Chapter 1

INTRODUCTION

In today's world of Information Technology, the markets are barrierless and the reach of businesses is expanding beyond cities to entire nation, and even across the globe. Large organizations in Banking, Automobile, Agriculture, Education etc. use huge amount of data and classify it to understand demographies, consumption, usage patterns etc. These businesses also generate huge amount of data themselves, which is increasing exponentially. For taking right decisions at right time, this huge amount of data needs to be processed efficiently and accurately for correct interpretations and decisions. This entire philosophy of storing, maintaining, classifying and interpreting data, to find patterns or trends for better business decisions, is an upcoming area of research.

1.1 DATA MINING

Data Mining is a process of drawing out useful patterns or knowledge from the huge data collected in information systems and to use these patterns in taking safe and smart decisions. It helps organizations to take right decisions at right time. Data mining provides techniques to process large amount of data efficiently and presents it in the required form. The predefined methods and algorithms that are used to extract these useful patterns are called Data Mining Techniques. Some popular data mining techniques include Frequent Pattern Mining, Association Rule Analysis, Classification and Clustering.

Clustering is a technique of dividing the given dataset into groups or clusters such that the objects in one group are more similar to each other than the objects in the other group. Clustering has its applications in variety of domains like health sector (Kaur et al. 2006; Sharma et al. 2014), agriculture (Ananthara et al. 2013; Cao et al. 2014), E-Commerce (Cheng et al. 2009; Li Mei et al. 2010), education (Bresfelean et al. 2009; Hung et al. 2011) etc. For example, in an organization, grouping and identifying the products which
are not in high demand may help in reducing their production to cut losses. Further, in Educational Institutions, grouping the students according to their academic performance may help to identify the students with lower grades. Such students can be motivated to attend remedial classes to overcome their difficulties. In Health sector, data mining and clustering help in identifying the relationship between the diseases and symptoms. In Banking sector, it is used to group customers with overdue credit card payments. In Market Survey, clustering is utilized to identify customers having certain buying patterns. A lot of work is being done to apply these techniques in various other sectors, apart from the above mentioned (Athanasopoulou et al. 2011; Kostia 2009).

Many clustering algorithms have been proposed in the literature (Gan et al. 2007; Wu et al. 2007). These clustering algorithms are broadly classified into two categories, Hierarchical and Partitional. The hierarchical algorithms find clusters by arranging them into hierarchy (top-down or bottom-up), as a result they are not suitable for large datasets. On the other hand, the partition based clustering algorithms find the clusters independently. So they can easily partition large datasets. In this method, the given dataset containing n objects is divided into required number of clusters, K, where K≤n such that these clusters must be non-empty and mutually disjoint. This is an iterative method and the clusters once created are further refined by moving the objects from one cluster to another. This movement of objects is done based on value of some objective function. The K-Means algorithm is one of the commonly used techniques in this category.

K-Means is a simple algorithm known for its speed. The algorithm is inexpensive in terms of computational cost and works well with high dimensional and large datasets. However, it has some limitations. One major limitation is that the clusters produced are highly dependent on the objects initially selected as centroids (cluster centers). As the initial centroids are selected randomly, K-Means algorithm may not provide same result for different runs of the algorithm on same dataset. A lot of work has been done to overcome this limitation. Another limitation is the requirement to specify a pre-defined value of K (number of clusters) as input. Providing value of K is domain specific. Sometimes the dataset is unknown or new and in that case incorrect input of number of
clusters required leads to inefficient grouping of data. To overcome this limitation, researchers are still exploring some ways. These limitations of K-Means are carried forward to its extensions K-Modes and K-Prototype which work on categorical and mixed data respectively. Further, K-Modes faces considerable challenge to its efficiency from high dimensional datasets. The K-Modes algorithm becomes computationally expensive and quality of the clusters is reduced with the increase in dimensions. Researchers are exploring ways to overcome this limitation of K-Modes.

1.2 KNOWLEDGE DISCOVERY FROM DATA (KDD)

In today’s world information technology is everywhere, to maintain account records in banks, to maintain records of the patients in hospitals, to record academic performance of the students in educational institutes etc. In late 1980s, a new trend has come up to extract information from the meaningful data collected in information systems of the organizations to take smart and safe decisions. The process of extracting useful patterns or knowledge is called Data Mining or KDD (Han et al. 2011). In Data mining, the process of knowledge discovery is as shown in Figure 1.1.

