PREDICATE BASED ASSOCIATION RULES MINING WITH NEW INTERESTINGNESS MEASURE

HAFIZ ISHFAQ AHMAD

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Doctor of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JULY 2022

DEDICATION

This thesis is wholeheartedly dedicated to my beloved parents, who have been my source of inspiration and gave us strength when we thought of giving up, who continually provide their moral, spiritual, emotional, and financial support.

To my brothers, sisters and friends who shared their words of advice and encouragement to finish this study.

Lastly, to Rahila for her advice, her patience, and her faith because she always understood.

ACKNOWLEDGEMENT

"**To Allah belong all glory and power. To Him mount up (all) words of purity: It is He Who exalts each deed of righteousness**". All Praises for ALLAH SWT who enabled me to complete this thesis. The One, All-Compassionate, All-Merciful, and Magnificent without Whose consent nothing can happen in the world.

First, my deep gratitude goes to Dr. Alex Tze Hiang Sim, who expertly guided me through my graduate education and who shared the excitement of four year of discovery. His unwavering enthusiasm for Data Mining and Machine Learning kept me constantly engaged with my research, and his personal generosity helped make my time at UTM enjoyable.

I would also like to give my sincere thanks to my supervisor Assoc. Prof. Dr. Roliana Ibrahim, who accepted me as her PhD student after Dr. Alex left UTM. Her advice and guidance help me move forward in the right direction. I have learned a lot from her, without her help I could not have finished my PhD successfully.

My sincere thanks goes to the Ministry of Higher Education of Malaysia for their financial support under the "Malaysia International Scholarship (MIS)".

My parents, brothers, and the sisters, who gave me love, support, and guidance throughout my life that enabled me to reach the stars and achieving my dreams. I thank you for your hardships and struggles that you bear for my success. I love you all.

Last but not the least, I am greatly indebted to my devoted wife, and my daughter. They form the backbone and origin of my happiness. Their love and support without any complaint or regret has enabled me to complete this PhD. I owe my every achievement to them.

ABSTRACT

Association Rule Mining (ARM) is one of the fundamental components in the field of data mining that discovers frequent itemsets and interesting relationships for predicting the associative and correlative behaviours for new data. However, traditional ARM techniques are based on support-confidence that discovers interesting association rules (ARs) using predefined minimum support (*minsupp*) and minimum confidence (*minconf*) threshold. In addition, traditional AR techniques only consider frequent items while ignoring rare ones. Thus, a new parameter-less predicated based ARM technique was proposed to address these limitations, which was enhanced to handle the frequent and rare items at the same time. Furthermore, a new interestingness measure, called g measure, was developed to select only highly interesting rules. In this proposed technique, interesting combinations were firstly selected by considering both the frequent and the rare items from a dataset. They were then mapped to the pseudo implications using predefined logical conditions. Later, inference rules were used to validate the pseudo-implications to discover rules within the set of mapped pseudo-implications. The resultant set of interesting rules was then referred to as the predicate based association rules. Zoo, breast cancer, and car evaluation datasets were used for conducting experiments. The results of the experiments were evaluated by its comparison with various classification techniques, traditional ARM technique and the coherent rule mining technique. The predicate-based rule mining approach gained an accuracy of 93.33%. In addition, the results of the g measure were compared with a state-of-the-art interestingness measure developed for a coherent rule mining technique called the h value. Predicate rules were discovered with an average confidence value of 0.754 for the zoo dataset and 0.949 for the breast cancer dataset, while the average confidence of the predicate rules found from the car evaluation dataset was 0.582. Results of this study showed that a set of interesting and highly reliable rules were discovered, including frequent, rare and negative association rules that have a higher confidence value. This research resulted in designing a methodology in rule mining which does not rely on the *minsupp* and *minconf* threshold. Also, a complete set of association rules are discovered by the proposed technique. Finally, the interestingness measure property for the selection of combinations from datasets makes it possible to reduce the exponential searching of the rules.

ABSTRAK

Perlombongan Peraturan Penyatuan (ARM) merupakan salah satu komponen asas dalam bidang perlombongan data yang menemui set item yang kerap dan perhubungan yang menarik bagi meramalkan tingkah laku asosiatif dan korelatif untuk data baharu. Walau bagaimanapun, teknik ARM tradisional adalah berdasarkan keyakinan sokongan yang menemui peraturan penyatuan (AR) yang menarik menggunakan ambang sokongan minimum yang dipratakrifkan (minsupp) dan keyakinan minimum (minconf) yang telah ditetapkan. Di samping itu, teknik AR tradisional hanya mempertimbangkan item yang kerap dan mengabaikan item yang jarang berlaku. Oleh itu, teknik ARM berasaskan penetapan predikat tanpa parameter baharu, telah dicadangkan untuk menangani batasan ini yang dipertingkatkan untuk mengendalikan item yang kerap dan jarang berlaku pada masa yang sama. Tambahan pula, satu pengukuran tahap daya tarikan yang baharu, disebut sebagai keputusan ukuran g telah dibangunkan untuk memilih peraturan yang berdaya tarikan tinggi sahaja. Dalam teknik yang dicadangkan ini, gabungan yang menarik telah dipilih terlebih dahulu dengan mempertimbangkan kedua-dua item yang kerap dan jarang berlaku daripada set data. Ia kemudiannya dipetakan kepada implikasi pseudo menggunakan kondisi logik yang telah ditetapkan. Selepas itu, peraturan inferens digunakan untuk mengesahkan implikasi pseudo untuk menemui peraturan dalam set implikasi pseudo yang dipetakan. Set peraturan yang menarik yang terhasil kemudiannya dirujuk sebagai peraturan penyatuan berasaskan penetapan predikat. Set data zoo, kanser payudara, dan penilaian kereta telah digunakan untuk menjalankan eksperimen. Keputusan eksperimen dinilai dengan perbandingannya dengan pelbagai teknik pengelasan, teknik ARM tradisional dan teknik perlombongan peraturan koheren. Pendekatan perlombongan peraturan berasaskan penetapan predikat telah mendapat ketepatan 93.33%. Selain itu, keputusan ukuran g telah dibandingkan dengan pengukuran daya tarikan terkini yang dibangunkan untuk teknik perlombongan peraturan koheren yang dinamakan sebagai nilai h. Peraturan penetapan predikat ditemui dengan nilai keyakinan purata 0.754 untuk set data zoo dan 0.949 untuk set data kanser payudara, manakala purata keyakinan peraturan penetapan predikat yang ditemui daripada set data penilaian kereta adalah 0.582. Keputusan menunjukkan bahawa satu set peraturan yang menarik dan sangat dipercayai telah ditemui termasuk peraturan perkaitan yang kerap, jarang berlaku dan negatif yang mempunyai nilai keyakinan yang lebih tinggi. Kajian ini menghasilkan pembentukan metodologi dalam perlombongan peraturan yang tidak bergantung pada ambang minsupp dan minconf. Ia juga, adalah satu set lengkap peraturan persatuan yang ditemui oleh teknik yang dicadangkan. Akhir sekali, ciri pengukuran daya tarikan untuk pemilihan gabungan daripada set data memungkinkan untuk mengurangkan pencarian eksponen peraturan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ARM	-	Association Rule Mining
NAR	-	Negative Association Rule
AR	-	Association Rule
minsupp	-	Minimum Support
minconf	-	Minimum Confidence
FP-Tree	-	Frequent Pattern Tree
TID	-	Transaction ID
AODE	-	Averaged One-dependence Estimators
LDA	-	Linear Discriminant Analysis
SVM	-	Support Vector Machine
supp	-	Support
GEP	-	Gene Expression Programming

