Parallel Association Rule Mining Algorithm Based on MapReduce by Using Lift Interestingness Measure for Big Data

Nour Easa Oweis

Field of study: Computer Science, Communication Technology and Applied Mathematics.
Supervisor: Prof. RNDr.Václav Snášel, CSc.

Ostrava, 2016
Abstract

**Background:** Big Data mining is an analytic process utilized to discover the hidden knowledge and patterns from a massive, complex, and multidimensional dataset. Single processors memory and CPU resources are very limited in this aspect, which makes the algorithm performance ineffective. Association rule mining (ARM) is traditionally used to uncover hidden knowledge in data sets. However, they were unable to handle huge big data sets. Therefore, scalable and parallel strategies for ARM based on Big Data approaches are needed. Example of this approach is parallel association rule mining algorithm based on MapReduce by using lift interestingness measure (LIM)

**Methods:** This thesis proposes two algorithms for data mining and optimization. The first is parallel association rule mining algorithm based on MapReduce by using LIM (MapReduce Lift Association Rule (MRLAR)), to provide high scalability over parallel execution. The second is reduce dimensionality by using multiple data reduction techniques including principle component analysis (PCA), singular value decomposition (SDD), semi-discrete decomposition (SVD), applied to reduce the data into fewer dimensions as pre-processing techniques for data optimization.

**Results:** The MRLAR was found to directly extract the association rule and type of correlation between Lift Hand Side (LHS) and Right Hand Side (RHS) in the ARM (Lift) without the need for additional computation on the confidence measure. It also provided the following advantages: High scalability by utilizing parallel execution (MapReduce), support big data, one scan dataset, no more post-processing techniques and fault tolerance. The study also proposed an algorithm for data reduction using PCA, SVD, and SDD. The SVD was also found to have better accuracy and less time execution than SDD.

**Conclusions:** The MRLAR performed effectively in data mining. The data reduction techniques enhanced the pre-processing of data by dimensionality reduction.

**Keywords:** Big Data; Data Mining; Association Rule; MapReduce; Lift Interesting Measurement; Data Reduction.
# TABLE OF CONTENTS

Table of Contents .................................................................................................................. 3

List of Figures .......................................................................................................................... 5

List of Tables ............................................................................................................................ 6

List of Abbreviations ............................................................................................................... 7

1 Chapter One: Introduction ..................................................................................................... 8
   1.1 Topics of research ............................................................................................................ 9
   1.2 Motivation and objectives of this thesis ......................................................................... 10
   1.3 Thesis content ................................................................................................................ 11

2 Association Rule Mining Algorithm ...................................................................................... 12
   2.1 A Novel MapReduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data.... 12
      2.1.1 Prototype Implementation ....................................................................................... 12
      2.1.2 Dataset Collection ................................................................................................. 13
      2.1.3 Software Used ......................................................................................................... 14
      2.1.4 Proposed Algorithm (MRLAR) ............................................................................... 14
      2.1.5 Experimental Results and Performance Evaluation .............................................. 18
      2.1.6 Data Processing Graphic User Interface ............................................................... 23

3 Dimensionality Reduction Algorithms .................................................................................. 27
   3.1 Feature Selection by Principle Component Analysis (FSPCA) for Mining Frequent
       Association Rules ............................................................................................................ 27
      3.1.1 Dataset Collection and Conversion ...................................................................... 27
      3.1.2 Software Used ........................................................................................................ 27
      3.1.3 Proposed Algorithm (FSPCA) ............................................................................... 27
      3.1.4 Experiment Results ............................................................................................... 28
3.2 Feature Selection Using Semi-Discrete Decomposition (SDD) and Singular Value Decomposition (SVD) ........................................................................................................................................................................ 30

3.2.1 Dataset collection and preprocessing ................................................................................................................. 30

3.2.2 Software Used .......................................................................................................................................................... 30

3.2.3 Proposed Algorithm ............................................................................................................................................... 30

3.2.4 Experimental Results ........................................................................................................................................... 32

a) Applying Semi-Discrete Decomposition (SDD) ...................................................................................................... 32

b) Applying Singular Value Decomposition (SVD) .................................................................................................... 32

3.3 Summary and Wrap-up of the thesis work: ............................................................................................................. 34

4 Conclusions ............................................................................................................................................................... 38

4.1 Limitation of the Study and Directions for Future work ......................................................................................... 39

References .................................................................................................................................................................... 41

List of Student’s Publications .................................................................................................................................... 44

Published Papers: ....................................................................................................................................................... 44

a) Related to the topics: ............................................................................................................................................... 44

b) Unrelated to the Topics: ...................................................................................................................................... 45
LIST OF FIGURES

Figure 2-1: Illustration of the MapReduce-Based for Lift Association Rule (MRLAR) .................. 13
Figure 2-2: Illustration of the small capture of the original airport dataset ..................................... 13
Figure 2-3: Illustration of the MRLAR flowchart ........................................................................ 16
Figure 2-4: Illustration of the Mapper flowchart ......................................................................... 17
Figure 2-5: Illustration of the Reducer flowchart ....................................................................... 18
Figure 2-6: Illustration of the single dimension ARM pie chart diagram for ID numbers 1 and 2 ..... 20
Figure 2-7: Illustration of the single dimension ARM pie chart diagram for ID numbers 3 and 4 ... 20
Figure 2-8: Illustration of the multi-dimensional ARM pie chart diagram for the arrival delay and October ID numbers 3 and 4 .................................................................................. 22
Figure 2-9: Illustration of the multi-dimensional ARM pie chart diagram for the departure delay and June ID numbers 4 and 5 ......................................................................................... 22
Figure 2-10: Illustration of the data processing GUI ................................................................. 24
Figure 2-11: Illustration of the results of data processing GUI .................................................... 25
Figure 2-12: Illustration of the execution time (Secs) for MRLAR, Apriori, and LBA. ................. 26
Figure 3-3: Illustration of the association rules extracted by PCA ............................................... 29
Figure 3-4: Illustration of the truncated singular value decomposition “rank-k” ......................... 31
LIST OF TABLES

Table 1: Illustration of the unique digital code and string decoding processes. .......................... 15
Table 2: Illustration of the ARM with their support, confidence, and lift measurements with one attribute test (Arrival Delay or Departure Delay). .................................................................................................................. 19
Table 3: Illustration of the ARM with their support, confidence, and LIM with two attributes test (Arrival Delay, Departure Delay, and month of the year). .................................................................................................................. 21
Table 4: Illustration of the comparison between (single-dimension and multi-dimension ARM for LIM). .............................................................................................................................................. 23
Table 5: Illustration of the execution time for MRLAR, Apriori, and LBA. ...................................... 26
Table 6: Illustration of the association rules extracted by PCA .......................................................... 29
Table 7: Illustration of the Semi-Discrete Decomposition experimental result. .............................. 32
Table 8: Illustration of the Singular Value Decomposition experimental result. ............................ 33
## List of Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARM</td>
<td>Association Rule Mining</td>
</tr>
<tr>
<td>ETS</td>
<td>Exponential Threshold Strategy</td>
</tr>
<tr>
<td>FSPCA</td>
<td>Feature Selection by Principle Component Analysis</td>
</tr>
<tr>
<td>GUI</td>
<td>Graphical User Interface</td>
</tr>
<tr>
<td>IoT</td>
<td>Internet of Things</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge Data Discovery</td>
</tr>
<tr>
<td>LBA</td>
<td>Lift Base Algorithm</td>
</tr>
<tr>
<td>LHS</td>
<td>Lift Hand Side</td>
</tr>
<tr>
<td>LIM</td>
<td>Lift interestingness measurement</td>
</tr>
<tr>
<td>MRLAR</td>
<td>MapReduce Lift Association Rule</td>
</tr>
<tr>
<td>NFC</td>
<td>Near Field Communication</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>RFID</td>
<td>Radio Frequency Identification</td>
</tr>
<tr>
<td>RHS</td>
<td>Right Hand Side</td>
</tr>
<tr>
<td>SDD</td>
<td>Semi-Discrete Decomposition</td>
</tr>
<tr>
<td>SVD</td>
<td>Singular Value Decomposition</td>
</tr>
<tr>
<td>WSN</td>
<td>Wireless Sensor Network</td>
</tr>
</tbody>
</table>
1 Chapter One: Introduction

