The output is discovered knowledge that can be directed to the user. Interactivity is perceived through user intervention when asked to specify the size of the database and his/her number of patterns desired in the discovered rules. The default values are adjusted by the system and the mining results are displayed and/or saved. If the results obtained are not the desired, the user is asked to adjust the threshold values. The algorithm that deals with rare items cannot be invoked unless it is desired by the system user. Classified association rules mining is not a choice on any domain, although previously proved of interest.

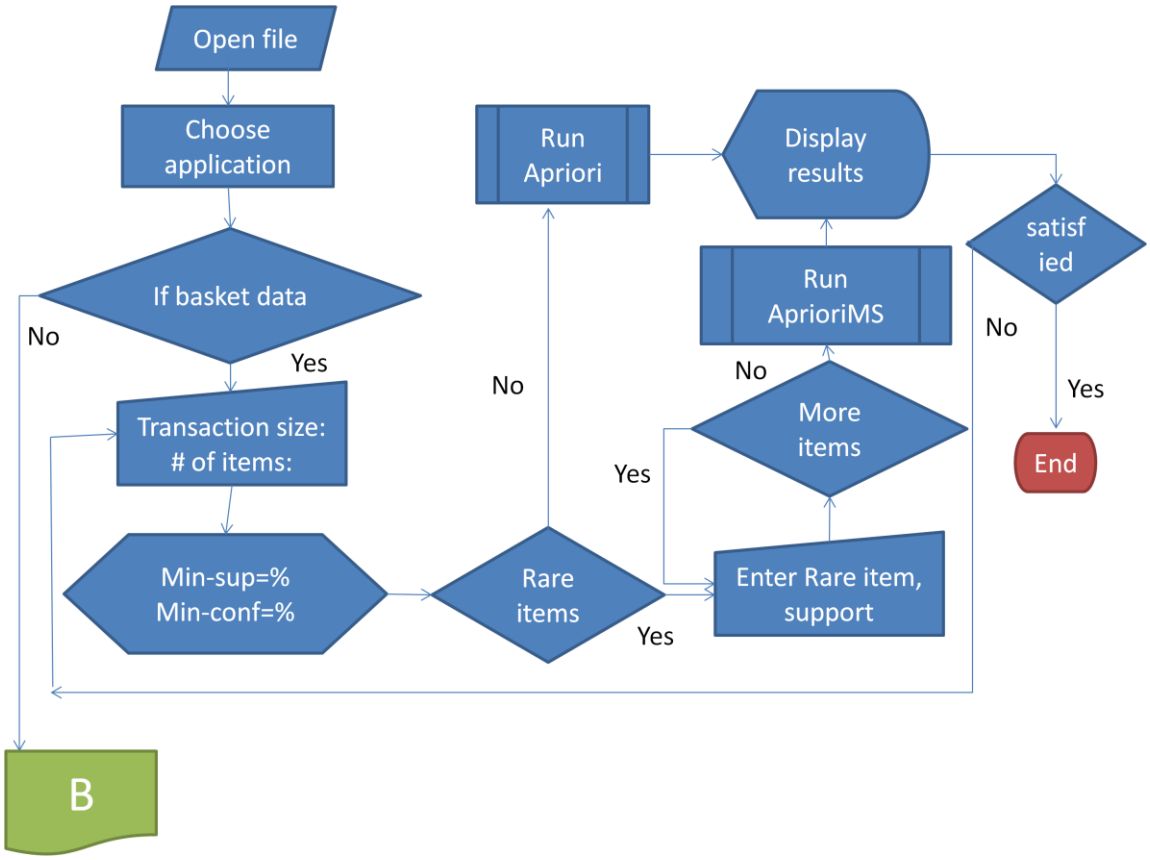