LIST OF SYMBOLS

$\equiv$	-	Logical Equivalence
$\neg$	-	Logical Not
λ	-	Lambda
$\rightarrow$	-	Logical Implication
$\cdot$	-	Logical Conjunction Operator
$\vee$	-	Logical Disjunction Operator
$\wedge$	-	Logical Conjunction Operator
I	-	Set of Items

CHAPTER 1

INTRODUCTION

1.1 Overview

Currently, vast amount of data is stored all over the Internet and the Web that has exceeded what human can comprehend on their own. Different tools and applications are needed to understand the data as the analysis of this data for turning it to useful information is crucial. This overwhelming state of the available data resulted in the discovery of data mining and has become a hot research topic. Data mining is defined by Witten *et al.* (2016) as the technique of extracting knowledge from the data. The knowledge gained from processing these large amounts of data can and has been used extensively for applications in various fields including market basket analysis, science exploration, production control and many more (Han *et al.*, 2011).

One of the dominant technique of data mining is the association rule mining (ARM). ARM deals with the discovery of relationships among frequently appearing item sets. The primary focus of ARM is to find rules with minimal human effort to predict co-occurrence of items and to discover relations, in the form of rules that are more interesting among items in large datasets. These interesting rules provide user with the ability of justifiable decision-making steps based on data pattern.

In ARM techniques the interestingness of association rules (ARs) is calculated based on different statistical and mathematical equations. In case of traditional ARM techniques (Brin *et al.*, 1997a; Park *et al.*, 1995; Agrawal and Srikant, 1994; Agrawal *et al.*, 1993), the interestingness of an AR is based on the idea to find all the rules that has a high support and high confidence threshold than a predefined minimum support (*minsupp*) and minimum confidence (*minconf*) threshold (Han *et al.*, 2011; Agrawal *et al.*, 1993). These measures select the most frequent item sets from the transaction records and rejects the non-frequent and the frequent absent item sets from the transaction records. However, in some cases, the non-frequent items which are not reported, due to the restrictions applied by the *minsupp* and *minconf*, can be useful (Koh *et al.*, 2008; Koh *et al.*, 2006). Therefore, the mined rules found are incomplete and making decision on these incomplete rules may have unfavourable effects on decision making. In addition, predefining an appropriate *minsupp* and *minconf* threshold requires an expert user because if an inappropriate *minsupp* and *minconf* value is predefined then redundant rules are discovered or there is a loss of important rules.

Moreover, ARM algorithms generate a huge number of rules, some of which are trivial and in some cases most of them are redundant. Understanding these large number of rules is difficult and in turn reduces the effectiveness of ARM algorithms. Thus, identification of the most useful rules and the filtering of the irrelevant rules must be carried out (Ju *et al.*, 2015). The practical application of ARM algorithms will be benefitted by implementation of techniques that filters redundant rules and discovery of hidden useful information from the dataset.

Multiple algorithms are developed for the discovery of ARs using propositional and predicate logic. For example, a rule-based expert system was developed by (Ikram and Qamar, 2015) to predict earthquakes based on previous data where the rules discovered are polished using predicate logic. In addition, mining predicate association by gene expression programming is also proposed for the discovery of association rules that cannot be expressed and discovered by traditional techniques (Zuo *et al.*, 2002). Propositional logic has also been used to overcome the limitations of minimum support and confidence model and to discover a set of coherent rules that are more reliable (Sim *et al.*, 2010)

In order to overcome the above-mentioned limitations, an ARM algorithm is introduced for the discovery of interesting ARs based on the concepts of predicate logic without presetting the *minsupp* and *minconf* threshold. The proposed technique discovers all frequent rules and rules where the consequence of the rules is rare to address the issue of discovering incomplete set of rules. Moreover, an interestingness measure is developed to identify and extract the most interesting and useful rules from datasets.

In this chapter, section 1.2 provides an explanation on ARM followed by the problem background of this research and its related solutions in section 1.3. The problem statement is elaborated in section 1.4 and the research questions of the study is explained in section 1.5. Next, the aims and objectives are outlined in section 1.6, and the motivation of the research study is summarized in section 1.7. Section 1.8 presents the scope of the study, while the significance of the study is discussed in section 1.9. The major contributions of this thesis are highlighted in section 1.10. Finally, the chapter concludes with a summary in section 1.11.

1.2 Association Rule Mining

Association rule mining (ARM), first introduced by Agrawal *et al.* (1993), is the process of finding frequent patterns, correlation, and associations among the items of a transactional database. It is one of the most frequently used tools to identify and extract relationships between items/attributes in a dataset/database. An association rule (AR) is generally in the $X \rightarrow Y$ form where X and Y are items or sets of items. The left-hand side of the rule is called the antecedent of the rule while the right-hand side of the rule is the consequence of the rule. This rule is read as X implies Y and it states that wherever X is present in a transactional record, Y will also be present in that transaction. Generally, the consequence of the rule consists of only one item found in combination with antecedent and the antecedent consist of one or more data items combined (Huang *et al.*, 2017; Makino *et al.*, 2017).

ARM discovers extensively large set of rules that also contains redundant and trivial rules. To reduce the set of rules generally a measure is used to remove these redundant and uninteresting rules. The traditional approach is based on the support-confidence concept as proposed by Agrawal *et al.* (1993), where the dependence of each item on the other is measured with two factors namely, support and confidence.

Support is the frequency of an item set appearing in a dataset, while confidence is the probability of both antecedent and consequence appearing in same transactions.

The process of mining ARs is often divided into two steps; (a) discovering all the frequent item sets whose frequency exceeds the *minsupp* threshold and (b) generation of ARs from the frequent item set using the constraints of *minsupp* and *minconf* threshold (Han *et al.*, 2011). In the first step, a given dataset is searched to discover the repeating pattern of attribute-value pairs that exceeds the preset *minsupp* threshold. These pairs of attribute-value are named as items which forms the frequent item sets. Next, these frequent item sets are analysed for the generation of ARs.

A major challenge for mining ARs is the generation of huge number of frequent item sets from a large dataset. The number of frequent item sets increase especially when the *minsupp* threshold value is set to low. In this case, many of the rules generated are similar to each other and no new information can be gained from those rules (Fournier-Viger *et al.*, 2017). On the other hand, if the *minsupp* threshold value is set to high, interesting rules can be lost and an incomplete set of rules will be discovered.

