Chapter 1

INTRODUCTION

1.1 Prologue

The development of hardware, software and scientific advancements made the rapid computerization of scientific activities and business. Scientific advancements have made easy to collect the data and digitizing the information which can be stored in database. This makes the collection and storing the data to grow at a phenomenal rate [5]. Such a raw data is rarely of direct use. Its true value is predicted on the ability to extract information useful for decision support or exploration by understanding the phenomenon governing the data source. One or more analysts may be intimately familiar with the data and with the help of computing techniques they can provide summaries and generate reports. Hence the analysts are acting as a sophisticated query processor. However such a manual query processing has its own limitation depending upon the size of the data and dimensions. Since
the scale of data manipulation, exploration and inferencing go beyond human capacities, advanced computing technologies become inevitable for automating process [92].

In order to identify hidden patterns or potential information from the large database, it is necessary to make partition of the large database using intelligent automation process. Thus the field of clustering has emerged. The clustering is defined as the problem of grouping a collection of similar objects into a set of natural groups without any prior knowledge. It is a technique, which is used to find the group of clusters that are some how similar in characteristics from the given data set for which the real structure is unknown. The clustering becomes intelligent clustering when it is applied for machine learning. Intelligent clustering has immense number of applications in every field of life. Especially, it plays an important role in the fields of Data Mining, Image Processing, Information Retrieval, Mobile Computing and Pattern Recognition [92].

The rest of the chapter is organised as follows: section 1.2 details the motivation for this thesis. The applications of clustering algorithms are presented in section 1.3 and section 1.4. The clustering process for pattern recognition is explained in section 1.5. Section 1.6 explains the steps involved in clustering algorithm. The research
opportunities are discussed in section 1.7 and the contributions are concisely given in section 1.8. And section 1.9 concludes the chapter with summary of the structure of thesis.

1.2 Motivation

Partitioning a set of objects into homogeneous clusters is a fundamental operation in pattern recognition [19] and this operation is needed in a number of machine learning tasks, such as data extraction, reduction, summation and unsupervised classification. It is also needed in segmentation of large heterogeneous data sets into smaller homogeneous subsets that can be easily managed, separately modelled and analyzed [61]. Clustering is a popular approach used to implement the partitioning operation and it provides an intelligent way of finding interesting groups when a problem becomes intractable for human analysis.

General references regarding clustering are explored in [43, 52, 59, 66, 80, 82, 89, 92, 99, 116, 131]. Clustering has always been used in statistical analysis [3, 47, 60] and it is viewed as a density estimation problem in multivariate statistical estimation [153]. It is also used in several exploratory pattern analysis, grouping and decision
making situations including data mining [2, 35, 53, 157, 185]; spatial database applications, for example, Geographical Information System data or astronomical data, [55, 154, 176], sequence and heterogeneous data analysis [25], Web applications [63, 77], and many others. Clustering is widely used for image segmentation [65, 91]. Data fitting in numerical analysis provides still another avenue in clustering [50]. The classical introduction into pattern recognition framework is discussed in [45]. Typical applications of pattern recognition include speech and character recognition.

1.3 Applications

Data clustering is a central task of Artificial Intelligence (AI) [91], Human beings are very good at finding clusters in low dimension with small amounts of data, but it is very difficult to instruct a computer to find such relationships. It is an interesting problem to solve from AI perspective. Here are some scientific and real life applications of data clustering [92].

i. Natural resources study and estimation.

ii. Agriculture: Automatic analysis of satellite pictures are used to determine the type and condition of agricultural crops, weather conditions, snow and water reserves, and mineral prospects.
iii. Hydrology.

iv. Forestry.

v. Geology: Automatic analysis of substances such as rock and soil are helpful to obtain the structure, origin and history of the earth.

vi. Environment.

vii. Cloud pattern.

viii. Urban quality.

ix. Stereological applications:

• Metal Processing.