Figure 5-6: Mining market basket data

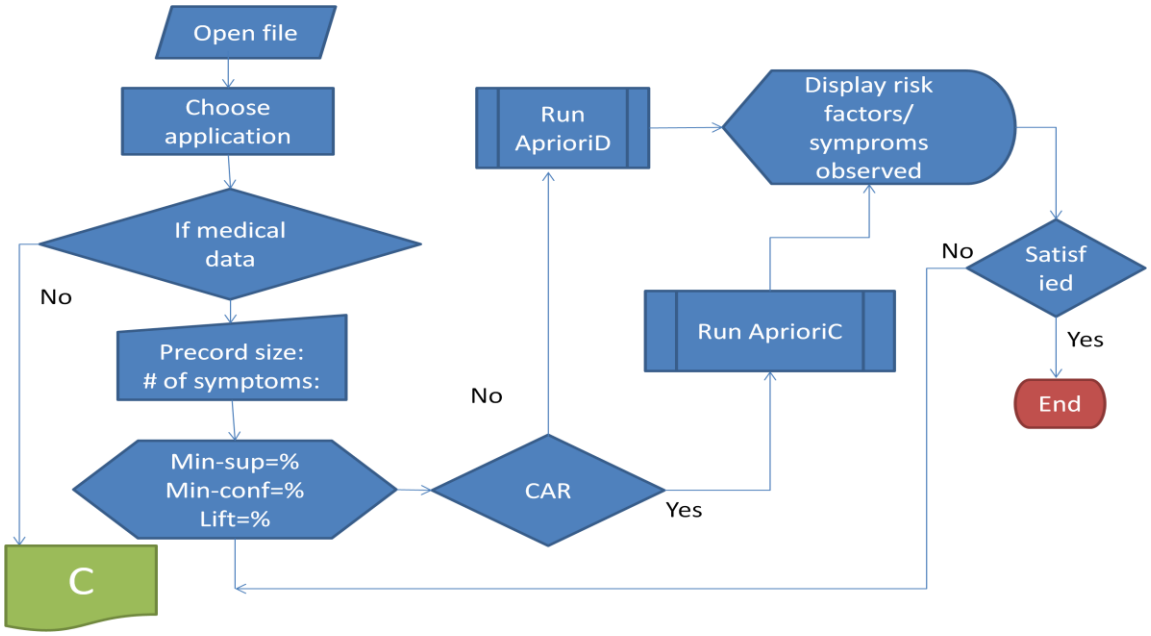


Figure 5-7: Mining medical records

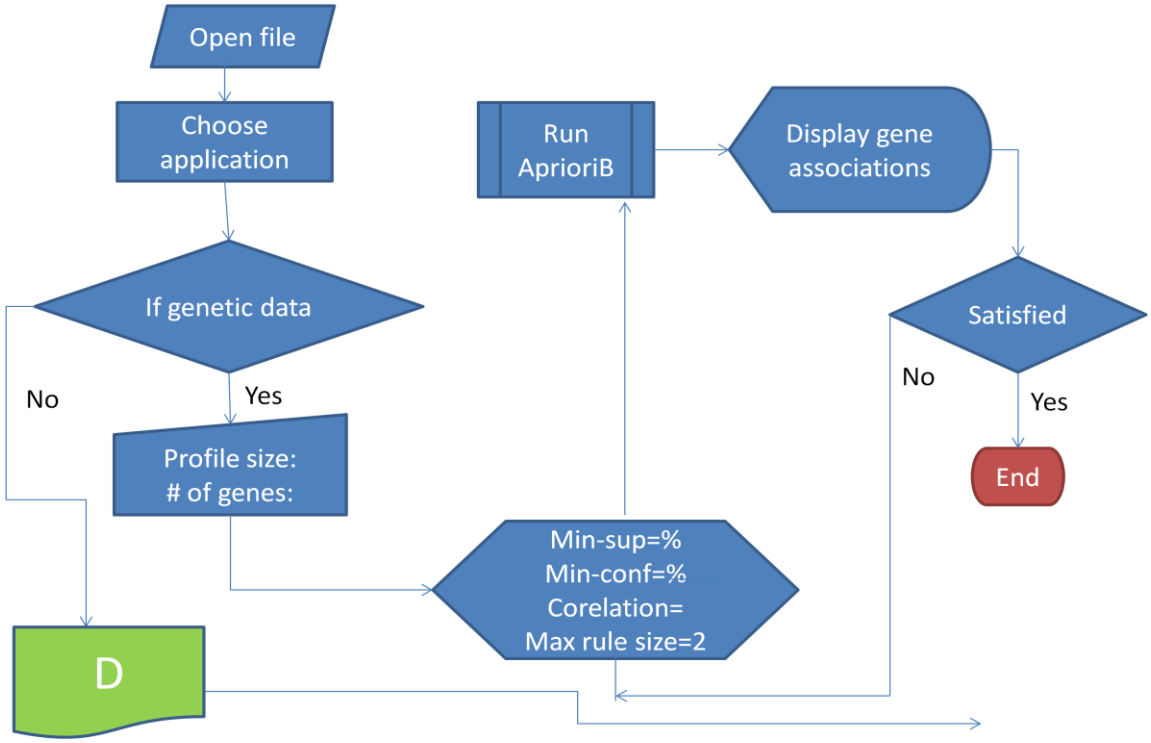


Figure 5-8: Mining Biological data

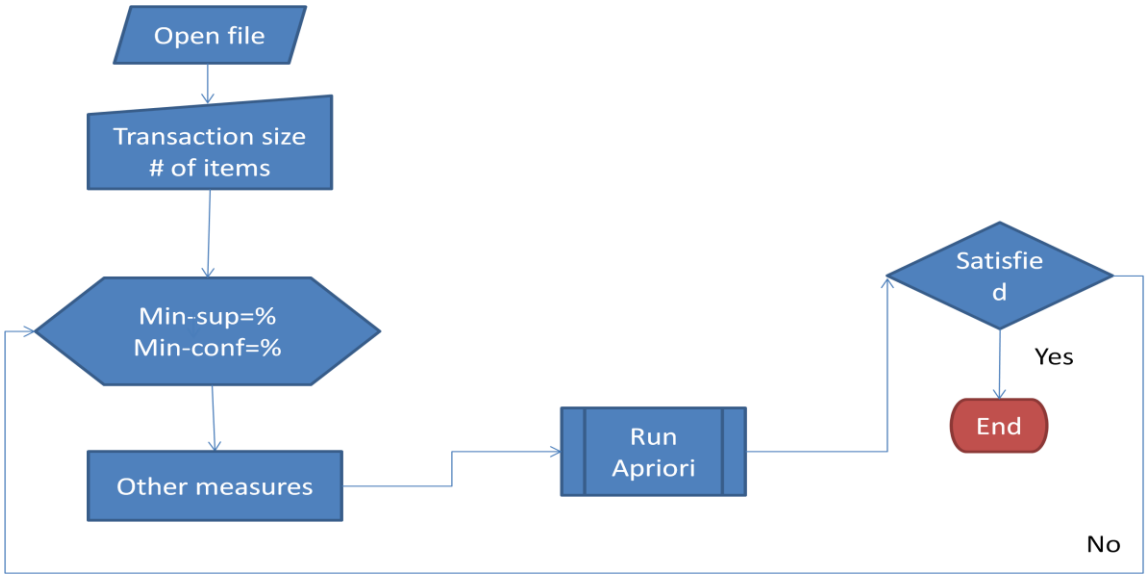
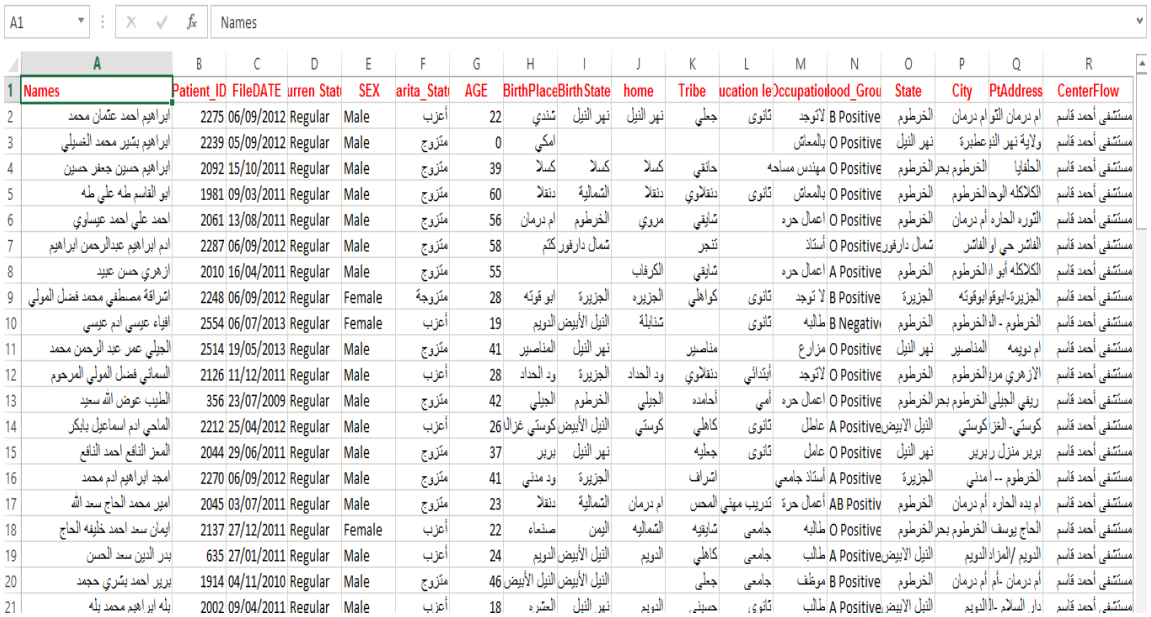


Figure 5-9: Mining general types of datasets

5.4.1 Framework Evaluation

Association rules are used in the medical domain, where data sets are generally high dimensional and small. The chief disadvantage about mining association rules in a high-dimensional data set is the huge number of patterns that are discovered, most of which are irrelevant or redundant. Several constraints are proposed for filtering purposes.