The recent advances in computers and communications have dramatically increased the number of associated relevant applications, such as Radio Frequency Identification Devices (RFID), Wireless Sensor Networks (WSNs), Internet of Things (IoT), and many others [1, 2, 20]. These applications create a huge stream of non-stop data that is currently well-known as big data, denoted as “Big Data” [3]. Big Data is a massive set of data that is too complex to be managed by traditional applications. Nowadays, it includes huge, complex, and abundant structured, semi-structured, and unstructured data as well as hidden data that are generated and gathered from several fields and sources [4]. There are many challenges to managing such complex sets of Big Data. These challenges include, but not limited to, extracting, analyzing, visualizing, sharing, storage, transferring and searching the “Big Data” [5].

Big Data are stored in powerful computers and include many hidden patterns’ indicators that may help in decision making. Data mining approaches have been introduced to facilitate decision making through determining and explanation of those patterns in a meaningful knowledge format [6]. The traditional data processing approaches and its applications cannot be directly implemented when working with the big data management [7]. Therefore, it is necessary to apply new and innovative techniques, algorithms, and frameworks in order to be able to manage, extract, and execute the big data mining approaches. This, in turn, will make these data mining techniques very helpful and more efficient. Frequent pattern mining is one of the well-known data mining techniques that focuses on discovering a number of interesting patterns from a large set of data items [8].

The association rule is a frequent pattern mining method that is applied to find or discover all the frequent co-occurrence relationships between variables in a large set of transactions [9]. The association rule strength is measured through two parameters: Support and Confidence. However, these two parameters may not be sufficient to discover some interesting patterns, thus, other measuring criteria are used which is the Lift Interestingness Measure, commonly known as the “Lift” [10]. Another approach that has been suggested to process large data sets is the MapReduce. MapReduce is usually proposed for parallel processing of large datasets.

This thesis investigates the efficiency of integrating both the Association Rule and the MapReduce approaches in processing large data sets in an innovative way. It proposes a parallel-based MapReduce approach that has been used for defining the association rule importance based on the Lift Interestingness Measurement (LIM). These objectives are addressed in the first part of this thesis work.

Nowadays, data mining techniques have to deal with a large number of features associated with high-dimensional settings with multiple types of data. There is a fear that due to the high dimensionality
of Big Data, there will be relations that are irrelevant or redundant among the studied variables in the analyzed set of data. This effect cases what is known as “the curse of dimensionality” [11]. Therefore, it is of paramount importance to remove any irrelevant or abundant features in order to achieve a more understandable model with easy visualization of the data [12].

The large size of data makes the extraction of association rules a hard task [13]. Dimension reduction methods based on multiple methods such as Singular Value Decomposition (SVD) [14], Semi-Discreet Decomposition (SDD), and Principle Component Analysis (PCA) [15,16] are widely used in several data mining aspects within different topics such as signal and image processing, social network analysis, healthcare and many more [13]. In this thesis as a second part, we propose a new method for dimensionality reduction and feature selection based on SVD, SDD, and PCA as a preprocessing technique for data optimization.

Experimental results of this thesis reveal that the parallel association rule mining algorithm based on MapReduce by using lift interestingness measure for Big Data was effective in data mining. It also showed that the reduction techniques, particularly the SVD, can discover the same rules obtained by the original data with high accuracy.

1.1 TOPICS OF RESEARCH

This section presents the main topics of this study. In this thesis, two main topics are examined: Data correlation and handling, and data processing and optimization. Each of them focuses on a different part of the main problem.

- A Novel MapReduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data:
  In this approach, the data mining technique used is the Lift Interestingness Measure (LIM) of the association rule. This rule was applied to find the correlation between the data samples (rows) and their parameters (columns). MapReduce process was added and implemented to handle the increased memory size needed to handle big data. The results achieved in combining these techniques (Lift and MapReduce) showed the effectiveness of this algorithm. The results of this part of the thesis have been published in an international peer-reviewed journal. [17]. This dissertation presents the details of the handling of the LIM algorithm and its application through the MapReduce functionality.

- Data Processing and Optimization:
  Several methods of data reduction and correlation techniques have been used to handle big data. A key benefit of using techniques of data reduction is to simplify the visualization of data and to eliminate irrelevant features and noise reduction. This dissertation examined few of these
methods such as; SVD, SSD, and PCA. Such methods were used as pre-processing and data reduction techniques. This thesis utilized mainly the SVD in the dimensionality reduction area. Moreover, other topics of my published research describe feature extraction and data reduction by the PCA. Another study reported a comprehensive comparison between the SVD and SSD and was accepted for publication in international conferences [18,19].

1.2 MOTIVATION AND OBJECTIVES OF THIS THESIS

In general, to discover relationships between data by using the association rule measurements, most of the algorithms use both confidence and support of the rules as interestingness measurements for association rules. However, in some cases, a number of patterns are discovered beyond human capabilities of determining the required results.

Based on the experimental results of this dissertation work, it was noted that the use of confidence is not effective enough to determine the association rules. This reason is that the confidence does not describe the type of correlation(s) between Left Hand Side (LHS) and Right Hand Side (RHS) in association rules. Therefore, this thesis uses LIM and integrates it with parallel algorithm under MapReduce approach instead of confidence as criteria for discovering meaningful association rules. The proposed algorithm is called “MapReduce-Based for Lift Association Rule” (MRLAR). In addition to the above criteria, the main positive characteristics of MRLAR are that it:

- Handles many tough big data problems, especially parallelization association rule, network communication and fault tolerance based on MapReduce approach.
- Once used, MRLAR eliminates the need for post-processing to identify rules of interest. Since this capability is integrated into the proposed MRLAR itself using parallel LIM instead of confidence, this solves the problem of confidence by directly extract association rule and type of correlation without using confidence).
- Does not require a pruning step. Pruning is usually required to make the number of candidate itemsets much smaller and then find the frequent patterns to generate an association rule.
- Faults tolerance feature (automatic data recovery based on MapReduce functionality).
- Solves two bottlenecks of the Apriori algorithm. The MRLAR is proposed to solve the following bottlenecks: First, the complex candidate generation process that uses most of the time, space and memory. And second, the multiple scans of the database.
- Requires less computational resources to produce a long sequence of items that convey potentially useful patterns.
Offers customizable options, based on previous choices collected from the user, MRLAR can narrow down the search space in order to extract only the association rules of interest to the user.

Improves the performance of Lift Base Algorithm (LBA) through parallel execution based on MapReduce approach. Details are presented in Chapter Two.

The second part of this thesis presents the implementation of a reduce dimensionality algorithm by providing new pre-processing techniques for finding the association rule. The proposed algorithms utilized a new method for dimensionality reduction and feature selection based on SVD, SDD, and PCA. Then find the association rules by using LIM Algorithm to reduce a high-dimensional dataset into fewer dimensions while retaining important information. This approach improves the performance and handles another big data mining problem.

1.3 THESIS CONTENT

Following the first chapter. This thesis is organized as follows:

- **Chapters Two and Three**: Covers the two proposed algorithms including the prototype implementation, methodologies, dataset collections, software used, algorithms step and experimental results. It also covered a proposal for a Graphic User Interface (GUI) used in the MapReduce, and finally presents a summary and wrap-up of the thesis work.