![Figure 1.1: Data Mining as a step in Knowledge Discovery (Han et al. 2011)](image)

Thus, data mining can be defined as the process of automatic retrieval of significant information from a database having current and historical data, using predefined methods
and algorithms. These methods and algorithms together are called data mining techniques. Data mining uses approaches and methods from multi-disciplinary areas like Statistics and Artificial Intelligence (AI). So Data Mining can be considered as a set of computer-assisted method to explore the data statistically and intelligently.

The organizations now a days can use data mining techniques to answer the queries like “Which two products should be sold in scheme together to increase the sales of the company” or “Which course should be launched in the coming year to increase the number of admissions”. Most of the business sectors such as Health sector, Agriculture, Banking and Educational organizations are using data mining techniques to make analytical decisions.

1.3 DATA MINING TECHNIQUES
The next section gives brief overview of the various popular data mining techniques. The popular data mining techniques are Frequent Pattern Mining, Association Rule Analysis, Classification and Clustering.

1.3.1 Frequent Pattern Mining and Association Rule Analysis
In Frequent Pattern Mining the patterns that appear in the dataset frequently are extracted. The association among these frequent patterns is found out by generating Association Rules. Market Basket Analysis is the most popular example of this technique. In this example, the buying habits of the customers are analyzed and the items frequently purchased are recorded. Using this information of frequent itemsets, association rules are generated.

The Association Rule is of the form:

\[
\text{buys}(X, \text{“computer”}) \Rightarrow \text{buys}(X, \text{“printer”})
\]

\[\text{[support=20\%, confidence=60\%]}\]

This rule shows that a person who buys “computer” tends to buy “printer” at the same time. There are two measures to find the interestingness of the association rules, Support and Confidence. These measures are expressed in %. In the above example, a support of 20% shows that 20% of the customers purchase “computer” and “printer” together. Similarly, a confidence of 60% reflects the possibility that the customers who have
purchased computer may purchase printer also. Using this association rule, the manager of the store can offer some scheme so that the persons buying “computer” are prompted to buy the “printer” also. This will in turn lead to increase of sale of “printer”. The Apriori algorithm is one of commonly used technique for finding frequent itemsets from the given dataset, from these itemsets Association rules can be generated (Han et al. 2011).

1.3.2 Classification

Classification technique is used to segregate the data into different groups or classes. For performing classification on a given dataset, the dataset should have at least one attribute as categorical attribute which will represent the class. This technique is then applied to divide the given dataset into groups. The number of values in the categorical attribute determines the number of groups. The process of classification in groups has two steps:

i) Generation of Classification Rules

ii) Classification of data using Classification Rule

The attributes in the dataset other than the attribute representing the class are called Input parameters and the attribute which represents the class is called Target attribute. This concept can be better understood with the example shown in Table 1.1.

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senior</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Middle-aged</td>
<td>Low</td>
<td>Yes</td>
</tr>
<tr>
<td>Senior</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>Middle-aged</td>
<td>High</td>
<td>Yes</td>
</tr>
</tbody>
</table>

In this example, ‘Age’ and ‘Income’ are Input parameters and ‘Buys_computer’ is Target parameter.

The first step of classification generates two rules:

If Age= Senior and Income=High Then Buys_computer =Yes

If Age=Middle_Aged Then Buys_computer =Yes
The first step is also called Supervised Learning as it is known to which class each object in the dataset belongs.

The second step classifies the data whose class is unknown. Suppose a new object is added in the dataset with value of **Buys_computer** as unknown as shown in Table 1.2.

<table>
<thead>
<tr>
<th>Age</th>
<th>Income</th>
<th>Buys_computer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Senior</td>
<td>Low</td>
<td>No</td>
</tr>
<tr>
<td>Middle aged</td>
<td>Low</td>
<td>Yes</td>
</tr>
<tr>
<td>Senior</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>Middle aged</td>
<td>High</td>
<td>Yes</td>
</tr>
<tr>
<td>Senior</td>
<td>High</td>
<td>?</td>
</tr>
</tbody>
</table>

Using the classification rules generated above, missing value will be predicted as “**Yes**”.

Some of the popular classification algorithms are ID3, C4.5 and Fuzzy C4.5 (Han et al. 2011).

### 1.3.3 Clustering

Clustering is a technique of dividing the given dataset into groups or clusters such that the objects in one group are more similar to each other than the objects in the other group. It differs from Classification in a way that it is unsupervised while Classification is supervised. Clustering is unsupervised as it does not depend on predefined classes and generation of rules from the training dataset. Some of the clustering methods are discussed below:

**Partitioning Method:**

In this method, the given dataset containing n objects is divided into required number of clusters, K, where K≤n such that these clusters must be non-empty and mutually disjoint. This is an iterative method and the clusters once created are further refined by moving the objects from one cluster to another. This movement of objects is done based on value of some objective function. The famous algorithms under this category are: K-Means and K-Medoids.
The advantage is that all the algorithms falling in this category are scalable. The limitation of this approach is that the algorithms under this category work well when clusters are globular. Also, this method is sensitive to noise.

**Hierarchical Method:**

In this method, the objects in the given dataset are organized into a tree of clusters. There are two ways of using this method:

- **Agglomerative method:** This method starts by putting every object in the dataset into a separate cluster. These clusters are then merged into large clusters until a single cluster containing all the objects is formed. AGNES (Agglomerative Nesting) algorithm is an example of this approach (Han et al. 2011).

- **Divisive method:** This method starts by putting all the objects in a single cluster which is a reverse of Agglomerative approach. This cluster is then divided into smaller clusters until the unit clusters are obtained. DIANA (Divisive Analysis) comes under this category (Han et al. 2011).

The advantage of Hierarchical method is that a tree of clusters is obtained and user can see the desired number of clusters by watching the tree at different levels. This approach works well when the clusters are globular. Also algorithms following this approach are sensitive to noise and outliers.

**Density-Based Method:**

As the name suggests, the method requires an initial input of density parameters like minimum number of objects needed to make a cluster and the minimum distance between objects within a cluster. The algorithms under this category produce clusters of arbitrary shape and can handle noise. The limitation of this approach is that the algorithms require some input parameters which sometimes become difficult to predict. The famous algorithms of this category are: DBSCAN and OPTICS (Han et al. 2011).
Grid-Based Method:

This method uses a multi resolution grid structure. The objects in the dataset are divided into finite number of cells. The density of each cell is calculated and then cells are sorted according to their densities.

The major advantage is that this method is efficient in terms of computational complexity. The limitation of this method is that the clusters produced are of low quality. The famous algorithm of this category is: STING (Han et al. 2011).

In our study, we shall focus on the famous partitioning method based clustering algorithm: K-Means and its extensions, K-Modes and K-Prototype.

1.4 INTRODUCTION TO K-MEANS ALGORITHM

The K-Means is a partition based iterative algorithm. This algorithm divides the given dataset into required number of clusters (K). The procedure starts with the random selection of K centroids one for each cluster. The objects are moved from one cluster to another and centroids are recomputed until all the objects in each cluster are at their minimum distance from their centroids. In this technique Euclidean distance is taken as the distance measure given in Equation 1.

\[ d(x, c) = \sum_{i=1}^{n} \sqrt{(x_i - c_i)^2}, \]

where \( x = (x_1, x_2, \ldots, x_n) \) is an object in a cluster and \( c=(c_1, c_2, \ldots, c_n) \) is the centroid of a cluster.

The basic algorithm works as follows:

<table>
<thead>
<tr>
<th>Algorithm: Basic K-Means</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong> Numerical dataset ((D)), Number of clusters ((K))</td>
</tr>
<tr>
<td><strong>Output:</strong> Elements of dataset classified into K clusters</td>
</tr>
</tbody>
</table>

1. Select K random points as initial centroids
2. Repeat
3. Create K clusters by assigning all data points to the closest centroid
4. Recompute centroids (mean value of the objects in the cluster) for each cluster to improve accuracy
5. Until the cluster centroids don’t change
The K-Means algorithm has been chosen by the author because of the following reasons:

- The K-Means algorithm is inexpensive in terms of time.
- Highly cohesive clusters are produced by the algorithm.

The algorithm has some limitations also:

- The final clusters produced are highly sensitive to initial centroids because the initial centroids are chosen randomly.
- The algorithm is sensitive to outliers because centroids are calculated as mean of the values lying in that cluster.
- The algorithm requires an initial input of K. Providing this value becomes difficult sometimes because the person using the algorithm may not be the domain expert.

Various extensions of K-Means have been discussed in literature to overcome the above mentioned limitations (Huang et al. 2005; Jain et al. 2012; Li et al. 2008; Li 2008; Panda et al. 2012; Singh et al. 2011; Visalakshi et al. 2009; Yuepeng et al.2011).

1.5 K-MODES ALGORITHM

The K- Modes extends the K-Means algorithm to cluster categorical data in the following manners (Khan et al. 2013):

- A Simple Matching Dissimilarity function suitable for categorical data is used instead of Euclidean distance.
- Modes are used to represent centroids instead of Mean values.
- A frequency based method is used to find centroids in each iteration of the algorithm.

The K-Modes algorithm carries in itself the benefits and limitations of the K-Means because it follows the same iterative process as that of K-Means. To overcome these

Basic K-Modes works as follows:

<table>
<thead>
<tr>
<th>Algorithm : Basic K-Modes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input</strong>: Categorical dataset (D), Number of clusters (K)</td>
</tr>
<tr>
<td><strong>Output</strong>: Elements of dataset classified into K clusters</td>
</tr>
<tr>
<td>1. Select arbitrarily K objects as initial centroids.</td>
</tr>
<tr>
<td>2. Allocate an object to the cluster whose mode is the nearest to it according to:</td>
</tr>
<tr>
<td>[ d(x,c) = \sum_{i=1}^{n} \delta(x_i, c_i) ]</td>
</tr>
<tr>
<td>where ( x=(x_1, x_2, \ldots, x_n) ) is an object in a cluster, ( c=(c_1, c_2, \ldots, c_n) ) is the centroid of a cluster and ( \delta ) is Kronker delta defined by</td>
</tr>
<tr>
<td>[ \delta(i,j) = \begin{cases} 0 &amp; \text{if } i = j \ 1 &amp; \text{if } i \neq j \end{cases} ]</td>
</tr>
<tr>
<td>Update the centroids (modes of the objects in the cluster) after each allocation.</td>
</tr>
<tr>
<td>3. After all objects have been allocated to clusters, retest the dissimilarity of objects against the current modes. If an object is found such that its nearest mode belongs to another cluster rather than its current one, reallocate the object to that cluster and update the modes of both clusters.</td>
</tr>
<tr>
<td>4. Repeat (3) until no object has changed clusters after a full cycle test of the whole dataset.</td>
</tr>
</tbody>
</table>

1.6 K-PROTOTYPE ALGORITHM

Earlier the clustering algorithms focused on single type of dataset either all attributes numerical or all attributes categorical. As mixed datasets are becoming common in real life, so clustering algorithms are being developed to cluster these datasets.

K-Prototype is a variant of K-Means that can be used with mixed datasets. K-Prototype uses the basic idea of the K-Means by applying Euclidean distance to numeric attributes.
and Binary distance to categorical attributes. The K-Prototype algorithm works as follows:

**Algorithm : Basic K-Prototype**

**Input:** Mixed dataset (D), Number of Clusters (K)

**Output:** Elements of Dataset classified into K clusters

1. Select arbitrarily K objects as initial centroids.
2. Allocate each object in D to a cluster whose centroid is nearest to it according to the following distance measure:
   \[ d(X_i, Q_l) = \sum_{j=1}^{m_r} (x_{ij}^r - q_{lj}^r)^2 + \gamma_l \sum_{j=1}^{m_c} \delta(x_{ij}^c, q_{lj}^c), \]  \( (3) \)

   Where \( X_i = (x_{i1}, x_{i2}, \ldots, x_{in}) \) is ith object in a cluster, \( Q_l = (q_{l1}, q_{l2}, \ldots, q_{ln}) \) is centroid of cluster \( l \) and \( \delta \) is Kronner delta defined by
   \[ \delta(i, j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases} \]

   \( x_{ij}^r \) and \( q_{lj}^r \) are values of numeric attributes for object i and the centroid of cluster \( l \),
   \( x_{ij}^c \) and \( q_{lj}^c \) are values of categorical attributes for object i and the centroid of cluster \( l \),
   \( m_r \) and \( m_c \) are the numbers of numeric and categorical attributes,
   \( \gamma_l \) is a weight for categorical attributes for cluster \( l \).

   Update the centroid (mean for numerical attributes and mode for categorical attributes) of the cluster after each allocation.

3. After all objects have been allocated to a cluster, retest the similarity of objects against the current centroid. If an object is found such that its nearest centroid belongs to another cluster rather than its current one, reallocate the object to that cluster and update the centroids of both clusters.

4. Repeat (3) until no object has changed clusters after a full cycle test of the whole dataset.

However, taking the binary distance for categorical attributes does not give accurate results because in many real life situations the categorical values may have some other degree of difference rather than just 0 or 1. To overcome this limitation various extensions of the K-Prototype algorithms have been presented in the literature (Hunt et al. 2011; Lui et al. 2006; Luo et al. 2006; Naija et al. 2008; Reddy et al. 2012; Roy et al.
2010; Sen et al. 2013; Shih et al. 2010; Wang et al. 2015; Yang et al. 2015; Yin et al. 2005).

1.7 CHALLENGES
Attempts have been made in the literature to deal with the limitations of the K-Means algorithm and its extensions. These limitations are discussed in section 1.1. Some of the work to overcome the limitation of providing number of clusters required as an input in K-Means algorithm and its extensions is discussed below:

Pelleg, Dan et al. (2000) suggested X-Means algorithm as an extension of K-Means which overcomes the limitation of inputting the value of K and is faster than the original K-Means. The algorithm starts with K as the lower bound of the given range and continues adding centroids until the upper bound is reached. During this process the centroid set that scores the best is recorded using the kd-tree data structure. The algorithm requires an input range suggesting the lower and upper bound of K.

Leela, V. et al. (2013) introduced Y-Means algorithm based on K-Means algorithm. Initially, it runs K-Means algorithm on the dataset and then follows the sequence of splitting, deleting and merging the clusters. This algorithm depends on K-Means algorithm to find the clusters initially.

Shafeeq, Ahamed, B., M. et al. (2012) presented an algorithm which overcomes the problem of providing the required number of clusters by finding the optimal number of clusters on the run. The user has to input the number of clusters (K) as 2 in the first run.

Abubaker, Mohamed et al. (2013) presented a new approach to overcome two of the major limitations of K-Means algorithm: One to select efficiently the initial centroids and second to remove the need of giving the number of clusters required as input to the algorithm. Their algorithm is based on the k-nearest neighbor method. They have suggested two versions of the algorithm, the first version takes $k_n$ (the number of nearest neighbors), and the number of clusters $k$ in the dataset as input parameters. Then the sorted list of data points is investigated until the number of obtained prototypes reached
the specified k. The second version takes $k_n$ as the only input parameter and obtains both the number of clusters and the prototypes simultaneously.

Liao, H. et al. (2009) in their algorithm extended the K-Modes clustering algorithm by introducing a regularization parameter to control the number of clusters in a clustering process. A suitable value of regularization parameter is chosen to generate the most stable clustering result.

Cheung, Y. et al. (2013) presented a similarity metric that can be applied to categorical, numerical, and mixed attributes. Based on this similarity metric an iterative clustering algorithm is developed to overcome the limitation of inputting K. The algorithm requires some initial value of K which should not be less than the original value of K.

Liang, Jiye et al. (2012) extended K-Prototype algorithm by proposing a generalized mechanism for characterizing within-cluster entropy and between-cluster entropy to identify the worst cluster in a mixed dataset, an effective cluster validity index to evaluate the clustering results and the K-Prototype algorithm with a new dissimilarity measure. The algorithm requires input parameters representing the minimum and maximum number of clusters that can be generated from the dataset.

All the methods discussed above for removing the limitation of providing the value of K initially in K-Means algorithm require some parameter other than K as input. However, the method suggested by Tibshirani et al. (2000) overcomes this limitation by using the technique of Gap Statistic. This technique utilizes the output generated by any clustering algorithm to compare the change in within cluster dispersion with that expected under an appropriate reference null distribution. This Gap method works well when the clusters are well separated. Also, the method is computationally complex. The new clustering technique named STep-wise Automatic Rival penalized (STAR) suggested by Cheung (2003) overcomes two major limitations of K-Means: One to select efficiently the initial centroids and second to remove the need of giving the number of clusters required as input to the algorithm. The algorithm consists of two separate steps. The first step provides each cluster center. The next step then adjusts the units adaptively by a learning rule. The limitation of this algorithm is the complex computation involved in it. These two methods work with numeric data.
All the suggested algorithms discussed above require some parameter to be input to perform clustering for the given dataset. This motivated the author to propose new algorithms.

1.8 MOTIVATION

The K-Means algorithm and its extensions K-Modes and K-Prototype have the following limitations:

- The final clusters produced are highly sensitive to initial centroids.
- The K-Means algorithm is sensitive to outliers because centroids are calculated as mean of the values lying in that cluster.
- The algorithms require an initial input of K. Providing this value becomes difficult sometimes because the person using the algorithm may not be the domain expert.
- The Binary distance for categorical attributes does not give accurate results because in many real life situations the categorical values may have some other degree of difference rather than just 0 or 1.
- The real world datasets with high dimensionality poses a considerable challenge to the efficiency of K-Means clustering especially for its extension K-Modes. The K-Modes algorithm becomes computationally expensive and quality of the clusters is reduced with the increase in dimensions.

These limitations motivated the author to extend the K-Means algorithm and its extensions to make them more efficient.

1.9 OBJECTIVES

- To extend the K-Means for numerical datasets algorithm to overcome the limitation of providing the number of clusters at the beginning of the algorithm.
• To modify the K-Modes algorithm to cluster categorical datasets so as to overcome the limitation of inputting the number of clusters required initially and work efficiently for high dimension datasets.
• To transform the K-Prototype algorithm for mixed datasets to overcome the limitation of providing the number of clusters required initially.
• To compare the accuracy of the clusters produced by the proposed algorithms with that of the original K-Means, K-Modes, K-Prototype and various other algorithms on different real world datasets of different sizes and dimensions.

1.10 METHODOLOGY

The new algorithms based on the K-Means algorithm are proposed to overcome the limitation of providing the value of K as input for numerical, categorical and mixed datasets. An algorithm which does not require K as input and works efficiently for high dimension datasets for categorical attributes is proposed. The proposed algorithms are implemented in C#.

The accuracy of the clusters produced by the proposed algorithms is compared with the original K-Means, K-Modes and K-Prototype algorithm and their various extensions presented by various authors. The results of the original K-Means, K-Modes and K-Prototype are obtained using software RapidMiner.

RapidMiner an open source solution is a complete business analytics workbench with a strong focus on data mining, text mining, and predictive analytics. It uses a wide variety of descriptive and predictive techniques to give you the insight to make profitable decisions. RapidMiner together with its analytical server RapidAnalytics also offers full reporting and dash boarding capabilities and, therefore, a complete business intelligence solution in combination with predictive analytics (rapidminer.com).

The experiments are performed on real datasets from (UCI Machine Repository: A website that maintains 300 datasets as a service to the machine learning community) and (KEEL data repository: A website that maintains 908 data sets).
1.11 CONTRIBUTION

The contribution of the thesis is to present new algorithms based on the partition based clustering algorithm K-Means, K-Modes and K-Prototype but with advance features of efficient data analysis and automatic generation of appropriate number of clusters for numerical, categorical and mixed datasets. The developed algorithms also generate clusters that are comparable to the clusters generated by the original algorithms and various other algorithms. An algorithm that extends the K-Modes algorithm that does not requires K as input and also works efficiently for high dimension datasets is also presented.

1.12 ORGANIZATION OF THESIS

The research work reported in the thesis have been organized into six chapters as given below:

Chapter 1 explains the basic concepts of Data Mining, Knowledge Discovery from data, Data Mining techniques, Clustering methods, K-Means, K-Modes and K-Prototype algorithms along with the Challenges, Motivation, Objectives, Methodology and Contribution of the research work.

The aim of Chapter 2 is to present the review of existing literature relevant to the theme of the research work.

In Chapter 3, the design and implementation of the proposed algorithm based on the K-Means algorithm to cluster the numerical data without providing any input to the algorithm is discussed. The results generated by the proposed algorithm have been compared with that of original K-Means and various other algorithms on many datasets.

The objective of chapter 4 is to discuss the design and implementation of the proposed algorithm based on the K-Modes algorithm to cluster the categorical data without providing the value of K initially. Further, the algorithm is extended to work on high dimension datasets. The results generated by the extended algorithms are compared with those generated by the original K-Modes and various other algorithms on datasets of
different sizes and dimensions. The datasets are taken from UCI and KEEL data repository.

In Chapter 5, the new algorithm based on the K-Prototype algorithm for clustering the mixed datasets without providing the value of K initially is discussed. The results are compared with those generated by the original K-Prototype and various other algorithms presented in the literature.

In Chapter 6, the conclusions of the work reported in the previous chapters along with the future scope for the work is discussed.

Towards the end, references of various research papers cited in the thesis have been given.