A dataset was constructed for the Sudanese Kidney Transplanted Association which keeps records about the patients who transplanted a new kidney. Thier database contains basic information and follow-up data during the patient visit after transplantation. 1,116 records were collected for the years 2009-2014, but after processing and removing missing value, the whole set turned to 326 records with 19 attributes. Figure 5-10 shows the excel sheet that contains the patient's record.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	Names	Patient_ID	FileDATE	urren Stat	SEX	arita_Stat	AGE	BirthPlace	BirthState	home	Tribe	ucation le	Occupatio	lood_Grou	State	City	PtAddress	CenterFlow
2	ابراهيم احمد عثمان محمد	2275	06/09/2012	Regular	Male	أعزب	22	شندي	نهر النيل	نهر النيل	جعلي	ثانوى	لاتوجد	B Positive	الخرطوم	ام درمان	ام درمان التوا	مستشفى أحمد قاسم
3	ابراهيم بشير محمد الفضيلي	2239	05/09/2012	Regular	Male	متزوج	0	امكي					بالمعاش	O Positive	نهر النيل	نهر النيل	ولاية نهر النو	مستشفى أحمد قاسم
4	ابراهيم حسين جعفر حسين	2092	15/10/2011	Regular	Male	متزوج	39	كسلا	كسلا	كسلا	حالقي		مهندس مساحه	O Positive	الخرطوم	الخرطوم بحر	الحلفايا	مستشفى أحمد قاسم
5	ابو القاسم طه علي طه	1981	09/03/2011	Regular	Male	متزوج	60	دنقلا	الشمالية	دنقلا	دنقلاوي	ثانوى	بالمعاش	O Positive	الخرطوم	الخرطوم	الكلاكله الوحد	مستشفى أحمد قاسم
6	احمد علي احمد عيساوي	2061	13/08/2011	Regular	Male	متزوج	56	ام درمان	الخرطوم	مروي	شايقي		اعمال حره	O Positive	الخرطوم	ام درمان	الثوره الحاره	مستشفى أحمد قاسم
7	ادم ابراهيم عبدالرحمن ابراهيم	2287	06/09/2012	Regular	Male	متزوج	58	كتم	شمال دارفور		تنجر		أستاذ	O Positive	شمال دارفور	الفاشر	الفاشر حي او	مستشفى أحمد قاسم
8	ازهري حسن عبيد	2010	16/04/2011	Regular	Male	متزوج	55			الكرقاب	شايقي		اعمال حره	A Positive	الخرطوم	الخرطوم	الكلاكله أبو ا	مستشفى أحمد قاسم
9	اشراقة مصطفي محمد فضل المولي	2248	06/09/2012	Regular	Female	متزوجة	28	ابو قوته	الجزيرة	الجزيره	كواهلي	ثانوى	لا توجد	B Positive	الجزيرة	الجزيرة	الجزيرة-ابوقو	مستشفى أحمد قاسم
10	افياء عيسي ادم عيسي	2554	06/07/2013	Regular	Female	أعزب	19	الدويم	النيل الأبيض	شنابلة		ثانوى	طالبه	B Negativ	الخرطوم	الخرطوم	الخرطوم - ال	مستشفى أحمد قاسم
11	الجيلي عمر عبد الرحمن محمد	2514	19/05/2013	Regular	Male	متزوج	41	المناصير	نهر النيل		مناصير		مزارع	O Positive	نهر النيل	نهر النيل	ام دويمه	مستشفى أحمد قاسم
12	السماني فضل المولي المرحوم	2126	11/12/2011	Regular	Male	أعزب	28	ود الحداد	الجزيرة	ود الحداد	دنقلاوي	أبتدائي	لاتوجد	O Positive	الخرطوم	الخرطوم	الازهري مرد	مستشفى أحمد قاسم
13	الطيب عوض الله سعيد	356	23/07/2009	Regular	Male	متزوج	42	الجيلي	الخرطوم	الجيلي	أحامده	أمي	اعمال حره	O Positive	الخرطوم	الخرطوم بحر	ريفي الجيلي	مستشفى أحمد قاسم
14	الماحي ادم اسماعيل بابكر	2212	25/04/2012	Regular	Male	أعزب	26	غزال	النيل الأبيض	كوستي	كاهلي	ثانوى	عاطل	A Positive	النيل الابيض	النيل الابيض	كوستي- الغز	مستشفى أحمد قاسم
15	المعز النافع احمد النافع	2044	29/06/2011	Regular	Male	متزوج	37	بربر	نهر النيل		جعليه	ثانوى	عامل	O Positive	نهر النيل	نهر النيل	بربر منزل ر	مستشفى أحمد قاسم
16	امجد ابراهيم ادم محمد	2270	06/09/2012	Regular	Male	متزوج	41	ود مدني	الجزيرة		اشراف		أستاذ جامعي	A Positive	الجزيرة	امدني	الخرطوم -- ا	مستشفى أحمد قاسم
17	امير محمد الحاج سعد الله	2045	03/07/2011	Regular	Male	متزوج	23	دنقلا	الشمالية	ام درمان	تدريب مهني المحس		أعمال حرة	AB Positiv	الخرطوم	ام درمان	ام بده الحاره	مستشفى أحمد قاسم
18	ايمان سعد احمد خليفه الحاج	2137	27/12/2011	Regular	Female	أعزب	22	صنعاء	اليمن	الشماليه	شايقيه	جامعي	طالبه	O Positive	الخرطوم	الخرطوم بحر	الحاج يوسف	مستشفى أحمد قاسم
19	بدر الدين سعد الحسن	635	27/01/2011	Regular	Male	أعزب	24	الدويم	النيل الأبيض	الدويم	كاهلي	جامعي	طالب	A Positive	النيل الابيض	النيل الابيض	الدويم /المزاد	مستشفى أحمد قاسم
20	بربر احمد بشري حجمد	1914	04/11/2010	Regular	Male	متزوج	46	النيل الأبيض	النيل الأبيض		جعلي	جامعي	موظف	B Positive	الخرطوم	ام درمان	ام درمان -أم	مستشفى أحمد قاسم
21	بله ابراهيم محمد بله	2002	09/04/2011	Regular	Male	أعزب	18	العشره	نهر النيل	الدويم	حسينيه	ثانوى	طالب	A Positive	النيل الابيض	النيل الابيض	دار السلام -ال	مستشفى أحمد قاسم

Figure 5-10: Sudanese Kidney Transplantation dataset

Table 5-9 illustrate the 19 rules extracted with the minsup and minconf threshold value of 80% and 90% respectively. This settings are within the range specified in (Wright et al. 2010; Tilman B. Drüeke, M.D., Francesco Locatelli, M.D., Naomi Clyne, M.D., Kai-Uwe Eckardt et al. 2006; M. J. Huang et al. 2007). Based on the domain expert opinion, minimum confidence was 70%. From the medical point of view, rules with 90% or higher confidence are preferable, but they are infrequent and generally do not involve all risk factors. Rules in the 80–90% range may be acceptable depending on which measurements are involved. Rules in the 70–80% range have borderline reliability and they indicate potential patterns to be discovered in subsets of patients. Rules with confidence lower than 70% are not medically reliable; in particular, rules with confidence 50% or lower are never considered interesting(Ordonez et al. 2005).

