CHAPTER 1
INTRODUCTION

This chapter deals with an introduction to data reduction in data mining, various categories of data reduction in data mining like clustering, classification and selection of attributes along with preprocessing filters methods in data mining, motivation of the proposed system, contribution and organization of the thesis.

1.1 OVERVIEW OF DATA REDUCTION METHOD IN DATA MINING

Usually when the data size becomes huge then the analysis becomes complex in terms data and mining which takes a very long period of time and finally making such analysis impracticable or infeasible (Jiawei Han & Micheline Kambar 2012). The data reduction techniques aim to obtain a reduced representation of the data set that is very small in volume, but it maintains the integrity of the original data. Therefore, the data reduction mines on the reduced data set which is efficient and produce the same analytical result. In the last two decades, databases and data mining techniques have proved to be a tool with great practical value for handling very large data. Data Mining can be done by many techniques like association rule mining, clustering, classification etc. (Tan et al. 2006). Clustering is one of the well known, popular data mining techniques. To group similar objects, clustering is a powerful technique in machine learning. It is popularly known for its efficiency and for discovering unknown knowledge from very large,
complex data sets. Data mining is the computerized process of thorough
digging and analyzing big volume of datasets and then extracting the
meanings of the data and also analyzing datasets from different perspectives
and summarizing it into useful information. Data mining plays a vital role in
terms of prediction and analysis. In data mining the clustering is a kind of
 technique which is used to group the data on the basis of similarity. The
groups are often called as Clusters. Clustering is the task of grouping a set of
objects in a group therefore the objects in the same group are more similar to
each other than to those in other groups. However, how good a cluster is
depending ultimately on the opinion of the user, based on the domain in
which it is applied. The three essential questions in data clustering are, how
many groups are there?" what are the groups?" (Fraley & Raftery 1998) and is
this the exact number of groups?”. The main goal is, using an enhanced
clustering technique with K-Means with Normalize preprocessing filter.

Currently clustering is the most active research in several
disciplines like statistics, pattern recognition and machine learning. With
respect to machine learning, the clusters give rise to hidden patterns. The
cluster forming in unsupervised learning represents a data concept. In
practical viewpoint clustering plays an exceptional role in data mining
applications such as scientific data exploration, information retrieval, mining
the text data, spatial data, web data and analysis, CRM, marketing, medical
diagnostics, computational biology, weather forecast, stock exchange and
many others. The main idea of feature reduction with respect to clustering is
to find a subset of features by completely removing the features with little or
no predictive information. Data mining is the study of collecting, cleaning,
processing, analyzing, and gaining useful insights from data. A wide variation
exists in terms of the problem domains, applications, formulations, and data
representations that are encountered in real applications. Therefore, “data
mining” is a broad umbrella term that is used to describe these different
aspects of data processing. Data mining strategies fall into two broad
categories namely Supervised Learning and Unsupervised Learning.
Supervised Learning methods are deployed when there exists a field or
variable (target) with known values and about which predictions will be made
by using the values of other fields or variable inputs (Nargess Memarsadeghi,
Dianne E O’Leary 2003). Unsupervised Learning methods tend to be
deployed on data for which there do not exist a field or variable with known
values, while fields or variables do exist for other fields or variables.

Data mining is the extraction of interesting nontrivial, implicit,
previously unknown and potentially useful information or patterns from large
databases (Jiawei Han et al. 2012).

1.2 DATA MINING TECHNIQUES

It is one of the phases in knowledge discovery process. The main
goal of this phase is pattern recognition. But this pattern recognition varies
based on user interests and based on the application. There exist several
mining techniques for pattern recognition which are described below.

Several effective data mining techniques have been developed for
detecting intrusions which perform close to or better than systems engineered
by domain experts (Xiaohong Guan et al. 2009). However, successful data
mining techniques are themselves not enough to create deployable IDS.
Despite the promise of better detection performance and generalization ability
of data mining-based IDS, there are some inherent difficulties in the
implementation and deployment of these systems. These difficulties can be
grouped into three general categories: accuracy i.e., detection performance,
efficiency, and usability. Typically, data mining-based IDS especially
anomaly detection systems have higher false positive rates than traditional
hand-crafted signature based misuse detection systems methods, making them
unusable in real environments. Also, these systems tend to be inefficient which is computationally expensive during both training and evaluation. This prevents them from being able to process audit data and detect intrusions in real time. Finally, these systems require large amounts of training data and are significantly more complex than traditional systems. In order to be able to deploy real time data mining based IDS, these issues must be addressed.