• Mineral Processing.

x. Biology: In this, clustering can be used to define taxonomies, categorize genes with similar functionality and gain insights into structures inherent in populations.

xi. Bio-medical Applications:

• Electrocardiograph

• Electrocardiogram

• Electromyogram Analysis

• Cytological, historical and other stereological applications

• X-ray analysis, mammogram analysis
• Magnetic Resonance Image analysis
• Diagnostics

xii. Man machine communication:

• Automatic speech recognition: Speaker identification
• Cursive script recognition system
• Optical character recognition systems
• Speech understanding
• Image understanding

xiii. Business: In business, clustering helps marketers to discover significant group in their customers database and characterise them based on purchasing patterns.

xiv. Web searching: In this case, clustering is used to discover significant groups of documents on the web page collection of semi structured documents. This classification of web documents assists in information discovery.

xv. Spatial data analysis: Owing to huge amount of spatial data that may be obtained from satellite images, medical equipment, Geographical Information Systems, image database exploration etc., examining spatial data in detail has become expensive and difficult for the users. It is used to identify and extract
interesting characteristics and patterns that may exist in large spatial databases.

xvi. Data Compression: In data clustering, each cluster has some set of objects that belongs to that cluster. If there is no need to store all the objects we can represent the set with abstract descriptions about the whole set of objects, then the objects of each cluster may be replaced with a set of descriptions.

1.4 Specific Applications

Underneath these specific application topic just discussed, the researcher would like to elaborate on some specific and important tasks in which data clustering plays a vital role.

Image Segmentation for Computer Vision

One of the most important issues in building intelligent, autonomous systems is that of image understanding. In order to understand an image, the first task of a computer is to segment the image into several parts. In a satellite image, one may like to distinguish the images such as building, water, forest and agriculture. Data clustering plays an integral role in image segmentation algorithms.
Customer Segmentation

The need to analyse data for decision making is growing exponentially, since data collections through electronic versions grow rapidly. Thus the field of data mining has emerged as the intersection of statistics, databases and machine learning for development of techniques to obtain information and knowledge from vast amount of micro-data, which are of numerical or categorical or mixture in nature.

1.5 Clustering Process for Recognition

A clustering operation involves the grouping of like objects/patterns/data with one another without any knowledge of pattern identity beforehand. Classification problems generally have this information. In the clustering problem, patterns must be separated solely on their specific attributes, whereas the classification problems have an access to error signals, which can be generated to guide the decision making of the machine. The clustering process is normally driven by similarity measure that the pattern in the same cluster are more similar to each other than the pattern in the different cluster according to some defined criteria.

The clustering process may result in different partitioning of a data set, depending on the specific criterion used for clustering. Thus,
there is a need of preprocessing before assuming a clustering task in a data set. The basic steps involved in clustering process are depicted in Figure 1.1.

The steps of clustering process [62] are summarised as follows:

**Feature selection**

The goal is to select the features properly on which clustering is to be performed so as to encode as much information as possible concerning the task of our interest. Thus, preprocessing of data may be necessary prior to their utilization in clustering task.

**Clustering algorithm**

This step refers to the choice of an algorithm that results in the
definition of a good clustering scheme for a data set. A proximity measure and a clustering criterion mainly characterize a clustering algorithm as well as its efficiency to define a clustering scheme that fits the data set. In most of the cases one has to ensure that all selected features contribute equally to the computation of the proximity measure and there are no features that dominate others.

Validation of the results

The accuracy of clustering algorithm results are verified using appropriate criteria [147], Since clustering algorithms define clusters that are not known well in advance, irrespective of the clustering methods, the final partition of data requires some kind of evaluation in most applications.

Interpretation of the results

In many cases, the experts in the application area have to integrate the clustering results with other experimental evidence and analyse in order to draw the right conclusion.

If each pattern is envisioned as a point in n-dimensional space, such as each star is seen as a point in the sky, then the stars in the sky are grouped, what measurement one has to use for this
determination?. The exact formulation of this answer often depends on the application of clustering [130, 145]. In some instances, the distances between patterns can be used to separate patterns. In other cases, pattern density in regions of space is used to indicate locations where patterns likely are drawn from the same class, the techniques known as histogram approaches.

It should be clear already that no clustering operation can ever be guaranteed to operate without error. The successful operation of the clustering method relies on the separability of the data from the attributes used as the pattern vectors. If two or more clusters overlap in n-dimensional space, they will never be perfectly separated. All clustering problems rely on the fidelity of these input data and generally are based on the separation from highly dense regions of patterns from one another. Each of these modes of the distribution is assigned a particular class label at a later time.

A pattern recognizer is a system to which a feature vector is given as input, and which operates on the vector to produce an output that is the unique identifier such as name, number, code word, vector, string, etc., associated with the corresponding group of objects. In a larger sense, each individual object is an atomic cluster, so that recognition includes identification. The pattern recognition process given may be
concrete and in that the recognition is based upon measurements of physical attributes. In the world of ideas, patterns are based upon the attributes of concepts and mental models, which is abstract pattern recognition. The work is more concerned here with concrete rather than abstract pattern recognition, but distinction is blurred and is one of degree.

