CHAPTER 1
CHAPTER 1
INTRODUCTION

This chapter gives an overview about data mining and the role of clustering in Data Mining. Problem specification being a sensitive area in Data Mining requires identification, usage and occurrence of problem. This chapter discusses the motivation, the objectives and the scope of the research work carried. The organization of the thesis is finally formulated

1.1 DATA MINING

Data mining has attracted a significant amount of research attention due to its usefulness in many applications, which includes selective marketing, decision support, business management, and user profile analysis, to name a few [10], [13], [37]. It has resulted in an unprecedented opportunity to develop automated data-driven techniques of extracting useful knowledge.

Extracting or ‘mining’ knowledge from large amounts of data is referred to as data mining [4]. Data mining should have been more aptly termed as ‘knowledge mining from data’. Mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material. Many terms carrying similar or slightly different meaning to data mining, go by the name knowledge mining from databases, knowledge extraction, data pattern analysis, data archaeology, and data dredging because data mining is about exploration and analysis. By automatic or semiautomatic means, quantities of data can help to uncover meaningful patterns and rules. These patterns and rules help corporations to improve their marketing, sales and customer support operations to better understand their customers.
1.1.1 Historical Perspective

Recently data mining comes in the form of articles in business and software magazines. But in earlier days, people were not familiar with the term data mining. Though data mining is an evolutionary field with a long history, the term itself was only introduced in the 1990s.

The roots of data mining can be traced back along three family lines. The longest of these three is the classical statistics. Without statistics, there would have been no data mining, as these statistics are the foundation of most technologies on which data mining is founded upon. Classical statistics embrace concepts such as regression analysis, standard distribution, standard deviation, standard variance, cluster analysis, and confidence intervals, all of which are used primarily to study data and data relationships. These are the building blocks on which more advanced statistical analysis are built upon. Even in today's data mining tools and knowledge discovery techniques, classical statistical analysis still play a significant role.

The second longest family line for data mining is the Artificial Intelligence (AI). This discipline, which is built upon heuristics, is just the reverse of statistics. It makes an attempt to apply human thought-like processing to statistical problems. This approach requires vast amount of computer processing power. It was not practical until the early 1980s, when computers began to offer useful power at reasonable prices. Artificial intelligence found relatively few applications at the very high-end scientific and government markets and it requires supercomputers, which makes it
available for everyone. The notable exceptions were that some high-end commercial products, such as query optimization modules for Relational Database Management Systems, adopted certain artificial intelligence concepts. Over time this changed, as artificial intelligence was used to create new ways in addressing and solving very complex and math-driven problems. At the Artificial Intelligence Laboratory at MIT, founded in the 1960s, an extensive research in many aspects of intelligence was made. Their aim was two-fold: to understand human intelligence at all levels, including reasoning, perception, language, development, learning, and social levels; and to build useful artifacts based on intelligence.

The third family line of data mining is machine learning, which is more accurately described as the union of statistics and artificial intelligence. While artificial intelligence was not a commercial success, and was therefore primarily used as a research tool, its techniques were largely co-opted by machine learning. Machine learning is able to take advantage of the ever-improving price/performance ratios offered by computers of the 1980s and 1990s, found in more applications because the entry price was lower than artificial intelligence. Machine learning could be considered as an evolution from artificial intelligence because it blends artificial intelligence heuristics with advanced statistical analysis. Machine learning attempts to let computer programs learn about the data studied. Thus the programs make different decisions based on the characteristics of the studied data, using statistics for fundamental concepts, and adding more advanced artificial intelligence heuristics and algorithms to achieve its goals.

As such, data mining, in many ways, is fundamentally the adaptation of machine learning techniques to business applications. Data mining can be well
described as the combination of historical and recent developments in statistics, artificial intelligence, and machine learning. These techniques are used together to study data and find previously hidden trends or patterns within. Data mining is widely accepted in science and business areas where large amounts of data are to be analyzed and to discover new trends that otherwise could not have been found. The three Generations of the Data Mining Systems [72] are as follows:

1\textsuperscript{st} Generation:

The first generation of what is now called data-mining systems appeared in the 1980s and it consisted of research-driven tools focused on single tasks. During this time, there was no proper understanding regarding the multi-dimensional layers of data as one-dimensional analysis tools that could solve these tasks. This Pre-internet era consisted of analyzing single problems in a largely maintained database. This task included building a classifier using a decision-tree or a neural network tool, finding clusters in data, and data visualization. These tools addressed a generic data-analysis problem and their intended user needed to be technically sophisticated to understand and interpret the results. Furthermore, making use of more than one of these tools were very complicated and involved significant data and metadata transformations, which was not an easy task to achieve even for expert users.