- **Chapter Four**: Finalized with the conclusions, limitations, and direction for future research, references, as well as the appendix including a list of the author publications.
2  ASSOCIATION RULE MINING ALGORITHM

2.1  A NOVEL MAPREDUCE LIFT ASSOCIATION RULE MINING ALGORITHM (MRLAR) FOR BIG DATA

This chapter describes the process used to find out the result of this thesis part. Section (2.1.1) defines the prototype implementation. Section (2.1.2) presents the dataset used. Section (2.1.3) describes the details of the software used. Section (2.1.4) displays the proposed algorithms steps, and the flowcharts of MRLAR. Section (2.1.5) illustrates the experimental results, and section (2.1.6) presents the data processing Graphic User Interface (GUI) and evaluation performance for MRLAR.

2.1.1  Prototype Implementation

This chapter describes the MapReduce approach applied for the Lift-Based Algorithm (LBA) [22].

This thesis developed a novel MapReduce approach for an association rule algorithm based on LIM which can handle a huge and complex data. This proposed algorithm has been entitled “MapReduce-Based for Lift Association Rule (MRLAR)”. The intention of this thesis was to improve the LBA through parallel executing. In other words, we worked to determine the type of correlation between LHS and RHS in parallel association rules mining. The MRLAR is illustrated in Figure (2-1) and consists of the following steps:

- **Map function:** This step combines two steps; the data splitting step and the Mapping step [23, 25].

  The splitting step performed to distribute the data across each separate Map nodes.

  The map step consists of map function that was established to find the association rules for some entities within a large sized database.

  The intended measurement is the LIM between some database key values, which can be selected by the user. Based on the choices collected from the user, MRLAR can narrow down the search space in order to extract only the association rules of interest.

- **Reduce function:** This step combined the outputs generated by each map node(s) to form the final collected association rule [25]. As mentioned previously, the rules are weighted by the reduce process not the confidence, but as an alternative to this parallel association rules, we utilized the Lift weight computation to define the correlation between LHS and RHS.
2.1.2 Dataset Collection

For this thesis, the algorithms used the dataset provided by the USA domestic airline flights between 1987 and 2008. The data comes originally from the Research and Innovative Technology Administration (RITA). This dataset is very large: there are nearly 120 million records in total, and takes up 1.6 gigabytes of space compressed and 12 gigabytes when uncompressed with a collection of records consisting of 29 variables of flight information for several airline carries, including arrival and departure times with CSV files format as shown in the small snippet in Figure (2-2) below. These files have derivable variables removed, are packaged in yearly chunks and have been more heavily compressed than the originals. The full dataset can be downloaded from (http://stat-computing.org/dataexpo/2009/).

A small subset of the dataset is also included with MATLAB libraries to allow you to run this and other examples without downloading the entire dataset.

*Figure 2-1: Illustration of the MapReduce-Based Lift Association Rule (MRLAR).*

*Figure 2-2: Illustration of the small capture of the original airport dataset.*
2.1.3 Software Used

For this part of the dissertation, the algorithms used the same MATLAB (2015a) because of its well-known ability to support big data enhancement especially by using MapReduce approach [21]. Currently, the execution environments in MATLAB support for MapReduce are serial (for local Desktop), the pool of parallel worker on Desktop (for local Desktop and multi-core), and Hadoop cluster, and you can easily run the same MapReduce algorithm in the different execution environments by minimal changes to your code [21]. In order to test the MRLAR algorithm, the experiment was done through this version of MATLAB (R2015a) as a high-level programming language on a pool of parallel on Desktop.

2.1.4 Proposed Algorithm (MRLAR)

As noted from the prototype implementation section, the MRLAR algorithm will use the LIM integrated with MapReduce approach to solves the problem of the impact LHS (left-hand side) on RHS (right-hand side) in parallel association rules to determine the type of data impact on each other. Thus, it classifies the correlations.

The MRLAR algorithm works as follows:

- **Pre-processing**: Once the data was gathered from the RITA site and before it underwent the decoding process, a pre-processing technique was used to clean it up in order to make sure the data goes to the “datastore” methods have been reviewed for the following: no empty cells, no Not a number (NaN) data, and no misrepresented data as strings which potentially could stop the execution of the program and/or results in wrong indication. The NaN and empty cells were assigned “zero” value.

- **Data Decoding**: Then the decoding process started for airport codes, which were identified as letters. Using MATLAB, we were able to find the unique list and assign a unique digital code to every element of the list as shown in Table (2) column A. Another way was used to decode the strings by assigning a digit from 1-26 to each letter in the alphabets A-Z as shown in Table (1) column B.
Table 1: Illustration of the unique digital code and string decoding processes.

<table>
<thead>
<tr>
<th>A. Unique Digital Code</th>
<th>B. String Decoding by Digit</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code</td>
<td>Decode</td>
</tr>
<tr>
<td>ABE</td>
<td>100</td>
</tr>
<tr>
<td>ABI</td>
<td>101</td>
</tr>
<tr>
<td>ABQ</td>
<td>102</td>
</tr>
<tr>
<td>ABY</td>
<td>103</td>
</tr>
<tr>
<td>ACK</td>
<td>104</td>
</tr>
<tr>
<td>ACT</td>
<td>105</td>
</tr>
<tr>
<td>ACV</td>
<td>106</td>
</tr>
<tr>
<td>ACY</td>
<td>107</td>
</tr>
<tr>
<td>ADQ</td>
<td>108</td>
</tr>
<tr>
<td>AEX</td>
<td>109</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Now, since we have a digital file with only strings as the column headers, we used the “datastore” to read the file in memory. After that, we scanned the file for duplicate entries using the command “unique” with which the data processing begin.

- **Initializing and key generation:** One of the most important reasons for designing a MapReduce algorithm based on LIM is to minimize the number of keys needed to be generated. This can be achieved by grouping the items by transactions, for example using “Day of Months” as a key from a collection of the tabular dataset.

- **MapReduce Functions:** MapReduce process is divided into map and reduce functions; their tasks are defined as follows: The map function does two parts; first, it reads a set of "records" from the database (file), searches and filters the selected set per the algorithm needs, and then outputs a set of records of the form (key, data). The second part is the split function which is a hash function, it splits the dataset into segments which are identified by a key. Once the segment is filled it get published to memory/disk, then a new segment is initialized. The MRLAR algorithm flowchart is illustrated in Figure (2-3).
In this dissertation, the map function as shown in Figure (2-4) was setup so that each segment of the data includes the following items in its key:

- **“Data Count”** this is the support count of the samples that support the test condition X, for example: “Arrival delay > \{30, 60, 90\}”.
- **“Data Set Count”** this is the support count of the samples that support the second test condition Y, for example: “Month is \{October or June\}”.
- **“ID”**: this is the segments ID.
The second part of the MapReduce is the execution of the reducer function. The reducer function processes the reduced segment as they are grouped according to their key value. In general, the reducer function can perform any operation on the data under the condition of all the reduced data be processed with the same reducer/hash function.

In this dissertation, the reducer function was setup to process as a cumulative sum of the segments Keys, see Figure (2-5).

In addition to this process, a third function (Result processing) was utilized to handle the final stages for both of map and reduce functions to find the Association Rule. This function handled the cumulative results and finds the Support, Confidence and Lift functions which are the main bone of the association rules. Then it handles the display of the final results.
2.1.5 Experimental Results and Performance Evaluation

a) Single-Dimension Association Rule

Traditionally, many of association rules use “Confidence” and “Support” measurements for testing the occurrence of relationships among variables of the dataset. However, it was found that relying solely on both of them may not be sufficient and meaningful. In this thesis part, a test was setup as shown in Table (2), for the association rules with their support, confidence, and lift measures with one attribute at a time (single-dimension association rule). The attributes that was tested (one at a time) were: the delay in flights arrival of more than 60 minutes or more than 120 minutes, as well as the delay in flights departure of more than 30, 40, or 90 minutes.