Table 5-9: Association rules of kidney transplantation society

Consequence	Antecedent			min sup	min conf
1. no genetic causes	Creatinine normal			81	95
2. Uric acid normal	Creatinine normal			81	96
3. no genetic causes	Urea normal	Uric acid normal		81	97

4. Alive	Urea normal	no genetic causes		82	96
5. no genetic causes	Urea normal	Alive		82	96
6. Alive	Urea normal	Uric acid normal		81	97
7. Uric acid normal	Urea normal	Alive		82	96
8. Uric acid normal	Platelets Normal	no genetic causes		85	96
9. no genetic causes	Platelets Normal	Uric acid normal		85	96
10. Alive	Platelets Normal	no genetic causes		85	96
11. no genetic causes	Platelets Normal	Alive		85	96
12. Alive	Platelets Normal	Uric acid normal		85	96
13. Uric acid normal	Platelets Normal	Alive		85	96
14. Alive	no genetic causes	Uric acid normal		92	97
15. Uric acid normal	no genetic causes	Alive		92	97
16. no genetic causes	Uric acid normal	Alive		93	96
17. Alive	Platelets Normal	no genetic causes	Uric acid normal	82	97
18. Uric acid normal	Platelets Normal	no genetic causes	Alive	82	97
19. no genetic causes	Platelets Normal	Uric acid normal	Alive	82	97

5.5 Summary

The efficient discovery algorithm is the core of the mining process. Many applications and programs were written for the sake of mining frequent patterns and association rules. The extracted rules should express novel, useful and non trivial knowledge. Each of these programs was specially developed for a certain domain of knowledge and proves efficiency. Parameter settings differs from one developed program to another. Since no one application can be beneficial to all areas of research, the proposed intelligent frame work will open wide range of future research.

Parameter setting for three different datasets was tested and evaluated with domain knowledge whaich was gotten from the working researchers in each domain. The IARMF can easily be understood by any user, the novice and expert users. It successfully drawn interesting knowledge or rules when applied to the constructed Sudanes dataset.

CHAPTER SIX

6.1 Conclusion

Discovery algorithms are procedures designed to extract knowledge from data. This activity involves two processes: identifying interesting patterns and describing them in a concise and meaningful manner. Efficient algorithms will be crucial. Incremental methods are needed to efficiently keep pace with changes in data. Interactive systems will provide, perhaps, the best opportunity for discovery.

Association rules mining has been the backbone for knowledge discovery for decades. Building a classification model can be done through features which are considered frequent in a certain data base or through use of association. The core to extract interesting association rules is selecting the right algorithm for a specific dataset. The databases can be found at many lengths and dimensions. Practioners on the field cannot neglect the effect of parameter setting on the number of association rules generated. The size of the database along with interestingness measure produces what can be called useful knowledge. So, decision upon the minimum threshold value is the key issue. The longer rules are usually more important than the rules that are shorter in size in the real applications but this is not always true when judged with subjective measures. Some time and in specific domains short rules are desirable but with specific patterns in the left or the right hand side of the rule.

A new intelligent framework was introduced based on well known and mostly used algorithms. The new frame work gets information from the user of the system about the type and the size of the dataset he/she wants to explore, and then execute an algorithm with pre-specified setting that match his preferences. Rules are mined for a specific target user, reducing the time required for the mining process. Experimental evaluation of a system based on our algorithm reveals performance that is significantly better than that of traditional correlation-based approaches.

6.2 Recommendation and future work

- Extend the framework for processing of other data types
- Add explanation modules
- Extend the modules for other applicable applications
- Extending the framework by adding new modules for other domains
- If the user of the system changes the parameter settings, the changes should be saved. To build a log for every execution of the different algorithms,
    - Building a recommender system that can learn from previous runs of the system.
- Classify the algorithms which might suits best each type of dataset.

References

Agrawal, R., Imielinski, T. & Swami, A., 1993. Database mining: A performance perspective. *Knowledge and Data ...*, pp.1–22. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=250074 [Accessed March 5, 2013].

Agrawal, R., Imieliński, T. & Swami, A., 1993. Mining association rules between sets of items in large databases P. Buneman & S. Jajodia, eds. *Proceedings of the 1993 ACM SIGMOD international conference on Management of data SIGMOD 93*, 22(May), pp.207–216. Available at: http://dl.acm.org/citation.cfm?id=170072 [Accessed May 27, 2012].

Agrawal, R. & Srikant, R., 1994. Fast algorithms for mining association rules. *Proc. 20th Int. Conf. Very Large Data Bases,*, pp.1–13. Available at: http://www.cs.cmu.edu/afs/cs/Web/People/ngm/15-721/summaries/12.pdf [Accessed July 4, 2012].

Almodaifer, G., Hafez, A. & Mathkour, H., 2011. Discovering medical association rules from medical datasets. *ITME 2011 - Proceedings: 2011 IEEE International Symposium on IT in Medicine and Education (2011)*, 2, pp.43–47. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6132053 [Accessed November 30, 2014].

Alves, R., Rodriguez-Baena, D.S. & Aguilar-Ruiz, J.S., 2010. Gene association analysis: a survey of frequent pattern mining from gene expression data. *Briefings in bioinformatics*, 11(2), pp.210–24. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19815645.

Anon, 2010. Breast Tissue Data Set. Available at: https://archive.ics.uci.edu/ml/datasets/Breast+Tissue.

Ashoka Savasere, Edward Omiecinski, S.B.N. et al., 1995. An efficient algorithm for

mining association rules in large databases. *Proc 1995 Int Conf Very Large Data Bases VLDB95*, 5(1), pp.432–444. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2912889&tool=pmcentrez&rendertype=abstract [Accessed August 6, 2012].

Becerra, D. & Vanegas, D., 2009. An association rule based approach for biological sequence feature classification. *... , 2009. CEC'09. IEEE ...*, pp.3111–3118. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4983337 [Accessed March 14, 2013].

Besemann, C., Denton, A. & Yekkirala, A., 2004. Differential Association Rule Mining for the Study of Protein-Protein Interaction Networks. *Bioinformatics*, pp.72–80.

Borgelt, C. & Kruse, R., 2002. Induction of Association Rules: Apriori Implementation. *15th Conference on Computational Statistics Physica Verlag, Heidelberg, Germany 2002*, 1, pp.1–6.

Brin, S. et al., 1997. Dynamic itemset counting and implication rules for market basket data. In *Proceedings of the 1997 ACM SIGMOD international conference on Management of data - SIGMOD '97*. New York, New York, USA: ACM Press, pp. 255–264. Available at: http://portal.acm.org/citation.cfm?doid=253260.253325.

Cai, C.H. et al., 1998. Mining Association Rules with Weighted Items. *International Database Engineering and Applications Symposium*, pp.68–77.

Chen, S.-C. & Bahar, I., 2004. Mining frequent patterns in protein structures: a study of protease families. *Bioinformatics (Oxford, England)*, 20 Suppl 1, pp.i77-85. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1201446&tool=pmcentrez&rendertype=abstract [Accessed May 13, 2012].