Suppose a grocery shop manager wants to know which items are frequently purchased together, then an AR mined from the transaction records of the grocery shop will look like:

$$buys(X, \text{"Milk"}) \rightarrow buys(\text{X}, \text{"Bread"})\ [support = 5\%, confidence\ =\ 50\%]$$

where a customer is represented as X, a support of 5% indicates that 5% of all the customers, under the analysis, bought Milk and Bread together. While the 50% confidence indicates that there is a 50% chance that if a customer buys Milk, then he/she will also buy Bread. In this example, if any of the rules does not satisfy the user defined threshold for *minsupp* and *minconf*, then those rules will be discarded.

1.3 Problem Background

ARM has various applications in different areas including market basket analysis, science exploration and many more. The rule mining process, introduced by Agrawal *et al.* (1993), uses the *minsupp* and *minconf* threshold for the discovery of ARs. The number of rules generated by mining ARs may be huge depending on how the *minsupp* threshold is set (Fournier-Viger *et al.*, 2017). This huge number of rules contains a lot of redundancy as well as a possibility of weak correlation between frequent item sets. Therefore, the analysis of these rules seems impossible and less productive.

The explosion in data growth in the recent history resulted in the birth of the data mining techniques for knowledge extraction. ARM was the initial development in the data mining to extract the meaningful information from the huge data (Agrawal, Imielinski, & Swami, 1993; Agrawal & Srikant, 1994). The ARM is then rigorously studied and improved by the researchers in the following years such as Yan *et al.* (2009), Djenouri *et al.* (2013), Soysal (2015), Goyal *et al.* (2015), Narvekar and Syed (2015), and in Yuan and Ding (2012). The key limitation of the Apriori algorithm was its performance issues with respect to the time, thus many researchers proposed efficient techniques such as Hong and Bian (2008), Yang (2004), Kuo and Shih (2007), Guo *et al.* (2017), and in Han *et al.* (2004). Among all these, the FP-growth was a breakthrough in the time of the extraction of frequent rules (Han *et al.*, 2004).

However, there are a number of other disadvantages identified by the researchers. a) Requirement of user input parameters such as minimum support *(minsupp)* and minimum confidence *(minconf);* b) the exponential number of candidate rules generation. Researchers are working on the improvement of both issues. Selection of the predefined values are input parameters is difficult, especially for large databases and handling of large candidate rules is humanly not possible. The generation of the large rules may also negatively impact the results (Yan *et al.*, 2009). Thus, researchers proposed automatic parameter setting approaches as well parameter less approaches as discussed by Yan *et al.* (2009), Kuoa *et al.* (2011), Dash *et al.* (2013), and Minaei-Bidgoli *et al.* (2013). Another important task is identifying interesting rules (Dash *et al.*, 2013; Minaei-Bidgoli *et al.*, 2013; Nandhini *et al.*, 2012). A large number of techniques are developed that evaluate the interestingness of frequent rules.

The selection of a subset of rules is always risky and its impact is twofold. The subset may consist of the strongest rules and the analysis will only considers subset of the rules. There is no guarantee that we found all the interesting rules, similarly there is also no guarantee that we have less interesting rule in the selected subset (Sim *et al.*, 2010). The rejection of hidden strong rules will lead to flawed conclusions of decision makers who are taken by the consideration that the selected rules are all the rules discovered from the data. Moreover, it can be considered that there always exists an appropriate *minsupp* threshold for each dataset, but it is hard to find. This leads to the major issue of discovering interesting rules in a dataset (Webb and Vreeken, 2014).

Therefore, algorithms have been designed for the extraction of concise representation from the discovered frequent item sets. This representation summarizes the whole set of the frequent item sets into a much smaller set of frequent item sets (Fournier-Viger *et al.*, 2017). It is also reported that the discovery of these concise representation is much faster as compared to the discovery of frequent item sets in the traditional ARM techniques as there is no need to scan the dataset multiple times for calculating the support of each item set (Aliberti *et al.*, 2015; Soulet and Rioult, 2014; Fournier-Viger *et al.*, 2014; Szathmary *et al.*, 2014; Lucchese *et al.*, 2006). Some of the most popular examples for extraction the summary set of the frequent item sets from a dataset are closed item sets (Aliberti *et al.*, 2015; Vo *et al.*, 2012), maximal item sets (Uno *et al.*, 2004), and generator item sets (Soulet and Rioult, 2014; Fournier-Viger *et al.*, 2014; Szathmary *et al.*, 2014).

Furthermore, in discovering frequent item sets the mining process considers that all items as equal (Fournier-Viger *et al.*, 2017), however, in real life items are much more different than each other (Liu *et al.*, 1999a). For example, milk and avocado are two different items with selling frequencies very much different than each other in a retail store. The sale of avocado is infrequent as compared to the milk, therefore, these both or not equal item in real life. This introduce the importance of the infrequent items which is almost negative of the frequent items and represent rarity of items set (Liu *et al.*, 1999a). Thus, meaning that some items from a transaction record are less likely to show up in frequent item sets as compared to others. These infrequent items or rare items are mostly discarded by the traditional ARM techniques by failing the *minsupp* and the *minconf* condition. The ARs that satisfies the *minsupp* threshold requirement are important to be identified in some cases, however, there is the possibility to discover more useful knowledge using the ARs of the infrequent items (de Sá *et al.*, 2018; Kim and Yun, 2016; Lin *et al.*, 2015; Troiano and Scibelli, 2014). The rejected rules that do not satisfy the condition of *minsupp* might be needed to improve data mining models for different datasets.

The infrequent item problem has been addressed by researchers using the technique of multiple *minsupp* threshold algorithms (Kiran and Reddy, 2011; Hu and Chen, 2006; Liu *et al.*, 1999a). In these algorithms, users can choose different *minsupp* threshold for each item based on their appearance in the transaction records. Furthermore, other methods have also been proposed where the infrequent item sets are searched in the dataset instead of searching the frequent items (Koh and Ravana, 2016; Szathmary *et al.*, 2012; Szathmary *et al.*, 2007).

Next, ARM techniques are working on the concept of the presence of the items in the dataset while it does not consider the absence of an item in a given transaction record. Ignoring the rules with stronger association that shows the absence of an item in the transaction can be misleading. For example, ignoring a stronger association among the presence of item X and absence of item Y is misleading in comparison with reporting the presence of item X with presence of item Y which has less strong association than the former association. Therefore, leading to inappropriate decisions as a consequence.

Moreover, selection of interesting ARs requires a detailed insight into the data as it is infeasible to include all the frequent item sets discovered. Through this step, all the ARs that are less interesting will be dropped while the most interesting rules will be used for decision making. Consequently, the evaluation methodology for mining association rules is of great significance in both theory design and practical application.

In the recent past, the researchers achieved a reasonable success in the development of the interestingness measure of association rules (Datta and Mali, 2017; Ju *et al.*, 2015; Chen, 2007). Chen (2007) introduced a data envelopment analysis (DEA) as a post-processing approach. The DEA is used after discovering of the AR to rank these rules based of a predefined criterion. Toloo *et al.* (2019) improved the DEA by introducing mixed integer linear programming (MILP). Objective interestingness mainly considers statistical significance features of objective data, including Support, Confidence and Lift, which are classic, as well as Validity, Conviction, Improvement, and Chi-square analysis, which are relatively new (Geng and Hamilton, 2006). The common objectives of all researchers are to discover the rules that truly reflect the users' interest in the generated rules set. However, various interestingness measures conflict with one another as they produce different results in different circumstances

In the recent past, the researchers introduced logic-based discovery of ARs. (Sim *et al.*, 2010) proposed a coherent rule based on predicate logic to extract knowledge from dataset. These techniques neither use background knowledge nor pre-set parameters. The coherent rules are then used to discover the ARs without setting the *minsupp* or *minconf*. Chen *et al.* (2013b) improved the logic-based rules by using fuzziness in the rules. In the process, it generates the fuzzy candidate coherent rules. Then, the confusion matrix is calculated that these candidates satisfy the four different criteria. If it passes all the criteria, then they are coherent rules. Chen *et al.* (2014) also used a logic-based AR to avoid the *minsupp* and *minconf* and used coherent rules to discover the hidden knowledge from dataset.

The above discussion generates the following hypothesis that need to evaluate.

H_1: The *minsupp* and *minconf* can be replaced by the logic-based techniques that will not require the domain knowledge for generating the association rules.

H_2: The infrequent and rare item also possess important information like frequent items.

H_3: The interestingness measure can identify the frequent and infrequent item and improve the quality of the ARs.

1.4 Problem Statement

The pre-setting of a *minsupp* threshold in the ARM leads to the rejection of rules falling below the threshold value. If the *minsupp* is not specified or the threshold is set to a very low value, it will result in huge number of rules. This makes it impossible to examine all the discovered rules and make decisions based on these rules. However, setting the *minsupp* results in loss of rules that may lead to information loss that can be yielded from a dataset. (Fournier-Viger *et al.*, 2017). On the other hand, there always exists an appropriate *minsupp* threshold for each dataset, but identifying it is almost impossible as it requires an in-depth knowledge of each domain to be mined. Sim *et al.* (2010) introduced logic-based AR that did not use the *minsupp* and *minconf*. Their technique has the ability to extract very strong rule from the dataset. However, the evaluation shows that their logic-based technique is losing information due to their strict internal criteria.

Moreover, traditional ARM does not consider the negative and infrequent rules. Infrequent rules possess the important information that highlight the rarity and casual activities such as fraud, however ARM techniques are not capable of capturing such rules. This also results in loss of rules that contains the association between the presence and absence of items in ARM. The rules discovered will be inadequate and may lead to flawed actions (Chen *et al.*, 2013b).

Besides the *minsupp* and *minconf*, the interestingness measure is also used to select the important rules. There are a number of interestingness measure that are used in the literature. However, these measure either effectively select the frequent ARs or infrequent ARs (Geng and Hamilton, 2006). To the best of the author knowledge, there is not a single technique that can measure the interestingness of both the frequent and infrequent interesting ARs at the same time. An appropriate interestingness measure increases the accuracy and efficiency of discovering interesting ARs by selecting rules using knowledge from a given dataset (Datta and Mali, 2017).

1.5 Research Questions

This research will focus on the following research questions.

(a) How to discover the association rules without *minsupp* and *minconf* using predicate logic?

(b) How to discover the infrequent and negative ARs from a dataset?

(c) How to devise an interestingness measure that can evaluate and improve the frequent and infrequent rules in dataset.

1.6 Aims and Objectives

The aim of the current research was to overcome the limitations of the traditional ARM technique that are connected with the selection of *minsupp* and *minconf* for the discovery of interesting ARs. An ARM algorithm is proposed for discovering interesting ARs without presetting the *minsupp* and *minconf* threshold. The discovery of the ARs is performed using the concepts of predicate logic and a new interestingness measure called the *g* measure that addresses the issue of discovering incomplete set of rules. Moreover, the *g* measure identifies and extracts the most interesting and useful rules from datasets.

This research will focus on achieving the following objectives:

(a) To develop a technique for discovering association rules without *minsupp* and *minconf* using predicate logic.

(b) To design a model for discovering infrequent and negative ARs from a dataset.

(c) To propose an interestingness measure to evaluate frequent and infrequent interesting ARs.

1.7 Motivation

The classical approach of ARM uses the *minsupp* and *minconf* threshold for the discovery of frequent item sets that are strongly associated. However, weakly associated rules that are non-interesting and exceeds the *minsupp* threshold are also generated in this support-confidence based techniques. Moreover, the rules which does not satisfy the *minsupp* are rejected. This results in the loss of rules that may not be frequent but are strongly associated, deriving an incomplete set of interesting ARs. Decision making on such incomplete set of rules leads to inappropriate and erroneous decisions.

In addition, the support-confidence does not consider the absence of items during the discovery of ARs even if they are strongly associated and has the interestingness property for a transactional database. Again, loss of strong and interesting rules occurs, providing the users with incomplete set of rules unknowingly misleading them to take flawed decisions about the relationship among items in a dataset.

Extensively large set of rules are generated in ARM for which algorithms are designed to extract only the strongest rules. Yet, extraction of the strongest rules from an incomplete set of rules does not guarantee a complete set of strong and very interesting rules as they can be hidden due to the rejection of rules in the previous steps.

The motivation of this research study is the limitations discussed above and to develop a technique that finds ARs without the requirement of *minsupp* and *minconf* threshold. This eliminates the adverse effect of missing ARs and discovers a complete set of rules by observing the presence as well as the absence of the item in a transaction record. Moreover, logic deals with understanding of how information is captured and how it is possible for one statement to be the consequence of another. It means that, how the information needed for the conclusion to be drawn is already present in the statement or a group of statements. Therefore, the integration of logic with ARM will lead to discover the complete, valid, and sound ARs from a given dataset.

1.8 Scope of Study

This research study deals with generation of ARs without presetting the threshold for *minsupp* and *minconf*. Thus, the proposed technique will not require the domain knowledge for the user that is required for setting the *minsupp* and *minconf* threshold. Although, theoretically, this study can deal with the huge size of the transactional data, certain assumptions are to define the scope of this study.

This study is introducing an innovative ARM technique that discover interesting predicate rules. This technique does not deal with the partition datasets. Thus, the assumption is made that entire dataset may be loaded in the memory. Moreover, Python programming language was used for the implementation of the proposed techniques and testing it against state-of-the-art techniques for discovering ARs. Due to its productivity, speed, extensive availability of support libraries and having open-source development capabilities, Python is selected as the programming tool.

According to the research field of associative classification, it is a common practice to constraint the right-hand side of the rule to be a single consequence. This research will use the same technique that will search only the rules with a single item set at the right-hand side. The reason for this restriction is to reduce the repeated scanning and evaluation of the rules where at some point the same subset of the rules appear at the right-hand side.

1.9 Significance of Study

In this research, a new technique is introduced for mining predicate rules in datasets without presetting a *minsupp* and *minconf* threshold. Implications and inference rules of predicate logic are used for discovering ARs from frequent as well as infrequent items that are interesting from datasets. Using the truth table values of items in a dataset for mapping implications to ARs eliminates the requirement of *minsupp* and *minconf* threshold. Thus, extracting ARs that are hidden but strong and making it a statistically sound process to prove the rules are valid.

Moreover, the techniques proposed in this study will enable user to consider all possible combination of item sets that are both present and absent in each transaction record. Therefore, providing a complete set of rules resulting in user confidence on the knowledge discovery from the data. However, minimum threshold requirement results in discovering incomplete rules from the data mining activities. Therefore, the decision made on the incomplete rules will lead to erroneous decisions.

A new interestingness measure proposed that has a property for reducing the exponential searching during the discovery of ARs. Thus, discarding the non-interesting ARs at the start of the rule discovery and limit the processing. In addition, a significant characteristic of this property is the discovery of ARs whose consequence is frequent or rare by comparing the reliability differences between the presence and absence of the consequence of the rule. Furthermore, the interestingness measure ranks the rules based on their interestingness and provides users with a set of complete and interesting rules building user's confidence for decision making.

1.10 Contributions

This research makes the following contribution by introducing a new ARM technique offering a complete set of rules discovered using a new interestingness measure:

(a) Parameter-less association rules mining technique: The concepts of predicate logic including implications and inference rules are used to discover ARs. This results in designing a methodology in rule mining which does not rely on the *minsupp* and *minconf* threshold.

(b) Negative and rare itemset ARs: Beside frequent, the negative and infrequent item sets are considered during the discovery of AR to overcome the limitation of traditional ARM techniques that only considers the discovery of frequent item set. A complete set of predicate rules are discovered that are interesting because the selection of rules is performed based on logical techniques.

(c) Eliminating / limiting exponential search space: The interestingness measure property for selection of combinations from datasets makes it possible to reduce the exponential searching of the rules. The rules discovered only considers the most interesting items from the dataset and discards the non-interesting item/item sets. This reduces the complexity of the searching process and also makes the rule mining process computationally adequate to be implemented.

(d) Measure for Predicate rules: A new measure is developed for selection of the interesting predicate rules from the dataset called the *g* measure. The measure is designed to find consequence of a rule that is rare or frequent by comparing the reliability difference between the presence and the absence of the consequence of the rule. Thus, rule discovered are highly reliable and decision made on these rules are correct in comparison to the incomplete set of rules discovered using traditional ARM techniques.

1.11 Summary and Thesis Organization

This chapter provided an overview of the research conducted, highlighting the problem background and the objectives of the research. The scope of the research to be covered is introduced and the contributions attained during the research study is described.

This thesis is divided into six chapters. The second chapter of the thesis describes the basic concepts and terminologies of ARM and a critical analysis of the state-of-the-art research conducted in the field of ARM. Chapter 3 focuses on the research methodology describing the steps to discover ARs without *minsupp* and *minconf* threshold. It is followed by chapter 4, which presents the ARM process in detail. The results and discussion are provided in chapter 5 and finally, chapter 6 presents the conclusion of the research study and the future directions for further improvements.

REFERENCES

Agarwal, R. C., Aggarwal, C. C., and Prasad, V. (2001). A tree projection algorithm for generation of frequent item sets. *Journal of parallel and Distributed Computing*, *61*(3), 350-371.

Agrawal, R., Imieliński, T., and Swami, A. (1993). *Mining association rules between sets of items in large databases.* Paper presented at the Acm sigmod record, 207-216.

Agrawal, R., and Srikant, R. (1994). *Fast algorithms for mining association rules.* Paper presented at the Proc. 20th int. conf. very large data bases, VLDB, 487-499.

Ait-Mlouk, A., Agouti, T., and Gharnati, F. (2017). Mining and prioritization of association rules for big data: multi-criteria decision analysis approach. *Journal of Big Data*, *4*(1), 1-21.

Al-Opahi, A. A.-W., and Al-Maqaleh, B. M. (2015). An Efficient Algorithm to Automated Discovery of Interesting Positive and Negative Association Rules. *IJACSA) International Journal of Advanced Computer Science and Applications*, *6*(6).

AL-Zawaidah, F. H., Jbara, Y. H., and Marwan, A. (2011). An improved algorithm for mining association rules in large databases. *World of Computer science and information technology journal*, *1*(7), 311-316.

Alatas, B. (2012). A novel chemistry based metaheuristic optimization method for mining of classification rules. *Expert Systems with Applications*, *39*(12), 11080-11088.

Aliberti, G., Colantonio, A., Di Pietro, R., and Mariani, R. (2015). EXPEDITE: EXPress closED ITemset enumeration. *Expert Systems with Applications*, *42*(8), 3933-3944.

Alwidian, J., Hammo, B. H., and Obeid, N. (2018). WCBA: Weighted classification based on association rules algorithm for breast cancer disease. *Applied Soft Computing*, *62*, 536-549.

Antonie, M.-L., and Zaïane, O. R. (2004). *Mining positive and negative association rules: An approach for confined rules.* Paper presented at the European Conference on Principles of Data Mining and Knowledge Discovery, 27-38.

Bache, K., and Lichman. (2013). UCI: Machine Learning Repository.

Berzal, F., Blanco, I., Sánchez, D., and Vila, M.-A. (2002). Measuring the accuracy and interest of association rules: A new framework. *Intelligent Data Analysis, 6*(3), 221-235.

Berzal, F., Cubero, J.-C., Marín, N., Sánchez, D., Serrano, J., and Vila, A. (2005). Association rule evaluation for classification purposes. *Actas del III Taller Nacional de Mineria de Datos y Aprendizaje*, 135-144.

Bodon, F. (2003). *A fast APRIORI implementation.* Paper presented at the FIMI, 63.

Borgelt, C. (2012). Frequent item set mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2*(6), 437-456.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997a). *Dynamic itemset counting and implication rules for market basket data.* Paper presented at the ACM SIGMOD Record, 255-264.

Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997b). *Dynamic itemset counting and implication rules for market basket data.* Paper presented at the Proceedings of the 1997 ACM SIGMOD international conference on Management of data, 255-264.

Cao, L., Dong, X., and Zheng, Z. (2016). e-NSP: Efficient negative sequential pattern mining. *Artificial Intelligence, 235*, 156-182.

Ceruto, T., Lapeira, O., Tonch, A., Plant, C., Espin, R., and Rosete, A. (2014). *Mining medical data to obtain fuzzy predicates.* Paper presented at the International Conference on Information Technology in Bio-and Medical Informatics, 103-117.

Chen, C.-H., Lan, G.-C., Hong, T.-P., and Lin, Y.-K. (2013a). Mining high coherent association rules with consideration of support measure. *Expert systems with applications, 40*(16), 6531-6537.

Chen, C.-H., Li, A.-F., and Lee, Y.-C. (2013b). A fuzzy coherent rule mining algorithm. *Applied Soft Computing, 13*(7), 3422-3428.

Chen, C.-H., Li, A.-F., and Lee, Y.-C. (2014). Actionable high-coherent-utility fuzzy itemset mining. *Soft Computing, 18*(12), 2413-2424.

Chen, J., and Gao, Y. (2009). Evaluation criterion for association rules with influence degree. *Computer Engineering and Application, 45*(8), 141-143.

Chen, M.-C. (2007). Ranking discovered rules from data mining with multiple criteria by data envelopment analysis. *Expert Systems with Applications, 33*(4), 1110-1116.

Chiclana, F., Kumar, R., Mittal, M., Khari, M., Chatterjee, J. M., and Baik, S. W. (2018). ARM–AMO: an efficient association rule mining algorithm based on animal migration optimization. *Knowledge-Based Systems, 154*, 68-80.

Dash, S. R., Dehuri, S., and Rayaguru, S. (2013). *Discovering interesting rules from biological data using parallel genetic algorithm*. Paper presented at the 2013 3rd IEEE International Advance Computing Conference (IACC), 631-636.

Datta, S., and Mali, K. (2017). *Trust: a new objective measure for symmetric association rule mining in account of dissociation and null transaction.* Paper presented at the 2016 Eighth International Conference on Advanced Computing (ICoAC), 151-156.

De Cock, M., Cornelis, C., and Kerre, E. E. (2005). Elicitation of fuzzy association rules from positive and negative examples. *Fuzzy Sets and Systems, 149*(1), 73-85.

de Sá, C. R., Azevedo, P., Soares, C., Jorge, A. M., and Knobbe, A. (2018). Preference rules for label ranking: Mining patterns in multi-target relations. *Information Fusion, 40*, 112-125.

Dechang, P., and Xiaolin, Q. (2008). A new fuzzy clustering algorithm on association rules for knowledge management. *Information Technology Journal, 7*(1), 119-124.

Delgado, M., Ruiz, M. D., Sanchez, D., and Serrano, J. M. (2011). *A fuzzy rule mining approach involving absent items.* Paper presented at the Proceedings of the 7th Conference of the European Society for Fuzzy Logic and Technology, 275-282.

Dharshinni, N., Azmi, F., Fawwaz, I., Husein, A., and Siregar, S. D. (2019). *Analysis of accuracy K-means and apriori algorithms for patient data clusters.* Paper presented at the Journal of Physics: Conference Series, 012020.

Djenouri, Y., Drias, H., and Chemchem, A. (2013). *A hybrid bees swarm optimization and tabu search algorithm for association rule mining.* Paper presented at the

2013 World Congress on Nature and Biologically Inspired Computing, 120-125.

Dong, X., Gong, Y., and Cao, L. (2018). F-NSP+: A fast negative sequential patterns mining method with self-adaptive data storage. *Pattern Recognition, 84*, 13-27.

Dong, X., Hao, F., Zhao, L., and Xu, T. (2020). An efficient method for pruning redundant negative and positive association rules. *Neurocomputing, 393*, 245-258.

El-Hajj, M., and Zaiane, O. R. (2003). *COFI-tree mining: a new approach to pattern growth with reduced candidacy generation.* Paper presented at the Workshop on Frequent Itemset Mining Implementations (FIMI'03) in conjunction with IEEE-ICDM.

Feddaoui, I., Felhi, F., and Akaichi, J. (2016). *EXTRACT: New extraction algorithm of association rules from frequent itemsets.* Paper presented at the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), 752-756.

Fournier-Viger, P., Wu, C.-W., and Tseng, V. S. (2014). *Novel concise representations of high utility itemsets using generator patterns.* Paper presented at the International Conference on Advanced Data Mining and Applications, 30-43.

Fournier-Viger, P., Lin, J. C. W., Vo, B., Chi, T. T., Zhang, J., and Le, H. B. (2017). A survey of itemset mining. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*.

Geng, L., and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR), 38*(3), 9-es.

Gopalan, R. P., and Sucahyo, Y. G. (2002). *ITL-MINE: Mining Frequent Itemsets More Efficiently*. Paper presented at the FSKD, 167-.

Goyal, V., Sureka, A., and Patel, D. (2015). *Efficient skyline itemsets mining.* Paper presented at the Proceedings of the Eighth International C* Conference on Computer Science & Software Engineering, 119-124.

Grahne, G., and Zhu, J. (2003). *Efficiently using prefix-trees in mining frequent itemsets.* Paper presented at the FIMI.

Guo, H., and Zhou, Y. (2009). *An algorithm for mining association rules based on improved genetic algorithm and its application.* Paper presented at the 2009

Third International Conference on Genetic and Evolutionary Computing, 117-120.

Guo, Y., Wang, M., and Li, X. (2017). Application of an improved Apriori algorithm in a mobile e-commerce recommendation system. *Industrial Management & Data Systems, 117*(2), 287-303.

Han, J., and Kamber, M. (2001). Data mining concept and technology. *Publishing House of Mechanism Industry*, 70-72.

Han, J., and Pei, J. (2000). Mining frequent patterns by pattern-growth: methodology and implications. *SIGKDD Explorations, 2*(2), 14-20.

Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*: Elsevier.

Han, J., Pei, J., Yin, Y., and Mao, R. (2004). Mining frequent patterns without candidate generation: A frequent-pattern tree approach. *Data mining and knowledge discovery, 8*(1), 53-87.

Hart, P. E. (1973). *Pattern classification and scene analysis*.

Hong, Z., and Bian, F. (2008). *A heuristic approach for fast mining association rules in transportation system.* Paper presented at the 2008 Fifth International Conference on Fuzzy Systems and Knowledge Discovery, 541-545.

Hoseini, M. S., Shahraki, M. N., and Neysiani, B. S. (2015). *A new algorithm for mining frequent patterns in Can Tree.* Paper presented at the 2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI), 843-846.

Hu, Y.-H., and 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), 1-24.

Huang, C., Lu, R., and Choo, K.-K. R. (2017). Secure and flexible cloud-assisted association rule mining over horizontally partitioned databases. *Journal of Computer and System Sciences, 89*, 51-63.

Huang, H. (2020). Research on feature classification method of network text data based on association rules. *International Journal of Computers and Applications, 42*(2), 157-163.

Hurley, P. (2014). *A concise introduction to logic*: Nelson Education.

Ikram, A., and Qamar, U. (2015). Developing an expert system based on association rules and predicate logic for earthquake prediction. *Knowledge-Based Systems, 75*, 87-103.