**Intrusion Detection**

Large amount of data exists in the system which could be gathered by network personnel to detect security policy violations. With this scenario, the analysis is a tedious one and network administrators do not have the resources to analyze the data for security policy violations especially in the presence of a high number of false positives that cause them to waste their limited resources. Data mining techniques are used in classification and identification of new patterns from large volume of training data that are collected from KDD -Knowledge Discovery in Data Mining CUP 1999 benchmark dataset in order to perform hybrid intrusion detection in host as well as in network (Cherkassky & Mulier 1998). Moreover, intrusion detection has been carried out using classification and clustering algorithms integrated with feature selection (Liu & Yu 2005). The recent rapid development in data mining has made available a wide variety of algorithms, drawn from the fields of statistics, pattern recognition, machine learning, and database.

Clustering is a general unsupervised classification procedure that divides a set of objects in different classes. Objects from the same class should be similar to each other. There is no initial indication about the classes or about their number but only the properties of the objects in the set of data.
1.2.1 Various Data Mining Technologies

1.2.1.1 Naive Bayes Classifier

A simple and widely used classifier based on the Bayes theorem used to classify documents/sentiments. The idea is to appraise the probability of a test document belonging to each category and then select the most probable category. This is mathematically stated as follows (Chaudhari & Govilkar 2015)

\[
P(c_j | d) = \frac{P(d | c_j) P(c_j)}{P(d)}
\]

Where, \( P(c_j | d) \) = probability of instance of \( d \) being in class \( c_j \)

\( P(d | c_j) \) = probability of generating instance of \( d \) in given class \( c_j \)

Naive Bayes algorithm is implemented to estimate data probability to be negative/positive. Thus, probability (conditional) of a word with positive/negative meaning is calculated in view of a slew of positive/negative examples and by calculating each class’ frequency.

\[
P(Sentiment \mid Sentence) = \frac{P(Sentiment)P(Sentence \mid Sentiment)}{P(Sentence)}
\]

So,

\[
P(Word \mid Sentiment) = \frac{\text{Number of word occurrence in class}}{\text{Number of words belonging to a class} + \text{Total nos of word} + 1}
\]
Uses of Naive Bayes classification (Mahendran et al. 2013):

- Naive Bayes text classification: used as a probabilistic learning method. Naive Bayes classifiers are the most successful known algorithms for classifying text documents.

- Spam filtering: the best known use of Naive Bayesian text classification. It uses a Naive Bayes classifier to identify spam email. Bayesian spam filtering distinguishes illegitimate spam email from legitimate email. Many modern mail clients implement Bayesian spam filtering.

- Hybrid Recommender System: Using Naive Bayes classifier and Collaborative Filtering Recommender Systems apply machine learning/data mining techniques to filter unseen information and predict whether a user likes a given resource.

1.2.1.2 Support Vector Machines (SVM)

It is a Vector space based machine-learning method which finds a decision boundary between two classes maximally far from a training data point. Though SVM performs the linear classification there exists a classification non linear method that implements a root trick which absolutely maps the inputs to high-dimensional feature spaces. This discriminative classifier is the best text classification method (Govindarajan & Romina 2013) SVM for testing different data set domains and for using weighting schemes. It also consists of experiments with different features from three corpora. The SINAI Corpus was built from Amazon.com specifically to prove SVM feasibility for different domains.
SVM creates a hyper plane or a hyper plane set in infinite dimension space. A document's SVM score $z_j$ is mathematically given as follows (Chaudhari & Govilkar 2015)

$$z_j = w_1 x_{j1} + w_2 x_{j2} + \ldots + w_d x_{jd} + b$$

\text{i.e.} \quad z_j = x_j^T w + b

Where,

- $X_i$ is a p-dimensional real vector.
- $w$ is a vector that contains the weights and is given as
  $$\vec{w} = \sum_j \alpha_j \vec{c}_j \vec{d}_j, \quad \alpha_j \geq 0, \quad c_j = \{1,-1\}$$
- $b$ is a constant

1.2.1.3 MLP

MLP is a technique that acts as a universal function approximator. MLP is a general, flexible and non-linear tool because a “back propagation” network has a hidden layer with various non-linear entities that learn every function/relationship between groups of input/output variables (whether variables are discrete/continuous).

MLP is a back propagation algorithm with two phases:

Phase I: It is a forward phase where activation is propagated from an input to an output layer.

Phase II: To change weight and bias value errors among practical & real values and the requested nominal value in an output layer is thus propagated backward.
1.2.1.4 Clustering Classifier

To identify the prominent features of humans/objects and recognizing those with a type one needs object clustering. Basically clustering is an unsupervised learning technique (Singh & Husain 2014).

Clustering determines a fresh or different set of classes/categories; new groups are of concern in themselves and their valuation is intrinsic. In this method, data objects/instances are grouped subsets that similar instances are grouped together and various different instances belong to other groups. As clustering is an unsupervised learning task, no class values which represent a former data instances combination are given situation of supervised learning.

Clustering assembles data objects/instances into subsets that similar objects are assembled together, while different objects belong to different groups. Objects are thus prepared and organized into efficient representations characterizing the population being sampled.

Clustering organization is denoted as a set of subsets $C = C_1 \ldots C_k$ of $S$, so that:

$$S = \bigcup_{i=1}^{k} C_i \text{ and } C_i \cap C_j = \emptyset \text{ for } i \neq j$$

Hence, an object in $S$ is related to exactly one subset.

1.2.1.5 Entropy

It is a conditional exponential classifier which maps a pair of feature sets and its label to a vector. It is called a log linear classifier as they work by extracting some features set from an input, combining them linearly
(each feature is multiplied by its weight and added) and using the sum as an exponent. It is parameterized by a set of weights to combine joint-features generated from a features set by encoding.

1.2.1.6 Decision Tree

It is a tree where internal nodes are represented by features, edges represent tests to be undertaken at feature weights and leaf nodes represent categories resulting from the above tests. It categorizes a document by starting at a tree root and moves successfully down via branches whose conditions are satisfied by a document till a leaf node is reached. The document is then classified in a category that labels the leaf node. Decision Trees were used in applications such as speech and language processing.

1.2.1.7 Nearest Neighbor Method

Nearest neighbor methods are the simplest and the most effective classes of classification algorithms used. They are based on the principle that for a set of training sentiment instances, the class of a new yet unseen occurrence is likely to be that of the majority of its closest “neighbor” instances from a training set (Kalaivani & Shunmuganathan 2015). Thus K-Nearest Neighbor algorithm works by inspecting k closest instances in a data set to a new occurrence that needs classification, and predicts based on what classes most k neighbors belong to. The notion of closeness is given by a distance function between two points in attribute space, specified a priori as a parameter to the algorithm. An example of distance function used is standard Euclidean distance between two points in an n-dimensional space, where n is a number of attributes in a data set.

Commonly used sentiment classification features are introduced below (Sadegh et al. 2012).
- Terms and their frequency: these consist of single words or word n-grams and their frequency/presence. The features were successfully used in sentiment classification and were effective for the task.

- POS information: an important indicator of sentiment expression. For example, adjectives carry much information regarding a document’s sentiment.

- Opinion words: Opinion words (sentiment words) and phrases are words/phrases expressing positive/negative emotions. For instance, good, fantastic, amazing and brilliant are words with positive emotion while bad, boring, slow, worst and poor are words having negative emotions. Though most opinion words are adjectives/adverbs, nouns and verbs also express opinion.

- Negations: negation words are important to evaluate sentence polarity as they transform sentiment orientation in a sentence. For instance, a sentence “I don’t like this mobile” has a negative orientation.

- Syntactic dependency: research work in this area used word dependency based features generated from dependency tree/parsing.

1.3 CHARACTERISTICS OF A GOOD CLUSTERING ALGORITHM

A specific method can perform well on one dataset, but very poorly on another, depending on the problem domain, size and dimensionality of the data as well as on the objective function used. The following are typical requirements of a clustering algorithm in data mining
• Find arbitrary-shaped clusters: Different types of algorithms will be biased toward finding different types of cluster structures/shapes and it is not always an easy task to determine the shape or the corresponding bias. Especially when categorical attributes are present, it may not be relevant to talk about cluster structures.

• Minimal requirements for input parameters: Many clustering algorithms require some user-defined parameters, such as the number of clusters, in order to analyze the data. However, with large datasets and higher dimensionality, it is desirable that a method requires only limited guidance from the user, in order to avoid biasing the result.

• Scalability: The ability of the algorithm to perform well with a large number of data objects. • Analyze mixture of attribute types: The ability to analyze dataset with mixtures of attribute types, as well as homogeneous ones.

• Handling of noise: Clustering algorithms should be able to handle deviations, in order to improve cluster quality. Deviations are defined as data objects that depart from generally accepted norms of behavior and are also referred to as outliers. Deviation detection is considered as a separate problem.

• Insensitivity to the order of input records: The same data set, when presented to certain algorithms in different orders, may lead to dramatically different clustering. The order of input mostly affects algorithms that perform only single scan over
the data set, leading to locally optimal solutions at every step. Thus, it is crucial that algorithms to be insensitive to the order of input.

- High dimensionality of data: The number of attributes/dimensions in many data sets is large, and many clustering algorithms can produce meaningful results only when the number of dimensions is small. The appearance of a large number of attributes is often termed as the curse of dimensionality (Leo Breiman & Jerome H Friedman 1997).

- Interpretability and usability: Most of the time, it is expected that clustering algorithms produce usable and interpretable results. But when it comes to comparing the results with preconceived ideas or constraints, some techniques fail to be satisfactory. Therefore, easy to understand results are highly desirable.

1.4 EVOLUTION OF CLUSTER ANALYSIS

According to the scholarly journal archive JSTOR, the word “cluster” was first used in its general sense to denote a group by (Bartram et al. 1739). The first hierarchical clustering method was developed by biologists to create hierarchy of various species for analyzing their relationship systematically. However, clustering exists in different types like Single-link clustering, complete-link clustering and average-link clustering (Lior Rokach, Oded Maimon 2005). The hierarchical clustering method used to optimize an objective function (Ward 1963). Partitional clustering, on the other hand, is closely related to data compression and vector quantization. The most popular partitional clustering algorithm, K-Means and Macqueen was proposed by (Laurence Morissette & Sylvain Chartier 2013). Later an
iterative approach to minimize the objective criterion of k-means method was proposed using Lloyd (Rafail Ostrovsky et al. 2012). Two well-known versions of K-Means in Pattern Recognition literature are ISODATA (H.Ball & J.Hall 1965). The graph-theoretic clustering method which is closely related to single link clustering (Zahn 1971). The first spectral clustering algorithm was proposed by (Shi & Malik 2000). The EM algorithm which is the standard algorithm for estimating a finite mixture model for mixture-based clustering, is attributed by (Dempster et al. 1977). Interest in mean-shift clustering was revived (Arnold Cheng 1995) The emergence of data mining leads to a new line of clustering research that emphasizes efficiency when dealing with huge database. DBSCAN by (Martin Ester et al. 1996) by for density based clustering and CLIQUE by (Rakesh Agrawal 2005) for subspace clustering are two well known algorithms in this community. Incremental clustering methods are designed to operate in a single pass over the data points to improve the speed and scalability of data clustering. The COBWEB system is an incremental conceptual clustering algorithm (Mattaios Theodorakis et al. 2004). The leaders clustering method is another simple and popular incremental clustering method which produces a partition of a dataset in linear time with respect to the size of the data set. Along with the clustering techniques that produce hard or crisp clustering, which means that each object is assigned to only one cluster, fuzzy clustering methods are developed wherein this restriction is relaxed. That is, in fuzzy clustering the object may belong to all of the clusters with a certain degree of membership. This is particularly useful when the boundaries among the clusters are not well separated and ambiguous. FCM (Fuzzy C Means) is one of the most popular fuzzy clustering algorithms in 1990s. FCM can be regarded as a generalization of ISODATA and was studied by (Dunn 1973). FCM variants were also developed to deal with other data types, such as symbolic data and data with missing values (Bezdek 1981). Recently, Rough set theory became
quite prominent to explore uncertainties (Lotfabadi et al. 2013). The incomplete information system using Rough K-Means (Kryszkiewicz & Babinski 2000). Many clustering methods have been evolved which use the fuzzy set and rough set theories in the recent years (Xu et al. 2010). Kernel based methods have been proposed to identify non-isotropic and linearly inseparable clusters in the input space (Lin & Zheng 2009). These techniques transform the given data in the input space to a high dimensional feature space called the induced space and find the clustering in that space. The kernel trick, arising from Mercers theorem proposed in is used to compute the distance between the patterns in the kernel induced feature space (Ha Quang Minh et al. 2006). The kernel K-Means algorithm in the online mode is depicted by (Thomas Hofmann et al. 2008). (Inderjith S Dhillon & Joydeep Ghosh 2005) presented a unified view of the kernel and spectral methods. Hybrid clustering methods combine two or more methods where the result of one method is given as input to the subsequent method. I-DBSCAN, Rough-DBSCAN, Leader-single-link are some examples of this type (Viswanath & Suresh Babu 2009). Ensemble based clustering methods are inspired by the ensemble of supervised learning methods (Zhang & Yunqian 2012). These hybrid and ensemble methods have been used not only to improve the clustering quality but also to speed-up the clustering process (Yogita Rani, Dr.Harish Rohil 2013). The literature on cluster analysis is vast, and hundreds of clustering algorithms have been proposed in the literature. It requires a tremendous effort to list and summarize all the major clustering algorithms.

1.5 APPLICATIONS OF CLUSTERING

Data clustering is prevalent in any discipline that involves analysis of multivariate data. Clustering has been widely used in various application domains which include artificial intelligence, pattern recognition, machine learning, information retrieval, image processing, biology, data mining,
marketing, medicine, psychology, recommender systems and statistics. In image processing, it is used to segment texture in images to differentiate between various regions or objects. It is also used for data compression in image processing, which is also known as vector quantization (Mukesh Mittal & Ruchika Lamba 2013). In computer vision, several important problems can be formulated as clustering problems. Clustering has always been used in statistics and science. The classic introduction into pattern recognition framework is given in (Bishop 2006).

Typical applications include speech and character recognition (Bor-Shenn Jeng et al. 1989). The statistical approaches to pattern recognition (Hastie et al. 2009). Documents can be clustered to generate topical hierarchies for information access or retrieval. Clustering is also used to perform market analysis (Bakaev & Avdeenko 2014). Clustering in data mining was brought to life by intense developments in text mining (Weiss et al. 2010) spatial database applications, such as, GIS or astronomical data sequence and heterogeneous data analysis (Adam et al. 2012), Web applications (Vakali et al. 2014), DNA analysis in computational biology (Langmead B Grapnel et al. 2009) and many others.

1.6 VARIOUS DEFINITIONS OF DATA CLUSTERING ALGORITHMS

Clustering algorithms can be classified along different independent dimensions (Dean & Ghemawat 2008). For instance, different starting points, methodologies, algorithmic point of view, clustering criteria, and output representations, usually lead to different taxonomies of clustering algorithms (Li G, Brasy et al. 2013). Different properties of clustering algorithms can be described as follows:
(i) Agglomerative vs. Divisive Clustering: This concept relates to algorithmic structure and operation. An agglomerative approach begins with each object in a distinct (singleton) cluster, and starts merging clusters together until a stopping criterion is satisfied (bottom-up hierarchical clustering). On the other hand, a divisive method begins with all objects in a single cluster and iteratively performs splitting until a stopping criterion is met (top-down hierarchical clustering).

(ii) Monothetic vs. Polythetic Clustering: Both the monothetic and polythetic issues are related to the sequential or simultaneous use of features in the clustering algorithm. Most algorithms are polythetic; that is, all features enter into the computation of distances or similarity functions between objects, and decisions are based on those distances, whereas, a monothetic clustering algorithm uses the features one by one.

(iii) Hard vs. Fuzzy Clustering: A hard clustering algorithm allocates each object to a single cluster during its operation and outputs a Boolean membership function either 0 or 1. A fuzzy clustering method assigns degrees of membership for each input object to each cluster. A fuzzy clustering can be converted to a hard clustering by assigning each object to the cluster with the largest degree of membership.

(iv) Partitional vs. Hierarchical Clustering: A Partitional clustering algorithm obtains a single partition of the data instead of a clustering structure, such as the dendogram produced by a hierarchical technique. Partitional methods have advantages in applications involving large data sets for which the construction of a dendogram is computationally prohibitive.
(v) Distance vs. Density Clustering: A distance-based clustering algorithm assigns an object to a cluster based on its distance from the cluster or its representatives, whereas a density-based clustering grows a cluster as long as the density or number of objects in the neighborhood satisfies some threshold. Distance-based clustering algorithms fail at discovering clusters of arbitrary shape, whereas density-based clustering algorithms are capable of finding arbitrary shape clusters.

(vi) Deterministic vs. Stochastic Clustering: This issue is most relevant to partitional techniques designed to optimize a squared error function. Deterministic optimization can be accomplished using traditional techniques in a number of deterministic steps. Stochastic optimization randomly searches the state space consisting of all possible solutions.

(vii) Intermediate vs. Original Representation Clustering: Some clustering algorithms use an intermediate representation for dimension reduction when clustering large and high dimensional datasets. It starts with an initial representation, considers each data object and modifies the representation. These classes of algorithms use one scan of the dataset and its structure occupies less space than the original representation of the dataset, so it may fit in the main memory.

Data reduction can significantly improve the clarity of the resulting classifier model. The clustering helps in the removal of the most irrelevant and redundant features from the dataset.

The clustering therefore helps in data reduction which improves the performance of the learning models by 1. Reducing the effect of the curse of
dimensionality 2. Enhancing generalization capability 3. Increasing the speed of the learning process 4. Improving model interpretability (Keum & Lee 2010). Feature reduction also helps people to acquire better understanding of the datasets by indicating important features and the relationship among the various features. To mine various datasets there exists different kinds of tools in data mining. They are WEKA, CIS TOOLBOX in MATLAB, R miner, Rapid miner.

Data reduction using clustering in unsupervised learning is a challenging task due to the presence of irrelevant attributes and high dimensionality of the data set (Cherkassky & Mulier 1998). Thus this work proposes a comparison of two different clustering algorithms using filter named Normalize in different levels. The clustering algorithms are 1. Expectation Maximization algorithm (EM) and K-Means algorithm proposed under WEKA tool. This comparison of algorithms identifies the one which performs better than the other. Finally, the comparison of the two clustering algorithms identifies that the K-Means algorithm preprocessed using Normalize filter in the instance level performs better than the Expectation Maximization algorithm which is also preprocessed in both the instance and attribute levels. The different clustering algorithms works differently for different data sets. WEKA (Waikato Environment for Knowledge Analysis) is the one among the various data mining tools which best clusters the datasets in unsupervised learning.

1.7 SCOPE OF THE WORK

The major objectives of this thesis are

To find and achieve the better data reduction in large data sets, here the preprocessing filters, selection of attributes and clustering and classification algorithms are applied on the large data sets in different combinations.
In the first combination, the preprocessing filters is implemented on the data set as the basic process. Secondly applying the selection of attributes on the preprocessed data set to identify the first form data reduction. Then on the reduced data the clustering algorithms are applied and the results are recorded to find the best form of data reduction in terms of the clustering algorithm. Finally, the results of the clustering algorithms like EM (Expectation & Maximization) and K-Means are compared and analyzed. In the second combination the selection of attributes is applied on the large data set to find the first of data reduction. Then on the reduced data set the preprocessing filters are applied to get next reduced of data reduction and finally the two clustering algorithms are applied on the data set and the results are recorded. Now the recorded results obtained in the two different combinations are compared and analyzed to identify with the combination which best results in efficient data reduction. With this analysis the identified combination of data reduction method is carried out for the large data set.

In another combination, the two different clustering algorithms are used along with different preprocessing filters. Here the selected dataset is not preprocessed initially but treated by two clustering algorithms and the output of the clustering algorithms are recorded. To compare with these obtained results the large data set is now preprocessed by two different filters and implemented with two clustering algorithms. The aim is to find whether usage of preprocessing filters provides with better efficiency in terms of data reduction. Finally, the results are compared and concluded.

Similarly, the data reduction is done in classification in two combination forms to find the better one. In the first form the data set processed with filters, then on the reduced data two types of classifier methods are implemented and the output is recorded. Here the comparison for data reduction is carried out between classifier methods and the method which yields the minimum time is selected for data reduction in large data sets.
The research work presented here is aimed to achieve the better efficient data reduction using and comparing various preprocessing filters in the data mining techniques like classification, clustering and selection of techniques and identifying the better result by comparing among them.

1.8 ORGANIZATION OF THE THESIS

The rest of the synopsis is organized as follows: Section II illustrates the objectives of the proposed data reduction system. Section III reviews some of the existing works related data reduction using clustering and classification in supervised and unsupervised learning in data mining. Section IV presents the detailed description for the proposed K-Means, Expectation Maximization clustering algorithm, Classification algorithm and Preprocessing filters. Section V shows the performance and comparison results of the proposed system. Finally, this synopsis concluded and the future work to be carried out is stated in Section VI.