Table (2) shows the reliability of the relationship between support and confidence at the single dimensionality level confirmed also by the lift measurement. For example, the first two tests ID numbers (1) and (2) in Table (2) showed that while the support drops from 0.045 to 0.013 the confidence also followed from 0.093 to 0.027 with a positive lift association’s value (22.3 and 75.95 respectively).
Another point which can be noted is that the confidence increased while the number of itemsets increased, indicating a positive proportional relationship between the confidence and the number of itemsets. Please see Figure (2-6) for details.

The same results can be seen in ID numbers (3) and (4) with the same trend between confidence and support as well as the lift interestingness measure, see Figure (2-7). In Summary, the confidence and support measures are still valid and reliable measures of association at the single dimensionality rule. This was confirmed also by the use of the lift measure.

*Table 2: Illustration of the ARM with their support, confidence, and lift measurements with one attribute test (Arrival Delay or Departure Delay).*

<table>
<thead>
<tr>
<th>a- Arrival Delay Cases (in minutes)</th>
<th>ID</th>
<th>Rule</th>
<th>Itemsets</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>&gt; 60</td>
<td>2946</td>
<td>0.045</td>
<td>0.093</td>
<td>22.3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>&gt; 120</td>
<td>865</td>
<td>0.013</td>
<td>0.027</td>
<td>75.96</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>b- Departure Delay Cases (in minutes)</th>
<th>ID</th>
<th>Rule</th>
<th>Itemsets</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>3</td>
<td>&gt; 30</td>
<td>5684</td>
<td>0.087</td>
<td>0.212</td>
<td>11.5586</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>&gt; 40</td>
<td>4260</td>
<td>0.065</td>
<td>0.159</td>
<td>15.4223</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>&gt; 90</td>
<td>1358</td>
<td>0.021</td>
<td>0.051</td>
<td>48.3792</td>
</tr>
</tbody>
</table>
b) Multi-Dimension Association Rule

In order to achieve more efficiency with the high performance testing, parallel processing based on MapReduce was added in this study. The next experimental part of this dissertation integrated multiple attributes with the same dataset. The experiments have been performed with the same two columns of interest (arrival delay and departure delay) but combined them with another attribute that was the month of the year (June and October). Table (3) shows the attributes, with the Lift interestingness measures successfully being able to determine the type of correlation between itemsets (positively, negatively, and independent) using LHS and RHS compared to using support and confidence. Support and confidence did not survive the performance testing when it became to increased dimensionality of the itemsets. For example, Items (3) and (4) in Table (3) showed inconsistency in the
“support and confidence” trend. While the support has decreased in departure delay and October case (3) from (0.0057 to 0.0044) in item (4) (Table 3), the confidence has increased between these two cases (0.066 to 0.067), see Figure (2-8) for more details. The same results can be seen in departure delay and June cases ID numbers (4 and 5) with the same trend between confidence and support positive correlation by using lift measure, see Figure (2-9). Hence, the use of “support and confidence” measures only can survive under the single dimension association rule, but this is not applicable at the multidimensional level. Therefore, there was a need for a novel measure that is applicable at the multidimensional level, which is the lift measure in this thesis work. It is important to state that relying on the lift adds benefits to the prediction process of the future consequence in future datasets with comparing to the current data.

Table 3: Illustration of the ARM with their support, confidence, and LIM with two attributes test (Arrival Delay, Departure Delay, and month of the year).

<table>
<thead>
<tr>
<th>a- Arrival Delay and October Cases</th>
<th>ID</th>
<th>Rule</th>
<th>Itemsets</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>&gt; 60 &amp; Oct.</td>
<td>182</td>
<td>0.0028</td>
<td>0.0618</td>
<td>0.6678</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>&gt; 120 &amp; Oct.</td>
<td>54</td>
<td>0.0008</td>
<td>0.0624</td>
<td>0.6748</td>
</tr>
<tr>
<td>b- Departure Delay and October Cases</td>
<td>ID</td>
<td>Rule</td>
<td>Itemsets</td>
<td>Support</td>
<td>Confidence</td>
<td>Lift</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>&gt; 30 &amp; Oct.</td>
<td>376</td>
<td>0.0057</td>
<td>0.0662</td>
<td>0.715</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>&gt; 40 &amp; Oct.</td>
<td>287</td>
<td>0.0044</td>
<td>0.0674</td>
<td>0.7282</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>&gt; 90 &amp; Oct.</td>
<td>87</td>
<td>0.0013</td>
<td>0.0641</td>
<td>0.6925</td>
</tr>
<tr>
<td>c- Arrival Delay and June Cases</td>
<td>ID</td>
<td>Rule</td>
<td>Itemsets</td>
<td>Support</td>
<td>Confidence</td>
<td>Lift</td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>&gt; 60 &amp; Jun.</td>
<td>266</td>
<td>0.004</td>
<td>0.0903</td>
<td>1.1883</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>&gt; 120 &amp; Jun.</td>
<td>99</td>
<td>0.0015</td>
<td>0.1145</td>
<td>1.5063</td>
</tr>
</tbody>
</table>
### Departure Delay and June Cases

<table>
<thead>
<tr>
<th>ID</th>
<th>Rule</th>
<th>Itemsets</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>&gt; 30 &amp; Jun.</td>
<td>501</td>
<td>0.0076</td>
<td>0.0881</td>
<td>1.16</td>
</tr>
<tr>
<td>4</td>
<td>&gt; 40 &amp; Jun.</td>
<td>382</td>
<td>0.0058</td>
<td>0.0897</td>
<td>1.1802</td>
</tr>
<tr>
<td>5</td>
<td>&gt; 90 &amp; Jun.</td>
<td>141</td>
<td>0.0021</td>
<td>0.1038</td>
<td>1.3665</td>
</tr>
</tbody>
</table>

**Figure 2-8:** Illustration of the multi-dimensional ARM pie chart diagram for the arrival delay and October ID numbers 3 and 4.

<table>
<thead>
<tr>
<th>ID</th>
<th>Number of Itemsets</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>376</td>
<td>0.0057</td>
<td>0.0662</td>
<td>0.715</td>
</tr>
<tr>
<td>4</td>
<td>287</td>
<td>0.0044</td>
<td>0.0674</td>
<td>0.7282</td>
</tr>
</tbody>
</table>

**Figure 2-9:** Illustration of the multi-dimensional ARM pie chart diagram for the departure delay and June ID numbers 4 and 5.

<table>
<thead>
<tr>
<th>ID</th>
<th>Number of Itemsets</th>
<th>Support</th>
<th>Confidence</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>382</td>
<td>0.0058</td>
<td>0.0881</td>
<td>1.18</td>
</tr>
<tr>
<td>5</td>
<td>141</td>
<td>0.0021</td>
<td>0.1038</td>
<td>1.36</td>
</tr>
</tbody>
</table>

By comparing both approaches at the single-dimension and multi-dimension association rules, Lift showed the ability to discriminate negative (< 1) and positive (> 1) relationships in the dataset at
the multi-dimension association rule level. In the case with single-dimension, Lift always showed high positive values which did not present a useful knowledge interpretation to discriminate the relationships. However, when the data was challenged at the multidimensional level, the interpretation of the “Support and Confidence” can be misleading as shown in Table (4), and the usefulness for the “Lift” is evident.

Table 4: Illustration of the comparison between (single-dimension and multi-dimension ARM for LIM).