Jia, X.-b., Jin, Y., Li, N., Su, X., Cardiff, B., and Bhanu, B. (2018). Words alignment based on association rules for cross-domain sentiment classification. *Frontiers of Information Technology & Electronic Engineering, 19*(2), 260-272.

Ju, C., Bao, F., Xu, C., and Fu, X. (2015). A novel method of interestingness measures for association rules mining based on profit. *Discrete Dynamics in Nature and Society, 2015*.

Kane, C. (2018). *Classification using Association Rules.* Unpublished Dissertation, Technological University Dublin, Dublin.

Kaur, M., and Kang, S. (2016). Market Basket Analysis: Identify the changing trends of market data using association rule mining. *Procedia computer science, 85*, 78-85.

Kaya, M., and Alhajj, R. (2004). *Mining multi-cross-level fuzzy weighted association rules.* Paper presented at the 2004 2nd International IEEE Conference on'Intelligent Systems'. Proceedings (IEEE Cat. No. 04EX791), 225-230.

Kim, D., and Yun, U. (2016). Efficient mining of high utility pattern with considering of rarity and length. *Applied Intelligence, 45*(1), 152-173.

Kiran, R. U., and Reddy, P. K. (2011). *Novel techniques to reduce search space in multiple minimum supports-based frequent pattern mining algorithms.* Paper presented at the Proceedings of the 14th international conference on extending database technology, 11-20.

Koh, Y. S., and Ravana, S. D. (2016). Unsupervised rare pattern mining: a survey. *ACM Transactions on Knowledge Discovery from Data (TKDD), 10*(4), 45.

Koh, Y. S., Rountree, N., and O'Keefe, R. A. (2006). *Mining Interesting Imperfectly Sporadic Rules.* Paper presented at the PAKDD, 473-482.

Koh, Y. S., Rountree, N., and O’Keefe, R. A. (2008). Mining interesting imperfectly sporadic rules. *Knowledge and Information Systems, 14*(2), 179-196.

Kuo, R., and Shih, C. (2007). Association rule mining through the ant colony system for National Health Insurance Research Database in Taiwan. *Computers & Mathematics with Applications, 54*(11-12), 1303-1318.

Kuoa, R. J., Chao, C. M., and Chiu, Y. (2011). Application of particle swarm optimization to association rule mining. *Applied soft computing, 11*(1), 326-336.

Lin, J. C.-W., Gan, W., Fournier-Viger, P., and Hong, T.-P. (2015). *Mining high-utility itemsets with multiple minimum utility thresholds.* Paper presented at the Proceedings of the Eighth International C* Conference on Computer Science & Software Engineering, 9-17.

Lin, K.-C., Liao, I.-E., and Chen, Z.-S. (2011). An improved frequent pattern growth method for mining association rules. *Expert Systems with Applications, 38*(5), 5154-5161.

Ling, C. X., Chen, T., Yang, Q., and Cheng, J. (2002). *Mining optimal actions for profitable CRM.* Paper presented at the 2002 IEEE International Conference on Data Mining, 2002. Proceedings., 767-770.

Liu, B., Hsu, W., and Ma, Y. (1999a). *Mining association rules with multiple minimum supports.* Paper presented at the Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining, 337-341.

Liu, B., Hsu, W., Mun, L.-F., and Lee, H.-Y. (1999b). Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering, 11*(6), 817-832.

Lucchese, C., Orlando, S., and Perego, R. (2006). Fast and memory efficient mining of frequent closed itemsets. *IEEE Transactions on Knowledge and Data Engineering, 18*(1), 21-36.

Mahmood, S., Shahbaz, M., and Guergachi, A. (2014). Negative and positive association rules mining from text using frequent and infrequent itemsets. *The Scientific World Journal, 2014*.

Makino, Y., Kato, K., and Tanimoto, S. (2017). *The unpredictable structure of risk chains using association rule mining.* Paper presented at the Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2017 18th IEEE/ACIS International Conference on, 507-511.

Mattiev, J., and Kavsek, B. (2020). Coverage-based classification using association rule mining. *Applied Sciences, 10*(20), 7013.

Minaei-Bidgoli, B., Barmaki, R., and Nasiri, M. (2013). Mining numerical association rules via multi-objective genetic algorithms. *Information Sciences, 233*, 15-24.

Nanavati, A. A., Chitrapura, K. P., Joshi, S., and Krishnapuram, R. (2001). *Mining generalised disjunctive association rules.* Paper presented at the Proceedings of the tenth international conference on Information and knowledge management, 482-489.

Nandhini, M., Janani, M., and Sivanandham, S. (2012). *Association rule mining using swarm intelligence and domain ontology*. Paper presented at the 2012 International Conference on Recent Trends in Information Technology, 537-541.

Narvekar, M., and Syed, S. F. (2015). An optimized algorithm for association rule mining using FP tree. *Procedia Computer Science*, *45*, 101-110.

Ng, R. T., Lakshmanan, L. V., Han, J., and Pang, A. (1998). Exploratory mining and pruning optimizations of constrained associations rules. *ACM Sigmod Record, 27*(2), 13-24.

Padmanabhan, B., and Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems, 27*(3), 303-318.

Páircéir, R., McClean, S., and Scotney, B. (2000). *Discovery of multi-level rules and exceptions from a distributed database.* Paper presented at the Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, 523-532.

Pandey, S. C., and Nandi, G. C. (2014). TSD based framework for mining the induction rules. *Journal of Computational Science*, *5*(2), 184-195.

Park, J. S., Chen, M.-S., and Yu, P. S. (1995). *An effective hash-based algorithm for mining association rules* (Vol. 24): ACM.

Pei, J., Han, J., Lu, H., Nishio, S., Tang, S., and Yang, D. (2001). *H-mine: Hyper-structure mining of frequent patterns in large databases.* Paper presented at the Data Mining, 2001. ICDM 2001, Proceedings IEEE International Conference on, 441-448.

Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. *Knowledge discovery in databases*, 229-238.

Pijls, W., and Bioch, J. C. (2000). Mining frequent intemsets in memory-resident databases. *ERIM Report Series Reference No. ERS-2000-53-LIS.*

Pyun, G., Yun, U., and Ryu, K. H. (2014). Efficient frequent pattern mining based on linear prefix tree. *Knowledge-Based Systems, 55*, 125-139.

Rathee, S., Kaul, M., and Kashyap, A. (2015). *R-Apriori: an efficient apriori based algorithm on spark.* Paper presented at the Proceedings of the 8th Workshop on Ph. D. Workshop in Information and Knowledge Management, 27-34.

Savasere, A., Omiecinski, E., and Navathe, S. (1998). *Mining for strong negative associations in a large database of customer transactions.* Paper presented at the Proceedings 14th International Conference on Data Engineering, 494-502.

Shana, J., and Venkatachalam, T. (2015). An Improved Method for Counting Frequent Itemsets Using Bloom Filter. *Procedia Computer Science, 47*, 84-91.

Silberschatz, A., and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and data engineering, 8*(6), 970-974.

Sim, A. T. H. (2009). *Discovering of Association Rules Without a Minimum Support Threshold-Coherent Rules Discovery*: Monash University.

Sim, A. T. H., Indrawan, M., Zutshi, S., and Srinivasan, B. (2010). Logic-based pattern discovery. *IEEE Transactions on Knowledge and Data Engineering, 22*(6), 798-811.

Singh, J., Ram, H., and Sodhi, D. J. (2013). Improving efficiency of apriori algorithm using transaction reduction. *International Journal of Scientific and Research Publications, 3*(1), 1-4.

Song, M., and Rajasekaran, S. (2006). A transaction mapping algorithm for frequent itemsets mining. *IEEE transactions on knowledge and data engineering, 18*(4), 472-481.

Soulet, A., and Rioult, F. (2014). *Efficiently depth-first minimal pattern mining.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 28-39.

Soysal, Ö. M. (2015). Association rule mining with mostly associated sequential patterns. *Expert Systems with applications, 42*(5), 2582-2592.

Srikant, R., Vu, Q., and Agrawal, R. (1997). *Mining association rules with item constraints.* Paper presented at the Kdd, 67-73.

Sucahyo, Y. G., and Gopalan, R. P. (2003). *CT-ITL: Efficient frequent item set mining using a compressed prefix tree with pattern growth.* Paper presented at the Proceedings of the 14th Australasian database conference-Volume 17, 95-104.

Szathmary, L., Napoli, A., and Valtchev, P. (2007). *Towards rare itemset mining.* Paper presented at the Tools with Artificial Intelligence, 2007. ICTAI 2007. 19th IEEE International Conference on, 305-312.

Szathmary, L., Valtchev, P., Napoli, A., and Godin, R. (2012). *Efficient Vertical Mining of Minimal Rare Itemsets.* Paper presented at the CLA, 269-280.

Szathmary, L., Valtchev, P., Napoli, A., Godin, R., Boc, A., and Makarenkov, V. (2014). A fast compound algorithm for mining generators, closed itemsets, and computing links between equivalence classes. *Annals of Mathematics and Artificial Intelligence, 70,* 81-105.

Tan, P.-N., Steinbach, M., and Kumar, V. (2016). *Introduction to data mining*: Pearson Education India.

Toloo, M., Tavana, M., and Santos-Arteaga, F. J. (2019). An integrated data envelopment analysis and mixed integer non-linear programming model for linearizing the common set of weights. *Central European Journal of Operations Research*, *27*(4), 887-904.

Tran, A. N., Duong, H. V., Truong, T. C., and Le, B. H. (2011). *Efficient algorithms for mining frequent itemsets with constraint.* Paper presented at the 2011 Third International Conference on Knowledge and Systems Engineering, 19-25.

Troiano, L., and Scibelli, G. (2014). A time-efficient breadth-first level-wise lattice-traversal algorithm to discover rare itemsets. *Data Mining and Knowledge Discovery, 28*(3), 773-807.

Uno, T., Kiyomi, M., and Arimura, H. (2004). *2: Efficient Mining Algorithms for Frequent/Closed/Maximal Itemsets.* Paper presented at the Proc. IEEE ICDM Workshop Frequent Itemset Mining Implementations.

Vaithiyanathan, V., Rajeswari, K., Phalnikar, R., and Tonge, S. (2012). *Improved apriori algorithm based on selection criterion.* Paper presented at the 2012 IEEE International Conference on Computational Intelligence and Computing Research, 1-4.

Vijayalakshmi, V., and Pethalakshmi, A. (2015). An efficient count based transaction reduction approach for mining frequent patterns. *Procedia Computer Science, 47,* 52-61.

Vo, B., Hong, T.-P., and Le, B. (2012). DBV-Miner: A Dynamic Bit-Vector approach for fast mining frequent closed itemsets. *Expert Systems with Applications, 39*(8), 7196-7206.

Wang, C., and Tjortjis, C. (2004). *PRICES: An efficient algorithm for mining association rules.* Paper presented at the International Conference on Intelligent Data Engineering and Automated Learning, 352-358.

Wang, K., Tang, L., Han, J., and Liu, J. (2002). *Top down fp-growth for association rule mining.* Paper presented at the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 334-340.

Webb, G. I., and Vreeken, J. (2014). Efficient discovery of the most interesting associations. *ACM Transactions on Knowledge Discovery from Data (TKDD), 8*(3), 15.

Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*: Morgan Kaufmann.

Wu, L., Gong, K., He, Y., Ge, X., and Cui, J. (2010). *A study of improving apriori algorithm.* Paper presented at the 2010 2nd International Workshop on Intelligent Systems and Applications, 1-4.

Wu, X., Zhang, C., and Zhang, S. (2004). Efficient mining of both positive and negative association rules. *ACM Transactions on Information Systems (TOIS), 22*(3), 381-405.

Yan, X., Zhang, C., and Zhang, S. (2009). Genetic algorithm-based strategy for identifying association rules without specifying actual minimum support. *Expert Systems with Applications*, *36*(2), 3066-3076.

Yang, M. (2004). *Association rules data mining in manufacturing information system based on genetic algorithms.* Paper presented at the Proceedings. ICCEA 2004. 2004 3rd International Conference on Computational Electromagnetics and Its Applications, 2004., 153-156.

Yuan, J., and Ding, S. (2012). *Research and improvement on association rule algorithm based on FP-growth.* Paper presented at the International Conference on Web Information Systems and Mining, 306-313.

Yuan, X. (2017). *An improved Apriori algorithm for mining association rules.* Paper presented at the AIP conference proceedings, 080005.

Yuan, X., Buckles, B. P., Yuan, Z., and Zhang, J. (2002). *Mining negative association rules.* Paper presented at the Proceedings ISCC 2002 Seventh International Symposium on Computers and Communications, 623-628.

Yuan, Y., and Huang, T. (2005). *A matrix algorithm for mining association rules.* Paper presented at the International Conference on Intelligent Computing, 370-379.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika, 2*(2), 121-134.

Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering, 12*(3), 372-390.

Zuo, J., Tang, C., and Zhang, T. (2002). *Mining predicate association rule by gene expression programming.* Paper presented at the International Conference on Web-Age Information Management, 92-103.

LIST OF PUBLICATIONS

Journal with Impact Factor

1. **Ahmad HI**, Sim AT, Shah SM, Abrar M, Gul A. Mining predicate rules without minimum support threshold. Kuwait Journal of Science. 2021 Aug 16;48(4). **(Q2, IF:0.948)**

Indexed Conference Proceedings

1. **Ahmad, H.I.**, Sim, A.T.H. & Hee, J.M. (2017). A metamodel for tourist intentions to visit a destination using systematic literature review. In *2017 International Conference on Research and Innovation in Information Systems (ICRIIS)* (pp. 1-6). IEEE. https://doi.org/10.1109/ICRIIS.2017.8002515. **(Indexed by SCOPUS)**