2\textsuperscript{nd} Generation:

Data-mining vendors developed the second-generation data-mining systems called suites, which was around 1995. These second generation tools were driven largely by the realization of the knowledge discovery process, which requires multiple types of data analysis, with most of the effort spent in data cleaning and preprocessing. This discovery process generally consisted of discovering patterns in data. The suites, such as SPSS’s (Statistical Package
for Social Sciences) Clementine, Silicon Graphics Mine set, IBM’s Intelligent Miner, made the user to perform several discovery tasks (such as classification, clustering, and visualization) and support data transformation and visualization.

3rd Generation:

The drawback with the 2nd generation systems was that the business users could not use the system directly. Instead, they required significant statistical theory, which was able to support multiple discovery tasks. This implies that in order to extract hidden patterns of information, substantial amount of time was needed to understand what type of algorithm should be used and how it should be applied to generate useful results.

As the result, the third generation came up for business user’s need, vertical data mining based applications and solutions in the 1990s. These tools were primarily driven to solve specific business problems such as predicting future customer purchases or inventory optimization for a specific organization. This knowledge discovery process was done by shifting through piles of information stored in large databases to discover hidden patterns. The end results were pushed to front-end applications such as a decision support system, which allowed the business users to determine the strategy based on specific problem. Ultimately, the data-mining tool was supposed to address the data mining tool specifics.

Data mining vertical applications were developed to provide a high payoff for the correct decisions. These applications often solved the most popular, and at the same time, the most critical business problems for managers. Despite the results generated from applications, it was up to the manager of an organization to understand the data presented to him or her and frames the important and critical business decisions based on this data.
Furthermore, with the change in the Internet and which well-established corporations battle with their competitors, data mining offers new potential to client organizations.

Now these established corporations are equipped with vast amounts of information. These corporations are largely accumulated with several orders for large amounts of data, which are applied to mining tools that helps the helping corporations to determine optimal decisions.

The Internet has infused many new types of data to be at the forefront of the database storage. With increasing numbers, digital libraries are being used to store data such as voice, text, video, and images.

Web data mining or web mining presents a new set of challenges for data mining companies. Analyzing the click stream data found as the customer logs to determine in real time the right advertisement or offering to a particular customer is a new problem faced by data mining companies. Several technologies are becoming the forefront of providing this service to corporations.

Researchers have developed new ways to make predictions for web sites. Collaborative filtering was originally developed at MIT and implemented in systems such as Firefly Network (later acquired by Microsoft in 1998) and Net Perceptions, which are currently used on web sites in order to predict what could be the future customer purchasing patterns. The technique of Collaborative filtering is based on the premise that people who are looking for information should be able to make use of what others have already found and evaluated. As such, these systems make use of information on items purchased or selected to predict what future customers may want to purchase.
The Current collaborative filtering systems provide tools for readers to filter documents based on aggregated ratings over a changing group of readers.

1.1.2 DATA MINING IN KNOWLEDGE DISCOVERY PROCESS

Data preprocessing is commonly used as a preliminary data mining practice. It transforms the data into a format that is more easily and effectively processed for the purpose of the users.

There are a number of data preprocessing techniques such as data cleaning, data integration, data transformation, and data reduction. Each of these techniques is explained: Data cleaning can be applied to remove noise and correct inconsistencies in the data. Data integration merges data from multiple sources into a coherent data store. Data transformation improves the accuracy and efficiency of mining algorithms involving distance measurements. Data reduction can reduce the data size.

These data processing techniques, when applied prior to mining, can substantially improve the overall quality of the patterns mined and/or the time required for the actual mining [45].