Creighton, C., 2003. Mining gene expression databases for association rules. *Bioinformatics*, pp.79–86. Available at: http://bioinformatics.oxfordjournals.org/content/19/1/79.short [Accessed July 11,

2012].

Cristofor, L., 2006. ARMiner Project. Available at: https://www.cs.umb.edu/~laur/ARMiner/.

Delen, D., Walker, G. & Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2), pp.113–27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15894176 [Accessed March 4, 2012].

Facca, F.M. & Lanzi, P.L., 2005. Mining interesting knowledge from weblogs: A survey. *Data and Knowledge Engineering*, 53(3), pp.225–241.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth Padharic, 1996. From data mining to knowledge discovery in databases. *Advances in Knowledge Discovery and Data Mining*, 17(3), pp.1–36.

Frawley, W. & Piatetsky-Shapiro, G., 1992. Knowledge discovery in databases: An overview. *AI magazine*.

Geng, L. & Hamilton, H.J., 2006. Interestingness measures for data mining. *ACM Computing Surveys*, 38(3), p.9–es. Available at: http://portal.acm.org/citation.cfm?doid=1132960.1132963 [Accessed July 13, 2012].

Georgii, E. et al., 2005. Analyzing microarray data using quantitative association rules. *Bioinformatics*, 21(Suppl 2), p.ii123-ii129. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bti1121 [Accessed May 6, 2012].

Goethals, B., 2003. Frequent itemset mining implementations repository. , p.http://fimi.ua.ac.be/. Available at: http://fimi.ua.ac.be/.

Gong, P. et al., 2007. The application of improved association rules data mining

algorithm Apriori in CRM. *Pervasive Computing and*, (3), pp.2–6. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4365419 [Accessed July 1, 2012].

Gupta, N., Mangal, N. & Tiwari, K., 2006. Mining quantitative association rules in protein sequences. *Data Mining*, pp.273–281. Available at: http://www.springerlink.com/index/Y3587T614772773T.pdf [Accessed July 1, 2012].

Ha, S.H.S., 2011. Medical Domain Knowledge and Associative Classification Rules in Diagnosis. *International Journal of Knowledge Discovery in Bioinformatics*, 2(1), pp.60–73. Available at: http://services.igi-global.com/resolvedoi/resolve.aspx?doi=10.4018/jkdb.2011010104 [Accessed May 23, 2012].

Hahsler, M., Gruen, B. & Hornik, K., 2005. arules: Mining Association Rules and Frequent Itemsets. Available at: https://cran.r-project.org/web/packages/arules/.

Han, J., Pei, J. & Yin, Y., 2000. Mining frequent patterns without candidate generation. *ACM SIGMOD Record*, 29(2), pp.1–12. Available at: http://dl.acm.org/citation.cfm?id=335372 [Accessed May 30, 2013].

Hu, Y.-H. & Chen, Y.-L., 2006. Mining association rules with multiple minimum supports: a new mining algorithm and a support tuning mechanism. *Decision Support Systems*, 42(1), pp.1–24. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0167923604002052 [Accessed May 11, 2012].

Hu, Z., Shen, L. & Chen, S., 2008. An Improved Apriori-Based Personal Recommendation Algorithm for E-commerce. *2008 Third International Conference on Pervasive Computing and Applications*, pp.60–64. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4783649.

Huang, M.-J., Chen, M.-Y. & Lee, S.-C., 2007. Integrating data mining with case-based

reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*, 32(3), pp.856–867. Available at: http://linkinghub.elsevier.com/retrieve/pii/S095741740600042X [Accessed July 22, 2014].

Huang, M.J., Chen, M.Y. & Lee, S.C., 2007. Integrating data mining with case-based reasoning for chronic diseases prognosis and diagnosis. *Expert Systems with Applications*, 32(3), pp.856–867.

Hughes, T.R. et al., 2000. Functional Discovery via a Compendium of Expression Profiles. *Cell*, 102(1), pp.109–126. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0092867400000155.

Hung, F.-H. & Chiu, H.-W., 2007. Protein-Protein Interaction Prediction based on Association Rules of Protein Functional Regions. *Second International Conference on Innovative Computing, Informatio and Control (ICICIC 2007)*, pp.359–359. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=4428001.

Imanuel, 2014. 40 Top Free Data Mining Software - Predictive Analytics Today. *predictiveanalyticstoday*. Available at: https://www.predictiveanalyticstoday.com/top-data-analysis-software/.

Jingfang, H. & Busheng, L., 2011. An improved algorithm of association rules in the application of web logs. *Energy Procedia*, 13, pp.1282–1286. Available at: http://linkinghub.elsevier.com/retrieve/pii/S187661021102995X [Accessed May 27, 2012].

Karabatak, M. & Ince, M.C., 2009. An expert system for detection of breast cancer based on association rules and neural network. *Expert Systems with Applications*, 36(2), pp.3465–3469. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0957417408001103 [Accessed May 23, 2012].

Kharya, S., 2012. Using data mining techniques for diagnosis and prognosis of cancer disease. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, 2(2), pp.55–66. Available at: http://arxiv.org/abs/1205.1923 [Accessed January 14, 2013].

Kosala, R. & Blockeel, H., 2000. Web mining research: a survey. *Sigkdd Explorations Newsletter*, 2(1), pp.1–15.

Koyutürk, M., Grama, A. & Szpankowski, W., 2004. An efficient algorithm for detecting frequent subgraphs in biological networks. *Bioinformatics (Oxford, England)*, 20 Suppl 1, pp.i200-7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15262800 [Accessed March 9, 2012].

Kumar, D., Sathyadevi, G. & Sivanesh, S., 2011. Decision Support System for Medical Diagnosis Using Data Mining. *International Journal of Computer Science*, 8(3), pp.147–153. Available at: http://ijcsi.org/papers/IJCSI-8-3-1-147-153.pdf [Accessed January 14, 2013].

Larrañaga, P. et al., 2006. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1), pp.86–112.

Lazcorreta, E., Botella, F. & Fernández-Caballero, A., 2008. Towards personalized recommendation by two-step modified Apriori data mining algorithm. *Expert Systems with Applications*, 35(3), pp.1422–1429. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0957417407003582 [Accessed May 6, 2012].

Lenca, P. et al., 2008. On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 184(2), pp.610–626. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0377221706011465 [Accessed May 23, 2012].

Li, L. et al., 2004. Data mining techniques for cancer detection using serum proteomic

profiling. *Artificial intelligence in medicine*, 32(2), pp.71–83. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15364092 [Accessed March 8, 2012].

Liu, B., Hsu, W. & Ma, Y., 1999. Mining Quantitative Association Rules with Multiple Minimum Supports. *Proceedings of the fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp.337–341.

Liu, Y.-C., Cheng, C.-P. & Tseng, V.S., 2011. Discovering relational-based association rules with multiple minimum supports on microarray datasets. *Bioinformatics (Oxford, England)*, 27(22), pp.3142–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21926125 [Accessed March 29, 2012].

Luksza, M., 2005. *A System for Predicting Protein Function from Structure*.