1.5.1 Mathematical Preliminaries

A simple, formal, mathematical definition of clustering, is as follows: Let $x \in R^{N \times n}$ a set of data items represent a set of $N$ points $x^i \in R^n$. The goal is to partition $N$ points into $K$ groups such that every data that belongs to the same group are more “alike” than data in different groups. Each of the $K$ group is called a cluster.

To define the problem formally, Figure 1.2 depicts a population $P$ of non-identical objects $(x^1, x^2, \ldots, x^N)$, along with the processing that recognizes a sample object $x^i$. The attributes of each object are sensed or measured to yield a feature vector, $x^i = (x^{i1}, x^{i2}, \ldots, x^{in})$

The objective of clustering is to find partitions of $P$ which divides the objects of $P$ into $K$ disjoint clusters. For a given $N$, the number of possible partitions is definite but extremely large [52]. It is
Figure 1.2: Pattern Recognition Process
impractical to investigate every partition in order to find a better one for a classification problem. A common solution is to choose a clustering criterion to guide the search for a best partition. A clustering criterion is needed to find the best partition and that is called a cost function. The widely used cost function $\mu$ for $K$ clusters $C_1, C_2, \ldots, C_K$ is the clustering metric which is as follows:

$$\mu(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{x^j \in C_i} d(x^j - z^i),$$

where $C_i$ are clusters, $z^i$ are cluster centers and $d$ is a similarity measure often defined as the square Euclidean distance. Here, $z^1, z^2, \ldots, z^K$ are the representative vectors or prototype for each cluster or centroid, and $x^j$ is an element of a cluster $C_i$. Finding optimum $\mu$ is the clustering optimization problem [116, 117, 20, 159].

1.6 Clustering Algorithm

This section describes an example of centroid based partitional clustering algorithm viz., widely used K-means and its issues. The variations of centroid based partitional algorithms are reviewed in the subsequent chapters.

The K-means clustering algorithm is very popular for data
clustering, since K-means has been used in a wide assorted applications. K-means clustering models were introduced in 1967 by MacQueen [126] and they are considered as unsupervised classification technique. Earlier a minimum distance criteria is used in this method. The K-means algorithm is the most popular clustering tool used in scientific and industrial applications. The name conies from representing each of K clusters by the mean z of its points, the centroid. It is very simple to implement, it is fast, and it is fairly easy to understand [10]. The steps involved in the K-means algorithm can be summarised as follows:

The step one randomly picks up K feature vectors as centers for K clusters. The step 2 finds similar feature vectors for each cluster by Euclidean distance. Updating the cluster centroid helps to make more similar and clear cluster. The K-means algorithm divides the feature vectors into K clusters by minimizing the total within the class sum of squares at step 3 and then it halts the process. Most K-means type algorithms have been proved convergent [15, 126, 156]. The K-means algorithm has the following important properties.

i . It is efficient in processing large data sets. The computational complexity of the algorithm is O(tkmn), where m is the number of attributes, n is the number of objects, k is the number of clusters and t is the number of iterations over the whole data set. Usually,
k. m. t < II. In clustering large data sets the K-means algorithm is much faster than the hierarchical clustering algorithms whose general computational complexity is $O(n^2)$ [123].

ii. It often terminates at a local optimum [126, 156]. To find the global optimum, techniques such as deterministic annealing [93, 144] and genetic algorithms [121] can be incorporated with the K-means algorithm.

iii. It works only on numeric values because it minimizes a cost function by calculating the means of clusters.

iv. The clusters have convex shapes [4], Therefore, it is difficult to use the K-means algorithm to discover clusters with non-convex shapes.

v. The main difficulty in using the K-means algorithm is to specify the number of clusters.

1.7 Research Opportunities

The class of iterative optimization algorithms for data clustering are very popular both in research and applications. They are simple to implement; fast to execute and can provide very meaningful and interpretable results. Even in the case of more sophisticated algorithm, iterative optimization clustering algorithms play an integral role and
are still important [38, 105]. Therefore, any improvements which we can make to iterative optimization algorithms will bring benefits in many other areas as well.

The wide popularity of K-means algorithm is well deserved. It is simple, robust, and is based on the firm foundation of analysis of variances. Yet the K-means algorithm and its variants [170] also suffer from all the usual suspects as follows:

• The result strongly depends on the initial guess of centroids.

• Computed local optimum is known to be a far cry from the global one.

• It is not obvious what is a good K to use.

• Curse of dimensionality.

• The algorithm lacks scalability.

• Only numerical attributes are covered.

Hence in the above aspects, the K-means algorithm and its variants have to be revised to solve the problem in the above corresponding domain.
1.8 Research Contributions

The amount of data are increasing almost every year in real life cases. Hence, there is an urgent need for a new generation of computationally intelligent techniques and tools to assist man in extracting useful hidden information (knowledge) from the rapidly growing volume of data.

In this thesis, centroid-based partitional clustering algorithm and intelligence based approach which have simple and fast characteristics are further studied with suitable modifications without losing their characteristics. The statistical measures such as standard deviation and combined standard deviation, intelligent computing approaches [14, 24] viz., Neural Networks and Rough sets are proposed in this thesis to overcome the issues faced in the existing clustering algorithms. These approaches focus on finding the best clusters and estimate the number of clusters automatically. The proposed algorithms perform well and optimize the clustering metric than the widely used K-means and its variant algorithms [100]. Further, the proposed algorithms are applied to segment the medical image, colour image and customer database with categorical data.
1.9 Structure of the Thesis

The thesis is conveniently organized into nine chapters and the chapter one is introductory in nature. The rest of the chapters discuss the existing and proposed algorithms with applications. The gist of each chapter is provided here under:

**Chapter 2 : Related works in clustering**

A detailed survey of the state of the art in various partitional clustering techniques is studied and this chapter identifies the drawbacks in the traditional approaches based on centroid.

**Chapter 3 : Intelligent Statistical Clustering**

In this chapter, it is analysed that what makes clustering good in finding high quality clusters and the number of clusters. The proposed approaches overcome the issues of K-means method based on the standard deviation and combined standard deviation [166, 167, 168] and they are analysed for different data sets.

**Chapter 4 : Neural Network Clustering**

This chapter discusses the Neural Network clustering techniques used for learning which exhibits competitive form of behaviour. A
popular K-means clustering based Kohonen’s Neural Network (KNN) is discussed in detail. Weighted Kohonens Neural Network (WKNN) Algorithm and CSD based Neural Network Algorithm are also proposed [161, 162, 163]. It is observed that the proposed algorithms resulted with significant performance than the KNN clustering algorithm for different, data sets. The CSD based Neural Network clustering algorithm automatically finds the number of clusters with the significant results whereas the number of clusters have to be specified in advance in KNN and WKNN.

Chapter 5 : Rough Clustering

Rough clustering is attracting and creating interest among researchers [110]. This chapter describes the properties of rough sets for developing interval representations of clusters and it also compares the proposed approaches of Weighted Rough Neural Network and Kohonen’s Rough Neural Network based on CSD. The experimental result shows that the KRNNCSD algorithm achieves the best performance.

Chapter 6 : Mammogram Image Segmentation

Image segmentation is an important challenging task in medical image analysis. Particularly microcalcification is a complex structure in
mammogram. The proposed algorithms [165] in the previous chapters are applied to segment the mammogram image in this chapter. The results are analysed and compared with K-means and its variants.

Chapter 7 : Colour Image Segmentation

The colour image segmentation has been used in many applications since the much of technology for creating and displaying colour is based on the observation that a wide variety of colours can be obtained by mixing basic colours of light such as Red, Green and Blue (RGB) in different proportions. Man perceives object surfaces in scene in spite of shading and highlight effects. Hence the perception of objects in the real world without illumination effect has been a major concern for research community. This chapter compares the perception of objects using true colour image segmentation [164] by K-means clustering algorithm, its variants and proposed algorithms. It is observed that the proposed algorithms are performed well over the widely used clustering algorithms.

Chapter 8 : Customer Segmentation

Online business makes huge data about customers and face difficulties in understanding the customers and their behaviours. It is worthwhile if the behaviours of customer’s are known in advance for similar group of
customers. Hence, it is necessary to segment the customer database. These customer databases are usually mixed in type viz., numerical and categorical. The centroid based methods are often limited to numerical data. This chapter presents a novel representation for clustering algorithm to segment the customer database with categorical data. The computational results show that the proposed algorithm produces better result than the existing clustering algorithms.

Chapter 9: Conclusions and Future Work

This chapter summarises the key findings of the research work carried out and it also provides the directions for further research.