<table>
<thead>
<tr>
<th>Single-dimension association rules</th>
<th>Multi-dimension association rules</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td>Lift</td>
</tr>
<tr>
<td>a- Arrival Delay Cases (in minutes)</td>
<td>a- Arrival Delay and October Cases</td>
</tr>
<tr>
<td>1</td>
<td>22.3</td>
</tr>
<tr>
<td>2</td>
<td>75.96</td>
</tr>
<tr>
<td>b- Departure Delay Cases (in minutes)</td>
<td>b- Departure Delay and October Cases</td>
</tr>
<tr>
<td>3</td>
<td>11.5586</td>
</tr>
<tr>
<td>4</td>
<td>15.4223</td>
</tr>
<tr>
<td>5</td>
<td>48.3792</td>
</tr>
<tr>
<td>c- Arrival Delay and June Cases</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1.1883</td>
</tr>
<tr>
<td>2</td>
<td>1.5063</td>
</tr>
<tr>
<td>d- Departure Delay and June Cases</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>1.16</td>
</tr>
<tr>
<td>4</td>
<td>1.1802</td>
</tr>
<tr>
<td>5</td>
<td>1.3665</td>
</tr>
</tbody>
</table>

2.1.6 Data Processing Graphic User Interface.

In the process of handling the data, the Matlab scripts, and functions, the task becomes challenging and the data becomes seemingly unmanageable due to the hard visualization. Therefore, a Graphic User Interface (GUI) as shown in Figure (2-10) was developed using the “Guide” toolbox in Matlab to make the visualization process easier for the results and the output. This GUI is a very efficient and handy tool for its assistance in presenting the task and output results in a visualized manner, which makes the calling for each function and task seamless. This GUI is of special importance to non-technical professionals who would like to read the results but without the hassle of going through the codes.
The GUI gives the user the option to select any dataset. As an exercise/example in this work, the dataset used in this simulation was the USA domestic airline flights between 1987 and 2008, as presented in section (2.1.2). After the data is selected by the user, two conditions are needed to be set “Condition 1 (X)” and “Condition 2 (Y)”. These conditions were then used for the association, with the assumption of the item relations that the user is trying to find between the columns, with samples of rating/values in rows.

The presented GUI runs the main application discussed in this thesis, particularly in chapter 2.1.6. The algorithms developed were for two forms of the association rule; with and without MapReduce implementation as shown in Figures (2-11). The results from both trials matched (with and without MapReduce). This advantage is very important for the scientist when working with the data and multiple configurations for testing in addition to the simplicity of use of this tool. For example, to test whether ArrDealy > 60 (minutes) & Month = 10 (October), the GUI application will easily find the results.
The above simulation results demonstrated the functionality and capabilities of the developed GUI. It also proposed its benefits to the user by providing assistance in handling the data processing, as well as being a suitable easy visual aid when analyzing the results. Moreover, the GUI is characterized by the ability and simplicity of defining the configuration parameters and variations, as defined by the user. This in a greater sense, shows the capability of the proposed algorithm in this thesis (MRLAR).

To further evaluate the performance of the proposed algorithm (MRLAR), Matlab was used to compare the execution time of this algorithm and to compare it to the execution time of two other algorithms (Apriori and LBA) [22, 26]. Table (5) shows the execution time for Apriori, LBA, and MRLAR. The time execution comparison results showed a clear difference in the execution time on a different number of transactions, with MRLAR being significantly the fastest. This confirms that the MRLAR algorithm can help in saving time compared with both the Apriori and LBA. Figure (2-12) presents the difference in execution time and performance of proposed algorithm.
### Table 5: Illustration of the execution time for MRLAR, Apriori, and LBA.

<table>
<thead>
<tr>
<th>Size of Data</th>
<th>Apriori Time (Secs)</th>
<th>LBA Time (Secs)</th>
<th>MRLAR Time (Secs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>3334</td>
<td>0.54</td>
<td>0.45</td>
<td>0.057</td>
</tr>
<tr>
<td>13336</td>
<td>0.82</td>
<td>0.96</td>
<td>0.058</td>
</tr>
<tr>
<td>53344</td>
<td>1.16</td>
<td>1.04</td>
<td>0.058</td>
</tr>
<tr>
<td>106688</td>
<td>3.92</td>
<td>1.73</td>
<td>0.06</td>
</tr>
<tr>
<td>213376</td>
<td>6.15</td>
<td>2.11</td>
<td>0.073</td>
</tr>
</tbody>
</table>

**Figure 2-12:** Illustration of the execution time (Secs) for MRLAR, Apriori, and LBA.
3 DIMENSIONALITY REDUCTION ALGORITHMS

3.1 FEATURE SELECTION BY PRINCIPLE COMPONENT ANALYSIS (FSPCA) FOR MINING FREQUENT ASSOCIATION RULES

This section describes the process used in the second part of this dissertation. The proposed algorithm is called Feature Selection by Principle Component Analysis (FSPCA).

Section (3.1.1) defines the Groceries dataset collection and conversion. Section (3.1.2) describes the software used. Section (3.1.3) presents the proposed algorithms steps and the flowchart of FSPCA, and finally, section (3.1.4) presents the experimental results of FSPCA in this dissertation.

3.1.1 Dataset Collection and Conversion

We conducted the experiment on the Groceries dataset of real-world point-of-sale transaction data from a typical local grocery outlet. The dataset contained 9835 transactions of items with 169 categories [27]. Every transaction contained 1-32 itemsets. This dataset can be downloaded from the USCOTS Workshop website http://course1.winona.edu/cmalone/workshops/uscots2015/ [28].

From the original file we created new datasets which the attributes are the names of the products (169 itemsets), and the instances are the 9835 transactions, to a shape of a binary matrix $D$ of 9835 $\times$ 169, such that:

$$d_{ij} = \begin{cases} 1 & \text{if itemset } j \text{ appears in transaction } i \\ 0 & \text{otherwise} \end{cases}$$

3.1.2 Software Used

For this part of the thesis, the algorithm used the open source operator Rapidminer (6.5) Weight by PCA for its well-known ability to support big data mining [29].

3.1.3 Proposed Algorithm (FSPCA)

The main purpose of using the PCA was to reduce the high-dimensional data with a dimension $m$ into a lower dimensional data of dimension $k$, where $k \leq m$. 

27
To determine the reduction size, we used the attributes weights generated by the components created by the PCA. The attribute weights reflect the relevance of the attributes with respect to the class attribute. The higher the weight of an attribute, the more relevant it is considered.

The Results of applying PCA gave numerical floating numbers which were not compatible with the FP-Growth algorithm. Therefore, it was necessary to convert the data to binomial.

Our proposed algorithm consisted of 5 phases:

1. The data transformation phase: The transaction Groceries dataset was transformed into a binary matrix, in which each column corresponded to an attribute (itemset), and each row corresponded to a transaction. The resulting dataset was $9835 \times 169$ binary matrix.

2. In the second phase, we passed the dataset to the weight by PCA operator which used the PCA to assign weights to the attributes and sorts them according to their weights in an ascending order.

3. The third phase was the feature selection (dimension reduction) based on the PCA weights. In this phase, we chose a weight threshold such that each attribute with a weight less than the threshold was dropped from the dataset, and only the attributes with weights above the threshold were considered for the next phase. Then a new dataset was created based on the selected attributes.

4. The fourth phase was data conversion to the binomial dataset. In this phase, the new datasets obtained from the previous phase were converted to the binomial dataset using the numeric to the binomial operator.

5. The fifth phase applied the FP-Growth algorithm to the new datasets to find the minimum frequent itemsets that satisfy the support parameter. Then, the FP-Growth passed the output to the create association rule operator to extract the rules by using lift interestingness measure.