Member, M.A., Ghose, M.K. & Gauthaman, K., 2010. Association Rule Mining in Genomics. *International Journal of Computer Theory and Engineering*, 2(2), pp.269–273.

Meyfroidt, G. et al., 2009. Machine learning techniques to examine large patient databases. *Best Practice and Research Clinical Anaesthesiology*, 23(1), pp.127–143. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1521689608000839 [Accessed November 26, 2012].

Mobasher, B. et al., 2002. Discovery and evaluation of aggregate usage profiles for Web personalization. *Data Mining and Knowledge Discovery*, 6(1), pp.61–82.

Moro, S., Rita, P. & Vala, B., 2016. Predicting social media performance metrics and evaluation of the impact on brand building: A data mining approach. *Journal of Business Research*, 69, pp.3341–3351. Available at: https://archive.ics.uci.edu/ml/datasets/Facebook+metrics.

Mukhopadhyay, A. et al., 2010. Mining Association Rules from HIV-Human Protein Interactions. *International Conference on Systems in Medicine and Biology*, (December), pp.344–348.

Ohsaki, M. et al., 2007. Evaluation of rule interestingness measures in medical knowledge discovery in databases. *Artificial intelligence in medicine*, 41(3), pp.177–96. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17851054 [Accessed April 18, 2012].

Ordonez, C., Ezquerra, N. & Santana, C. a., 2005. Constraining and summarizing association rules in medical data. *Knowledge and Information Systems*, 9(3), pp.295–283. Available at: http://link.springer.com/10.1007/s10115-005-0226-5 [Accessed May 22, 2013].

Oyama, T. et al., 2002. Extraction of knowledge on protein--protein interaction by association rule discovery. *Bioinformatics*, 18(5), pp.705–714. Available at: http://bioinformatics.oxfordjournals.org/content/18/5/705.short [Accessed July 11, 2012].

Park, S.H. et al., 2009. Prediction of protein-protein interaction types using association rule based classification. *BMC bioinformatics*, 10(1), p.36. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2667511&tool=pmcentrez&rendertype=abstract [Accessed March 14, 2012].

Prompramote, S., Chen, Y. & Chen, Y.P., 2005. 5 Machine Learning in Bioinformatics. *Machine learning in bioinformatics. In Bioinformatics technologies . Springer Berlin Heidelberg.*, pp.117–153.

Railean, I. et al., 2013. Closeness Preference-A new interestingness measure for sequential rules mining. *Knowledge-Based Systems*, 44.

Sahar, S., 2010. Interestingness Measures-On Determining What Is Interesting. In *Data Mining and Knowledge Discovery Handbook*. Available at: http://www.springerlink.com/index/WLN5RV73Q26T254J.pdf.

Schafer, J. Ben, 2009. The Application of Data-Mining to Recommender Systems. *Encyclopedia of data warehousing and mining*, pp.44–48.

Soni, J. et al., 2011. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction. *International Journal of Computer Applications*, 17(8), pp.43–48. Available at: http://search.ebscohost.com/login.aspx?direct=true&profile=ehost&scope=site&authtype=crawler&jrnl=09758887&AN=74184303&h=KvSpngkTsPtbMPAj4I04aBMcE1j6ncmWWYXI6hOisilndCDh+okd4iogsITZm98dgyGvOs6SWp3bjXiVImPjPg==&crl=c [Accessed March 18, 2014].

Stelle, D., Barioni, M.C. & Scott, L.P., 2011. Using data mining to identify structural rules in proteins. *Applied Mathematics and Computation*, 218(5), pp.1997–2004. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0096300311009301 [Accessed May 23, 2012].

Tang, Y., Jin, B. & Zhang, Y.-Q., 2005. Granular support vector machines with association rules mining for protein homology prediction. *Artificial intelligence in medicine*, 35(1–2), pp.121–34. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16024240 [Accessed November 19, 2012].

Tilman B. Drüeke, M.D., Francesco Locatelli, M.D., Naomi Clyne, M.D., Kai-Uwe Eckardt, M.D., Iain C. Macdougall, M.D., Dimitrios Tsakiris, M.D., Hans-Ulrich Burger, P.D. & and Armin Scherhag, M.., 2006. Normalization of Hemoglobin Level in Patients with Chronic Kidney Disease and Anemia. *The New England Journal of Medicine*, 355(20), pp.2071–2084. Available at: http://www.nejm.org/doi/pdf/10.1056/NEJMoa062276.

Toivonen, H., 1996. Sampling large databases for association rules. *Proceedings of the International Conference on Very Large Data Bases*, 10(7), pp.134–145. Available at: http://www.ict.griffith.edu.au/~vlad/teaching/kdd.d/readings.d/toivonen96sampling.pdf [Accessed June 24, 2012].

Tuzhilin, A. & Adomavicius, G., 2002. Handling very large numbers of association rules in the analysis of microarray data. *Proceedings of the eighth ACM SIGKDD*

international conference on Knowledge discovery and data mining - KDD '02, p.396. Available at: http://portal.acm.org/citation.cfm?doid=775047.775104.

Vo, B. & Le, B., 2011. Interestingness measures for association rules: Combination between lattice and hash tables. *Expert Systems with Applications*, 38(9), pp.11630–11640. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0957417411004507 [Accessed May 27, 2012].

Wang, D. et al., 2010. Association Rule Mining Based on Concept Lattice in Bioinformatics Research. *and Computer Science (*. Available at: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5462360 [Accessed July 1, 2012].