Following are the steps involved in knowledge discovery process. Data Mining is the essential step in the process. [79]

1. Data cleaning (to remove noise and inconsistent data)
2. Data integration (where multiple data sources may be combined)
3. Data selection (where data relevant to the analysis task are retrieved from the database)
4. Data transformation (the data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations, for instance)
5. Data mining (an essential process where intelligent methods are applied in order to extract data patterns)

6. Pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interesting measures)

7. Knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user)

The data-mining step may interact with the user or a knowledge base. Interesting patterns being presented to the user may be stored as new knowledge in the knowledge base [79].

![Diagram of data mining process](image)

**Fig 1.1: Data mining process of knowledge discovery**

### 1.1.3 MAJOR COMPONENTS OF DATA MINING

Following are the major components of Data Mining [86]:

**Database, data warehouse, or other information repository:**

This includes one or more sets of databases, data warehouses, spreadsheets, or any other kind of information repositories. Data cleaning and data integration techniques are performed on data.
Database or data warehouse server:
This is responsible for fetching any relevant data, as requested by the user, the database or the data warehouse server.

Knowledge base:

Domain knowledge is used to guide search, or evaluate the resulting patterns. Such knowledge includes concept hierarchies, which are used to organize attributes or attribute values into different levels of abstraction. Knowledge can be used to assess patterns. Other examples of domain knowledge are additional interestingness constraints or thresholds, and Meta data (e.g., describing data from multiple heterogeneous sources).

Data mining engine:

This is essential to a data mining system, which ideally consists of a set of functional modules for tasks such as characterization, association, classification, cluster analysis and evolution and deviation analysis.

Pattern evaluation module:

This component typically employs interestingness measures and interacts with data mining modules focusing on search towards interesting patterns. It may use thresholds to filter out discovered patterns. Alternatively, the pattern evaluation module may be incorporated with the mining module, depending on the implementation of the data mining method used. For efficient data mining, it is highly recommended to push the evaluation of pattern interestingness as deep as possible into the mining process, confining search to interesting patterns only.

Graphical user interface:

This module communicates between the user and the data mining system, allowing the user to interact with the system by specifying a data-
mining query or task and providing information, which helps in, streamline search. Also performing exploratory data mining based on the intermediate data mining results. In addition, this component allows the user to browse database, data warehouse schemes or data structures, evaluate mined patterns, and visualize patterns in different forms.

1.1.4 DATA MINING FUNCTIONALITIES

Data mining functionalities are used to specify the kind of patterns that are found in data mining tasks [79]. Data mining functionalities and the kinds of patterns that can be discovered are described below:

**Concept or Class Description:**

Characterization and Discrimination of data can be associated with classes or concepts. For example, in all electronic stores, classes of items for sales include computers and printers, and the types of customers include big spenders and budget spenders. It would be useful to describe individual classes and concepts in summarized, concise, and precise terms. Such descriptions can be derived by

- Data characterization, by summarizing the data of the class under study, often known in general terms as the target class or

- Data discrimination, by comparing target class with one or more sets of comparative classes, which is often called as the contrasting classes or

- Both data characterization and discrimination.
**Association Analysis:**

Association analysis is the discovery of association rules which shows attribute-value conditions that occur frequently together in given set of data. Association analysis is widely used in market basket data or transaction data analysis.

**Classification:**

Classification is the process of finding set of models (or functions) that describe and distinguish data classes or concepts which enables to use models to predict class of objects, whose class label is unknown. The derived model is based on analysis of a set of training data i.e., data objects whose class label is known.

**Clustering Analysis:**

Clustering analyzes data objects without consulting a known class label. In general, class labels do not find a place in the training data because their beginning is not known. Clustering can be used to generate such labels. Objects are clustered or grouped based on principle of maximizing the intra class similarities and minimizing the interclass similarities.

**Outlier Analysis:**

A database may contain data objects that do not comply with the general behavior or model of the data. These data objects are outliers. Most data mining methods discard outliers as noise or exceptions. However, in some applications such as fraud detection, the uncommon events can be more interesting than more regularly occurring ones. The analysis of outlier data is referred to as outlier mining.
Evolution Analysis:

Data evolution analysis describes and model regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such an analysis include time series data analysis, sequence or periodicity pattern matching, and similarity-based data analysis.

1.2 CLUSTERING ROLE IN DATA MINING PROCESS

Clustering is used to classify homogeneous and well-separated groups of objects in databases [112]. It is a classical problem in database, artificial intelligence, theoretical literature and it also plays an important role in many fields of business and science. However, it is often necessary to interpret the clusters for knowledge discovery in various contexts such as customer profiling, text mining image and video categorization.