3.1.4 Experiment Results

This section explains the experiment we applied on the Groceries dataset to extract the association rule based on the PCA. As mentioned earlier, we ran the experiment in Rapidminer (6.5). We configured the Weight by PCA operator such that the outputs weights are normalized between $[0, 1]$, and the attributes were sorted in an ascending order based on their PCA weights. The attribute with the maximum weight was the whole milk with weight 1, followed by other vegetables, yogurt, root vegetables as 0.75, 0.47, and 0.42 respectively. The (sound storage medium) was the attribute with the minimum PCA weight valued at 0. We adjusted the selected attributes by weight operator to select only
the attributes with PCA weights ≥ 0.01. This resulted in selecting 107 attributes (itemsets) out of the original 169 attributes, a reduction by 36.6% of the data attributes was achieved.

For mining the association rules, we used the FP-Growth operator with minimum support value set to 0.0018 to find the minimum frequent itemsets by building FP-tree, then we used the create association rules operator which takes these frequent itemsets and find the rules that satisfy the confidence parameter, which we set to 0.95.

The following Figure (Figure 3-3) and Table (6) describe the Association Rules found by applying the FP-Growth algorithm.

Table 6: Illustration of the association rules extracted by PCA

<table>
<thead>
<tr>
<th>Consequence(LHS)</th>
<th>Antecedent(RHS)</th>
<th>Lift</th>
</tr>
</thead>
<tbody>
<tr>
<td>citrus fruit, domestic eggs, sugar →</td>
<td>whole milk</td>
<td>3.652</td>
</tr>
<tr>
<td>sugar, rice →</td>
<td>whole milk</td>
<td>3.91</td>
</tr>
<tr>
<td>other vegetables, yogurt, root vegetables, oil →</td>
<td>whole milk</td>
<td>3.91</td>
</tr>
<tr>
<td>yogurt, root vegetables, tropical fruit, sausage →</td>
<td>whole milk</td>
<td>3.669</td>
</tr>
<tr>
<td>whole milk, yogurt, root vegetables, oil →</td>
<td>other vegetables</td>
<td>4.824</td>
</tr>
<tr>
<td>root vegetables, tropical fruit, citrus fruit, whipped/sour cream →</td>
<td>other vegetables</td>
<td>5.168</td>
</tr>
</tbody>
</table>

Figure 3-1: Illustration of the association rules extracted by PCA
3.2 Feature Selection Using Semi-Discrete Decomposition (SDD) and Singular Value Decomposition (SVD)

This section describes the process used to find out the results of the second part of this thesis. Section (3.2.1) describes the Hepatitis dataset used and its preprocessing. Section (3.2.2) presents the software and technique used. Section (3.2.3) covers the proposed algorithm and flowchart. Section (3.2.4) covers the experimental results including a comparison between applying SDD and SVD to select significant features from Hepatitis dataset, and finally, this section is concluded by the summary of the proposed algorithms.

3.2.1 Dataset collection and preprocessing

For testing of semi-discrete decomposition (SDD) and singular value decomposition (SVD) as dimensionality reduction and features selection, we used the hepatitis dataset downloaded from UCI Machine Learning Repository [30]. The dataset total number of instances 155 and 20 attributes including class (died and alive), age, sex, steroid, antivirals, fatigue, malaise, anorexia, liver big, liver firm, spleen palpable, spiders, ascites, varices, bilirubin, alkaline phosphate, sgot, albumin, protime and histology.

The pre-processing stage was applied on hepatitis dataset by cleaning it from missing values and noise “which includes errors, and outlier/extreme values”. Hepatitis dataset contained 167 missing values; all of which were replaced by the mean value. Then there is 28 outliers and 5 extreme values that were removed/ excluded. After pre-processing the remaining instances were 122 and 20 attributes.

3.2.2 Software Used

In this experiment, by using Weka data mining tools, a Multilayer Perceptron Neural Network (MLPNN) has been chosen to be used for the classification part of the task. MLPNN is a feedforward neural networks model that maps a set of input data onto a set of appropriate outputs without cycling (looping), MPLNN utilizes the standard backpropagation algorithm for training the network [32].

3.2.3 Proposed Algorithm

The Truncated SVD for matrix $A$ is decomposed as multiplication of three matrixes $U_k, D_k, and V_k^T$, where $k (0 < k < n)$. Please see (Figure 3-4) for details.

The Truncated SVD was used to reduce dimensionality by removing part of $A$ which was considered as noise in original data. Even with applying dimensionality reduction, storage cost may be large since $U_k and V_k^T$ contain many small entries. It is possible to replace part of the small entries in
$U_k$ and $V_k^T$ with zeros and remove them without significant effect on query accuracy performance on the dataset.

\[
\begin{pmatrix}
\vdots & \mathbf{A} & \vdots \\
. & m \times n & .
\end{pmatrix}
\cong
\begin{pmatrix}
\vdots & \mathbf{U}_k & \vdots \\
. & m \times k & .
\end{pmatrix}
\begin{pmatrix}
\vdots & \mathbf{D}_k & \vdots \\
. & k \times k & .
\end{pmatrix}
\begin{pmatrix}
\vdots & \mathbf{V}_k & \vdots \\
. & k \times n & .
\end{pmatrix}
\]

Figure 3-2: Illustration of the truncated singular value decomposition “rank-k”

In this part of study, reduce dimensionality used to label features with small entries by applying threshold value $T$ on $U_k$ and $V_k^T$. The small entries in $U_k$ and $V_k^T$ are set to zero when they are smaller than the threshold value and treated as nonessential features. This in turn results in a sparse matrix.

There are many strategies used to generate sparse matrix using threshold namely single threshold strategy, Column Threshold Strategy, Exponential Threshold Strategy and $L_1$ regularization [31, 32].

The Exponential Threshold Strategy (ETS) was used to find a smooth threshold function because it calculates threshold for each column in $U_k$ and $V_k$. However, the other techniques apply one threshold value to all matrix. The ETS for matrix $\mathbf{A}$ with $m$ rows and $n$ columns can be calculated by the (Equation 3.1):

\[
T_j = \frac{\epsilon}{m} \sum_{i=0}^{m} |a_{ij}| e^{(k)j^{-2}}
\]

\text{Equation 3.1}

Where $\epsilon$ is scalar factor and $k$ is a number of singular values. The equation 3.1 computes threshold for each column in $U_k$ and $V_k^T$ and adjusted it by scalar $\epsilon$. And for any $u_{ij}$ in $U_k$ if $u_{ij} < T$ we set $u_{ij} = 0$ result in $U_k$. The same thing is applied to $V_k^T$ to give $\mathbf{V}^T_k$. The resulted sparse matrix is as shown in (Equation 3.2):

\[
\mathbf{\tilde{A}}_k = \mathbf{\tilde{U}}_k \mathbf{D}_k \mathbf{\tilde{V}}_k
\]

\text{Equation 3.2}

The zero vectors in new matrix $\mathbf{\tilde{A}}_k$ can be seen as nonessential features, and can be removed from original data to reduce dimensionality.
3.2.4 Experimental Results

a) Applying Semi-Discrete Decomposition (SDD)

First, the semi-discrete decomposition was applied on the hepatitis dataset to produce three matrices $U_k$, $D_k$ and $V_k^T$ with different values for $k$, according to the value of $k$ some features value are set to zero. These zero vectors can be seen as nonessential features; it is difficult to identify them in original data.

The evaluation parameters to compare results between applying SVD and SDD on original dataset and set of selected features using MLPNN on hepatitis dataset are accuracy in term of a number of correctly classified instances, reduction of storage and training time.