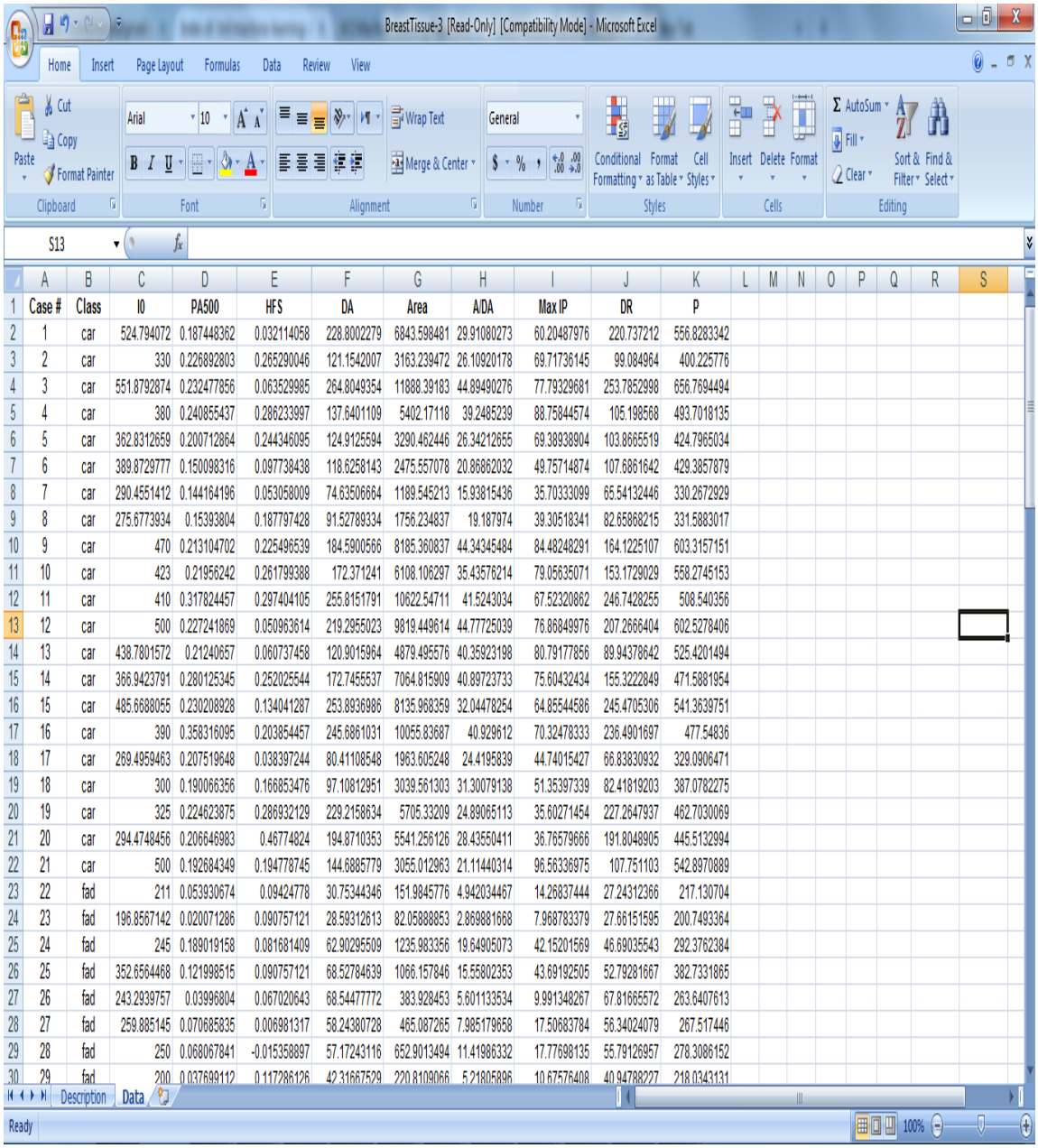
Wright, A., Chen, E.S. & Maloney, F.L., 2010. An automated technique for identifying associations between medications, laboratory results and problems. *Journal of Biomedical Informatics*, 43(6), pp.891–901. Available at: http://dx.doi.org/10.1016/j.jbi.2010.09.009.

Zaki, M., 2004. Report on BIOKDD04: workshop on data mining in Bioinformatics. *ACM SIGKDD Explorations ...*, 6(2), pp.153–154. Available at: http://dl.acm.org/citation.cfm?id=1046486 [Accessed July 26, 2012].

Zhang, M. et al., 2009. Study on the recommendation technology for tourism information service. In *ISCID 2009 - 2009 International Symposium on Computational Intelligence and Design*. Ieee, pp. 410–415. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5370162 [Accessed May 23, 2012].

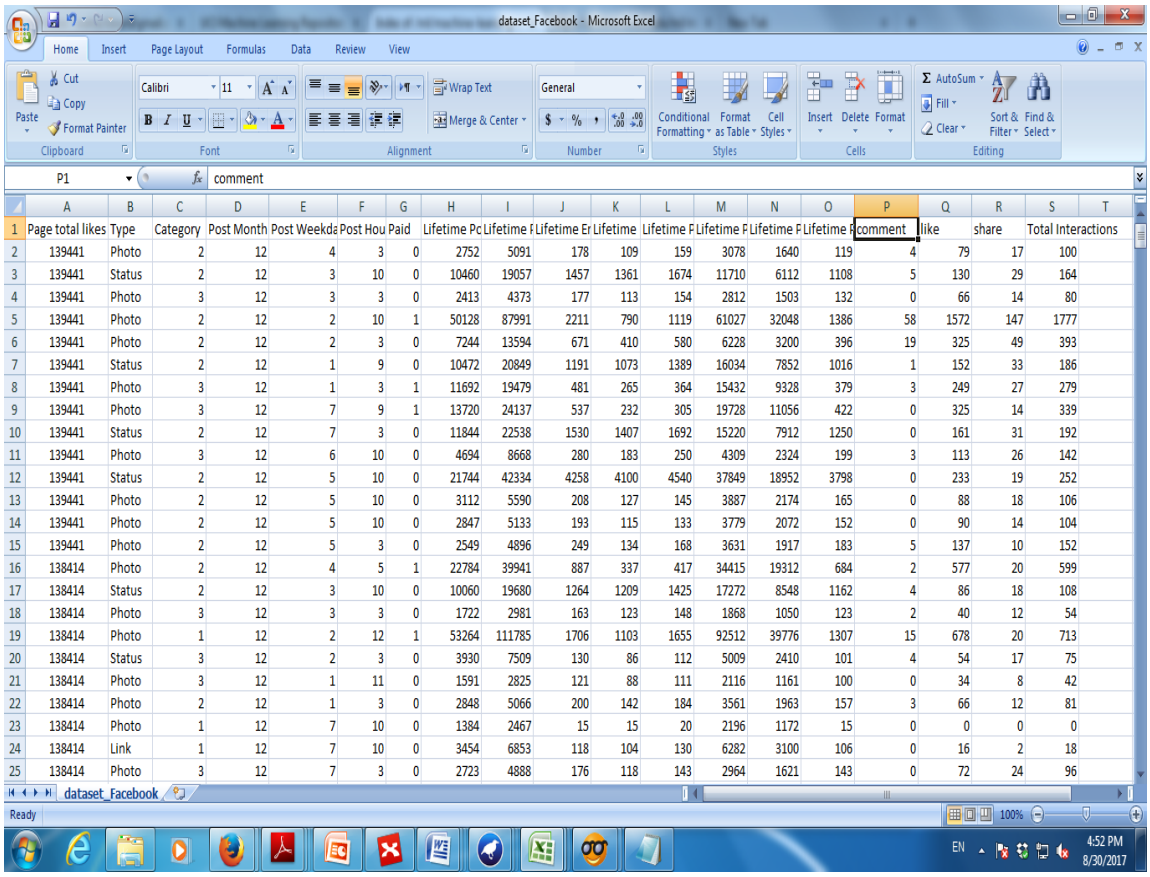
Zuhtuogullari, K. & Allahverdi, N., 2011. An improved itemset generation approach for mining medical databases. In *International Symposium on Innovations in Intelligent Systems and Applications*. Ieee, pp. 39–43. Available at: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5946123.

Appendix A



	A	B	C	D	E	F	G	H	I	J	K
1	Case #	Class	I0	PA500	HFS	DA	Area	A/DA	Max IP	DR	P
2	1	car	524.794072	0.187448362	0.032114058	228.8002279	6843.598481	29.91080273	60.20487976	220.737212	556.8283342
3	2	car	330	0.226892803	0.265290046	121.1542007	3163.239472	26.10920178	69.71736145	99.084964	400.225776
4	3	car	551.8792874	0.232477856	0.063529985	264.8049354	11888.39183	44.89490276	77.79329681	253.7852998	656.7694494
5	4	car	380	0.240855437	0.286233997	137.6401109	5402.17118	39.2485239	88.75844574	105.198568	493.7018135
6	5	car	362.8312659	0.200712864	0.244346095	124.9125594	3290.462446	26.34212655	69.38938904	103.8665519	424.7965034
7	6	car	389.8729777	0.150098316	0.097738438	118.6258143	2475.557078	20.86862032	49.75714874	107.6861642	429.3857879
8	7	car	290.4551412	0.144164196	0.053058009	74.63506664	1189.545213	15.93815436	35.70333099	65.54132446	330.2672929
9	8	car	275.6773934	0.15393804	0.187797428	91.52789334	1756.234837	19.187974	39.30518341	82.65868215	331.5883017
10	9	car	470	0.213104702	0.225496539	184.5900566	8185.360837	44.34345484	84.48248291	164.1225107	603.3157151
11	10	car	423	0.21956242	0.261799388	172.371241	6108.106297	35.43576214	79.05635071	153.1729029	558.2745153
12	11	car	410	0.317824457	0.297404105	255.8151791	10622.54711	41.5243034	67.52320862	246.7428255	508.540356
13	12	car	500	0.227241869	0.050963614	219.2955023	9819.449614	44.77725039	76.86849976	207.2666404	602.5278406
14	13	car	438.7801572	0.21240657	0.060737458	120.9015964	4879.495576	40.35923198	80.79177856	89.94378642	525.4201494
15	14	car	366.9423791	0.280125345	0.252025544	172.7455537	7064.815909	40.89723733	75.60432434	155.3222849	471.5881954
16	15	car	485.6688055	0.230208928	0.134041287	253.8936986	8135.968359	32.04478254	64.85544586	245.4705306	541.3639751
17	16	car	390	0.358316095	0.203854457	245.6861031	10055.83687	40.929612	70.32478333	236.4901697	477.54836
18	17	car	269.4959463	0.207519648	0.038397244	80.41108548	1963.605248	24.4195839	44.74015427	66.83830932	329.0906471
19	18	car	300	0.190066356	0.166853476	97.10812951	3039.561303	31.30079138	51.35397339	82.41819203	387.0782275
20	19	car	325	0.224623675	0.286932129	229.2158634	5705.33209	24.89065113	35.60271454	227.2647937	462.7030069
21	20	car	294.4748456	0.206646983	0.46774824	194.8710353	5541.256126	28.43550411	36.76579666	191.8048905	445.5132994
22	21	car	500	0.192684349	0.194778745	144.6885779	3055.012963	21.11440314	96.56336975	107.751103	542.8970889
23	22	fad	211	0.053930674	0.09424778	30.75344346	151.9845776	4.942034467	14.26837444	27.24312366	217.130704
24	23	fad	196.8567142	0.020071286	0.090757121	28.59312613	82.05888853	2.869881668	7.968783379	27.66151595	200.7493364
25	24	fad	245	0.189019158	0.081681409	62.90295509	1235.983356	19.64905073	42.15201569	46.69035543	292.3762384
26	25	fad	352.6564468	0.121998515	0.090757121	68.52784639	1066.157846	15.55802353	43.69192505	52.79281667	382.7331865
27	26	fad	243.2939757	0.03996804	0.067020643	68.54477772	383.928453	5.601133534	9.991348267	67.81665572	263.6407613
28	27	fad	259.885145	0.070685835	0.006981317	58.24380728	465.087265	7.985179658	17.50683784	56.34024079	267.517446
29	28	fad	250	0.068067841	-0.015358897	57.17243116	652.9013494	11.41986332	17.77698135	55.79126957	278.3086152
30	29	fad	200	0.037699112	0.117286126	42.31667529	220.8109066	5.21805896	10.67576408	40.94788227	218.0343131

Figure A-1: Breast Tissue Data.xls



Page total likes	Type	Category	Post Month	Post Weekda	Post Hou	Paid	Lifetime Pc	Lifetime F	Lifetime Er	Lifetime	Lifetime F	Lifetime F	Lifetime F	Lifetime R	comment	like	share	Total Interactions
139441	Photo	2	12	4	3	0	2752	5091	178	109	159	3078	1640	119	4	79	17	100
139441	Status	2	12	3	10	0	10460	19057	1457	1361	1674	11710	6112	1108	5	130	29	164
139441	Photo	3	12	3	3	0	2413	4373	177	113	154	2812	1503	132	0	66	14	80
139441	Photo	2	12	2	10	1	50128	87991	2211	790	1119	61027	32048	1386	58	1572	147	1777
139441	Photo	2	12	2	3	0	7244	13594	671	410	580	6228	3200	396	19	325	49	393
139441	Status	2	12	1	9	0	10472	20849	1191	1073	1389	16034	7852	1016	1	152	33	186
139441	Photo	3	12	1	3	1	11692	19479	481	265	364	15432	9328	379	3	249	27	279
139441	Photo	3	12	7	9	1	13720	24137	537	232	305	19728	11056	422	0	325	14	339
139441	Status	2	12	7	3	0	11844	22538	1530	1407	1692	15220	7912	1250	0	161	31	192
139441	Photo	3	12	6	10	0	4694	8668	280	183	250	4309	2324	199	3	113	26	142
139441	Status	2	12	5	10	0	21744	42334	4258	4100	4540	37849	18952	3798	0	233	19	252
139441	Photo	2	12	5	10	0	3112	5590	208	127	145	3887	2174	165	0	88	18	106
139441	Photo	2	12	5	10	0	2847	5133	193	115	133	3779	2072	152	0	90	14	104
139441	Photo	2	12	5	3	0	2549	4896	249	134	168	3631	1917	183	5	137	10	152
138414	Photo	2	12	4	5	1	22784	39941	887	337	417	34415	19312	684	2	577	20	599
138414	Status	2	12	3	10	0	10060	19680	1264	1209	1425	17272	8548	1162	4	86	18	108
138414	Photo	3	12	3	3	0	1722	2981	163	123	148	1868	1050	123	2	40	12	54
138414	Photo	1	12	2	12	1	53264	111785	1706	1103	1655	92512	39776	1307	15	678	20	713
138414	Status	3	12	2	3	0	3930	7509	130	86	112	5009	2410	101	4	54	17	75
138414	Photo	3	12	1	11	0	1591	2825	121	88	111	2116	1161	100	0	34	8	42
138414	Photo	2	12	1	3	0	2848	5066	200	142	184	3561	1963	157	3	66	12	81
138414	Photo	1	12	7	10	0	1384	2467	15	15	20	2196	1172	15	0	0	0	0
138414	Link	1	12	7	10	0	3454	6853	118	104	130	6282	3100	106	0	16	2	18
138414	Photo	3	12	7	3	0	2723	4888	176	118	143	2964	1621	143	0	72	24	96

Figure 0-2: Facebook Metrics dataset