Based on the data collected, data mining algorithms are used to either produce a description of the data stored, or predict an outcome. Different kinds of algorithms are used to achieve either one of these tasks. However, in the overall KDD (Knowledge Discovery in Databases) process, any mixture of these tasks may be called upon to achieve the desired results.

The tasks involved in KDD are [39]:

i) **Description tasks:** These tasks describe the data being mined using various methods which are given below.

   a) **Segmentation or Clustering:** To separate data items into subsets that is similar to each other, Partition-based clustering algorithms are used to achieve this task.
b) **Summarization:** To extract compact patterns that describes subsets of data. The method used to achieve this task is Association Rule algorithms.

c) **Change and Deviation Detection:** To detect changes in sequential data (such as protein sequencing, behavioral sequences, etc.).

d) **Dependency Modeling:** To construct models of casualty within the data.

ii) **Prediction tasks:** To predict some field(s) in a database based on information in other fields by using various methods, such as:

a) **Classification:** To predict the most likely state of a categorical variable (its class).

b) **Regression:** To predict results that is numeric continuous variables.

### 1.3 MOTIVATION OF THE THESIS

The analysis and processing of large datasets that arise from simulations and from sensors plays an increasingly important role in many domains of scientific research. Typical examples of very large scientific datasets include long running simulations of time-dependent phenomena that periodically generate snapshots of their state (e.g., hydrodynamics and chemical transport simulation for estimating pollution impact on water bodies, magneto hydrodynamics simulation of planetary magnetospheres, simulation of a flame sweeping through a volume, airplane wake simulations), archives of raw and processed remote sensing data (e.g. Thematic Mapper) and archives of medical images (e.g. high resolution light microscopy, CT imaging, MRI, sonography) [90].
These datasets are usually multi-dimensional. Multidimensional data refers to the increase in dimension of data, which is the essential feature in clustering. The data dimensions can be spatial coordinates, time, or varying experimental conditions such as temperature, velocity or magnetic field. The increasing importance of such datasets has been widely recognized. A number of systems have been developed for supporting storage, retrieval and visualization of such datasets, however, support for efficiently processing these datasets has been lacking. Every application developer has to implement complex support for managing and scheduling the processing. When the dimension of the data increases the complexity also increases correspondingly [70].

Scalability refers to the ability of clustering algorithms to work under increasingly large databases. Since clustering deals with large databases, scalability is a desirable feature. Literally, scalability means that as a system gets larger, its performance goes correspondingly.

While there exists many algorithms for clustering, such methods are suffers in high dimensions. To overcome the drawbacks in existing methods, new approaches are devised to cluster, multidimensional data.

1.4 PROBLEM SPECIFICATION

In statistics and engineering field, to arrange set of vectors (measurements) into a number of groups (clusters), clustering is a well-known problem. Clustering is an important area of application for several fields, which includes data mining, statistical data analysis, vector quantization and many more [20].

Technological advancements have paved the way for evolution of multidimensional data. Existing clustering algorithms face difficulty in handling multidimensional data. The inherent scarcity of the points makes
multidimensional data a challenge for data analysis. The most common problem is the rapid degeneration of performance with increasing dimensions because the approaches are originally designed for low dimensional data.

However the difficulty of high dimensional clustering is primarily due to the following reasons [103]:

- High dimensional data often contains large amount of noise (outliers), resulting in not well-separated clusters, further degrading the effectiveness of the clustering algorithms.
- In high-dimensional spaces, clusters are mostly of various densities.
- Clusters in high dimensional spaces rarely have well defined shapes.
- High dimensional data points, when not well separated can cause common clustering methods in generating inconsistent results.

To conclude, as the dimension of data increases, the problem becomes more complex. Obtaining meaningful result will be time consuming. To solve the high dimension problem, new algorithms are proposed, which can be used for clustering huge multidimensional data.

1.5 OBJECTIVES OF THIS RESEARCH

The primary objective of this research is to devise clustering algorithms for multidimensional data that can be applied to multi disciplinary approach in data mining. While preserving the competitive-clustering performance, the proposed algorithm achieves class accuracy and CPU time, which can be compared to the best of clustering accuracy and CPU time of other algorithms. The specific objective of the study consists of the following algorithms:
• The Scalable Incremental Dimensional Complexity (SIDC) clustering algorithm which can be applied for gene sequence clustering, image clustering and document clustering,
• The probabilistic clustering algorithm for multi dimensional data, which can be applied towards Facility Location problem.

**Partitional Algorithm**

The proposed partitional algorithm is designed with the huge gene sequence databases of high dimension, gathered from Internet and National Center for Biotechnology Information (NCBI) for gene sequence clustering. The database used, gives the scalability of the proposed partitional clustering algorithm and the performance is discussed.

Image clustering using proposed partitional algorithm has been studied with different images and the performances of the system are discussed. In this algorithm, the clustering is discussed to be most promising, making the algorithm suitable for practical applications. Appropriate measures were formulated to evaluate performance of the system.

The proposed partitional algorithm for document clustering has been designed to detect plagiarism in the documents, in which the Joy and Luck method is used as a standard metric to measure the amount of shared information between two computer electronic documents.

**Probabilistic Algorithm**

The Proposed probabilistic clustering technique for multi dimensional data is designed, with the basis of Genetic algorithm to find the centroid and the Euclidean distance metric to find the closest objects to the centroids. The algorithm offers many potential solutions to the multi dimensional data clustering.
The t-test is used to perform the statistical analysis for this thesis. Sir William Gosset gave this test. The t-test is in an application to access whether the mean of a sample drawn from a normal population deviate significantly from a stated value. According to the t-test analysis, if the calculated value modulus $t$ exceeds $t_{0.05}$, then the difference between $X$ and $M$ is significant at 5% level, if it exceeds $t_{0.01}$, the difference is said to be significant at 1% level. If modulus $t$ is less than $t_{0.05}$, it can be concluded that the difference between $X$ and $M$ is not significant and hence the sample might have been drawn from a population with mean=$M$.

1.6 SCOPE OF THE RESEARCH

Clustering is applied in many areas, which includes artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, recommended systems and statistics [2]. In biology, clustering is used to automatically build taxonomy of species based on their features. Currently, there is a considerable interest on the estimation of phylo-genetic trees from gene sequence data. Another application of clustering is to understand better gene functions through biological processes in a cell.

A key step in analysis of gene expression data is detection of groups of genes that manifest similar expression patterns [21]. Another growing application area is the customer relationship management, where data is collected from multiple touch-points (e.g., web surfing, cash register transactions, call center activities), which has become readily available. This data contains valuable knowledge of customer behavior that can help to optimize marketing, bundling, and pricing strategies. Due to the immense size
of data, extracting such knowledge is difficult, and often even obvious insights are overlooked.

Clustering plays a vital role in the mining process because of its ability to summarize data to a manageable level. In this area of research there is a whole lot of scope and need for innovative and efficient data clustering algorithms.

1.7 ORGANIZATION OF THE THESIS

The First chapter discusses the Introduction, Motivation for the present work, objectives of this research and the Scope of the Thesis.

The second chapter discusses fundamentals of clustering, need for clustering, the Review of literature in which the perspective of Clustering and Multidimensional data clustering.

The third chapter explains the widely accepted algorithms like k-means, fuzzy C-means, Hierarchical and Mixture of guassians in detail.

The Fourth chapter discusses the proposed partitional algorithm namely, Scalable Incremental Dimensional Complexity (SIDC) Algorithm in detail. k-means, Fuzzy C-means and Hierarchical methods are implemented in MATLAB (Matrix Laboratory) and all these algorithms are being compared with the proposed partitional algorithm using synthetic data. The performance of proposed partitional algorithm is discussed by taking into consideration of the CPU time and the accuracy.

The Fifth chapter discusses the applications of the SIDC in Gene-Sequence Clustering, Image Clustering, and Document Clustering.

The Sixth chapter discusses the probability based Genetic Algorithm for multidimensional data clustering, which can be trapped by local extrema problem.
The Seventh chapter consolidates findings and results of the entire thesis.

The Eighth chapter concludes the dissertation with further enhancements.

1.8 SUMMARY

Data mining has been discussed in this chapter. Clustering and its role in data mining have also been discussed, highlighting the problems in a detailed manner. The cope and the objectives of the thesis have been explained elaborately. The thesis organization has also been presented.