Table (7) below shows selected features with different $k$ values, reduction rate, accuracy (number of correctly classified instances) and time needed for training for original dataset and newly selected attributes set:

<table>
<thead>
<tr>
<th></th>
<th>Original dataset</th>
<th>Selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$k = 7$</td>
<td>$k = 13$</td>
</tr>
<tr>
<td>No of attributes</td>
<td>20</td>
<td>4</td>
</tr>
<tr>
<td>Feature Reduction Rate</td>
<td>-</td>
<td>80%</td>
</tr>
<tr>
<td>Training Time - MLP</td>
<td>0.66</td>
<td>0.1</td>
</tr>
<tr>
<td>Accuracy</td>
<td>86%</td>
<td>83.6%</td>
</tr>
<tr>
<td>Storage Reduction Rate</td>
<td>-</td>
<td>50%</td>
</tr>
</tbody>
</table>

b) Applying Singular Value Decomposition (SVD)

The truncated singular value decomposition was applied on Hepatitis dataset which resulted in the three matrixes $U_k$, $D_k$ and $V_k^T$. After that, Exponential Threshold Strategy was used to calculate the threshold for each column in $U_k$ and $V_k^T$ with $\epsilon = 0.04$ for all experiments and different values for each $k$.

The elements $u_{ij}$ and $v_{ij}$ in each column was compared with column threshold. If the value was less than column threshold, then the values of the element were set to zero. The zero vectors were
treated as a nonessential feature and removed from original data. The remaining set of significant features were used in training.

Table (8) shows selected features set from hepatitis dataset with different $k$ using SVD, reduction rate, time taken for training using multi-layer perceptron and accuracy.

<table>
<thead>
<tr>
<th></th>
<th>Original dataset</th>
<th>Selected features</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>$k = 7$</td>
</tr>
<tr>
<td>No of attributes</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>Feature Reduction Rate</td>
<td>-</td>
<td>5%</td>
</tr>
<tr>
<td>Training Time - MLP</td>
<td>0.66</td>
<td>0.56</td>
</tr>
<tr>
<td>Accuracy</td>
<td>86%</td>
<td>87%</td>
</tr>
<tr>
<td>Storage Reduction Rate</td>
<td>-</td>
<td>0%</td>
</tr>
</tbody>
</table>

From the previous experiments, the accuracy of original hepatitis dataset after pre-processing was 86% which is good compared to other algorithms in the literature [24]. We tried to improve it using dimensionality reduction techniques.

First, we applied semi-discrete decomposition with $k = (7, 13)$ on the dataset. The number of selected features had slightly decreased from 20 to 4, 5 features respectively. Many important features were removed so the number of correctly classified instances (accuracy) was also decreased to 83 and 84. However, the multi-layer perceptron training time and storage reduction rate were significantly improved.

In the second experiment, we applied the singular value decomposition with the same values for $k = (7,13)$ on hepatitis dataset. There were a few attributes are treated as nonessential and set to zero and in real they were not important to diagnose hepatitis like malaise, anorexia, steroid and sgot. When we removed them, the accuracy was stable and sometimes increased. However, the training time had a slight decrease and there was no reduction in the storage as our dataset was small.

It is noted from the above experiment that while both SVD and SDD had the same starting accuracy when applied to the original dataset, SVD proved to have more accuracy than SDD when applied to the same datasets with selected features such as $k = 3$, or $k = 7$. They both started with 86% accuracy level then the SDD accuracy went down when we had selected features $k = (7,13)$ while the
accuracy of the SVD has increased under the same circumstances. This means that SVD gave more accurate results than the SDD.

### 3.3 Summary and Wrap-up of the Thesis Work:

This thesis presented an innovative approach that attempted to deal with the challenges associated with the Big Data mining, processing and visualizing. This work came in two parts: utilizing the association rules especially the Lift Interstingness Measure “Lift” for data mining and then utilizing the MapReduce to manage the huge amount of data efficiently.

The thesis work is presented in four main sections (Chapters 2-5). Chapter Two presented the background of the Big Data; its history and present state, sources and types, characteristics and analytics as well as its applications and notable uses. After this the Data mining types, techniques, frequent patterns and association rules. This leads to the necessity of explaining the data reduction methods needed to manage the data efficiently. This part concentrated on the concept of dimensionality reduction and feature selection, SVD, SDD and PCA.

In relation to Chapter Two, we published four papers in this topic in international conferences proceeding indexed in Scopus, Dblp, and Thomsmon ISI. Following is the list of publications topics:


Chapter Two revealed how Big Data have widely changed several scientific research trends. Multiple analytic techniques have been developed to extract and visualize this huge and complex amount of non-stop data. Such techniques are being used for optimization processes and decision making based on the improvement of the traditional data mining and machine learning algorithms. It also presented the challenge of the growing aspects of Big Data with the necessity to utilize infrastructures in various application aspects to manage these data. It briefly consisted of management aspect (processing) and its platforms such as IBM, SAP, Amazon, etc and the extraction, analysis, and visualization such as NoSQL, MapReduce, Hadoop, and R Language. This processing concludes with the presentation of the storage aspects including DAS, NAS, SAN, Cloud, and Data Center. Then this chapter presented the pre-processing aspects of the paradigm including the Extracting association rule based on MapReduce by using Lift of the rule as well as SVD and PCA. It was concluded at the end of this chapter that mining association rule by using LIM based on MapReduce have been developed for extracting relationship between the itemset instead of using confidence. Reduce dimensionality was also utilized to reduce the attributes into fewer dimensions as a pre-processing technique deployed for association rule mining.

Chapter three reviewed the related work and literature for the two main components of this dissertation and is presented in two parts: The association rules mining algorithms and the dimensionality reduction algorithms. The details of each part is presented in a separate following chapter. The parallel approach (MapReduce) was presented in the work of several investigators who basically came to the following conclusions: The mapReduce provide scalability and time over the traditional algorithms, provided higher performance, paralleled the association rule extraction phase and could be easily utilized with the Hadoop platform, had a higher accuracy and efficiency. Moreover, the literature on the data reduction algorithm had shown that the serial approach was working in compression of the size of the formal context investigated. The nonnegative matrix factorization (NMF) was studied and found to be applied in healthcare settings for data reduction. The SVD alone or combined with the MapReduce, and more variety of combinations is presented in this chapter.

Based on the presented literature, the parallel association rule and reduce dimensionality algorithms are two of the best choices for high-performance data mining techniques. Several studies were developed to improve the performance of traditional association rule mining algorithms by using support and/or confidence as global measurements of the rule for solving different big data mining problems. It is also evident that MapReduce approach are widely used and considered as one of the best adoptions to improve such achievements in data mining.

Therefore, this study developed a novel MapReduce algorithm by using LIM to find the correlation between LHS and RHS in association rule mining, unlike the other studies that either used confidence and/or support of the rule to determine the relationship between the data as a first part of this
study. In its second part, it investigates a new dimension reduction technique by using feature selection based on multiple data reduction techniques as a pre-processing to find the association rules.

The author of this thesis work has published three papers in this area of research as follows:

1. **Nour E. Oweis**, Mohamed Mostafa Fouad, Sami R. Oweis, Suhail S. Owais and Vaclav Snasel, “A Novel MapReduce Lift Association Rule Mining Algorithm (MRLAR) for Big Data” International Journal of Advanced Computer Science and Applications(IJACSA), 7(3), 2016. **Indexed in Thomson Reuters Emerging Sources Citation Index, a new edition of Web of Science (ISI).**

2. Tayseer M.F Taha, Eltayeb Shomo, **Nour E. Oweis**, and Václav Snášel “Feature Selection by Principle Component Analysis for Mining Frequent Association Rules”. In 1st International Scientific Conference “Intelligent Information Technology for Industry”. Springer International Publishing, Russia, (In-press). **Indexed in Scopus and will be indexed in ISI.**


Chapter four was mainly to present the association rule mining algorithms, and gave special attention to our novel MapReduce Life Association Rule (MRLAR). It presented the methods used in this thesis for prototype implementation, dataset collection, software used, MRLAR description, and experimental results. In this part, we were able to develop a parallel association rule mining algorithm based on MapReduce approach by using LIM, denoted as (MRLAR). The experimental results presented in this thesis show that MRLAR performed effectively in prediction by integrating the uses of MapReduce and LIM instead of using confidence to determine the correlation between LHS and RHS in association rule: positive, negative or independent correlation in a parallel association rules to help the user make important decisions through determining and explaining those patterns in a meaningful knowledge format.

In Summary, the proposed algorithms had the ability to provide pre-processing techniques to reduce the dimensionality of the dataset and parallelism the LIM that shown to solve the problem of impact LHS on the RHS in association rules. In other words, it determines the type of correlation between LHS and RHS in parallel association rules which will be easily applied to many commodity machines that deal with big data mining without considering the synchronization problem.

Chapter Five presented the second part of the thesis work and concentrated on dimensionality reduction algorithms: Feature Selection by Principle Component Analysis (FSPCA) for Mining
Frequent Association Rules, and Feature Selection Using Semi-Discrete Decomposition and Singular Value Decompositions. In both sections, we presented the dataset collection and preprocessing, software used, the algorithm description and experimental results.

The main goal of this chapter was to introduce feature selection method based on the PCA weight, where the attributes are ranked according to their weight. Then, the attributes with weights less than a certain pre-determined weighting threshold are eliminated from the data, resulting in new dataset with fewer dimensions. FSPCA showed that with this data reduction technique one can obtain the same association rules that are obtained by the original data. In a real application where datasets contain a huge number of transactions and itemsets, extractions of the association rules may become a hard problem to solve due to the high dimensionality of the data. It was found that not all the transactions have the same level of importance. Therefore, eliminating some of them might not affect the information gained from the transaction dataset.

Additionally, in this chapter, a singular value decomposition and semi-discrete decomposition were applied on hepatitis dataset for dimensionality reduction and features selection by finding nonessential features using Exponential Threshold Strategy and remove them from original data to improve performance. After that multi-layer perceptron for classification was applied on selected features set. Finally, a comparison between two techniques in term of training time, storage and accuracy revealed that the experiment features selection using Singular Value Decomposition is more appropriate than Semi-Discrete Decomposition because it improves accuracy and had acceptable computation cost.
4 CONCLUSIONS

The era of Big Data is very wide and considered as one of the current research topics. This thesis focused on dealing with two major issues in Big Data processing, interpretation, and visualization. It proposed two algorithms for data mining: finding meaningful relationships using the Lift interestingness measure (LIM) and Reduction of dimensionality by eliminating any non-meaningful data. It also developed a Graphic User Interface (GUI) for the MapReduce so that data processing becomes more user-friendly, and more attractive to non-technical professionals. This thesis proposes

The first proposed algorithm is a novel algorithm designed for parallelism of association rule mining. It is based on MapReduce approach using LIM. This approach has allowed the data mining processes to be used with higher efficiency and better performance in handling some of the big data mining problems by producing correlations. By the development of MapReduce algorithm by using Lift interestingness measure (MRLAR), we were able to achieve a high powerful process, very fast execution environments over vast amounts of data, fault tolerance and extremely scalable scalability utilizing parallel execution. The proposed algorithm with only one scan database provided a high measurable benefit that allowed a high capability to directly extract association rule and determine the type of correlation between LHS and RHS by using LIM without the need to calculate confidence values, hence eliminating the need for additional calculations. The LIM shown its reliability measurement at the multidimensional level compared to the “support and confidence” model that could only survive under the single dimension association rule. Additionally, LIM added benefits to the prediction process of the future consequence in future datasets. Another achievement in the first part of the thesis is the development of the Graphic user Interface in the first algorithm. This has provided us with a more user-friendly data management option especially in non-technical professionals such as professionals in the health care systems and marketing.

The second proposed algorithm offered the ability to provide a pre-processing technique to reduce the dimensionality of the dataset by working only with meaningful data. This approach comes with several advantages upon application: It reduces the data capacity, reduces the amount of time and memory required by data mining algorithms, provides easy visualization of data, and eliminates irrelevant features and noise reduction, thus reducing costs of data preprocessing. Feature Selection by Principle Component Analysis (FSPCA) is a pioneering approach used in the mining of the association rules. In this thesis, we used the PCA for dimensionality reduction and for feature extraction in an attempt to beat the curse of dimensionality. It is important to note that the attributes’ weights were used by PCA in the following manner: The higher the weight of an attribute, the more relevant it is considered. Moreover, the dimension was reduced without losing the relevant information. FP-Growth
algorithm was applied to find the frequent itemsets in order to discover the useful rules and lead to better visualization of the datasets. In the presented work earlier, we adjusted the selected attributes by weight operator to select only the attributes with PCA weights ≥ 0.01. This resulted in selecting 107 attributes (itemsets) out of the original of 169 attributes, a reduction by 36.6% of the data attributes was achieved without losing any relevant information.

Moreover, the second proposed algorithm of this thesis presented the dimensionality reduction algorithms by using the Singular Value Decomposition (SVD) data reduction techniques in order to handle another challenge of big data by finding nonessential features using Exponential Threshold Strategy and remove them from original data to improve performance. Based on this thesis work, we were able to confirm that the SVD approach was more appropriate than the Semi-Discrete Decomposition (SDD) because it improved accuracy and has acceptable computation cost. When compared, both SVD and SDD had the same starting accuracy when applied to the original dataset, but SVD proved to have more accuracy than SDD when applied to the same datasets with selected features such as $k = 7, or k = 13$. They both started with 86% accuracy level then the SDD accuracy went down when we had selected features ($k = 7, or k = 13$) while the accuracy of the SVD has increased under the same circumstances. This means that SVD gave more accurate results than the SDD.

Big data is a huge challenge in this data production era, and it was an enthusiastic challenge during this work to find out that we can still manage the Big Data by applying the scientific knowledge and combining the different mining algorithms. LIM, MapReduce, dimensionality reduction and SVD can all be used in the data mining of Big Data in various aspects of our daily life including health care systems, groceries, and industries.

4.1 Limitation of the Study and Directions for Future Work

Although there were many benefits of MapReduce algorithm, but it has also some limitations at the present time. Some of these limitations are specific to MRLAR such as an extracting the association rule mining with single dimension as shown in the results chapter four. Another limitation which is based on MapReduce approach itself that it operates only on data structures of type (key, value) pairs, so that all the input datasets must be adapted to such structure.

It has been well known that SCIENCE can never answer questions with absolute certainty but can give us the best answer based on the knowledge at the time. Therefore, based on the findings of this thesis and assessment of the future needs, the following list is the proposed future work to complement what have been started and accomplished:
- Optimization of the current proposed approach to avoid the limitation of extracting the association rule mining with a single dimension.

- To parallelize data reduction techniques by using SVD, SDD, and other data reduction techniques to remove the unnecessary data with better time based on MapReduce approach. This algorithm has the ability to provide pre-processing techniques to reduce the dimensionality of the dataset which reduces the data capacity, thus reducing costs. Hence, MapReduce dimensionality reduction algorithm by SVD will handle another challenge of big data to avoid data dimensionality problems in parallel approach, therefore, reduce the amount of time and memory required by data mining algorithms, easy visualization of data, and eliminate irrelevant features and noise reduction.
REFERENCES


[8] Aggarwal, Charu C., and Jiawei Han, eds. Frequent pattern mining. Springer International Publishing, 2014.


[27] Hahsler, M., Hornik, K., & Reutterer, T. “Implications of probabilistic data modeling for mining association rules”. In From Data and Information Analysis to Knowledge Engineering (pp. 598-605). Springer International Publishing, 2006.


LIST OF STUDENT’S PUBLICATIONS

PUBLISHED PAPERS:

a) Related to the topics:


7) Intisar Hussien, Sara Omer, **Oveis, N. E.**, and Václav Snášel, “Feature Selection Using Semi Discrete Decomposition and Singular Value Decompositions”. In 1st International Scientific

b) Unrelated to the Topics:

