1. INTRODUCTION

Data is plural word of Datum which have number of meaning in different context raw facts and figures are called data, which can be used as primary input in processing to get a refined and meaningful information Data are interpreted by human to derive a meaningful information.

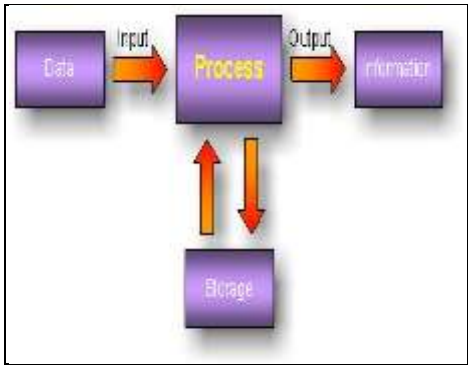


Fig.-1-1**Information**

Mining is the process of digging and searching the valuable minerals from the earth. At different level of mines, we get different minerals and metals and other valuable products.Now data mining refers to the mines of data "Data mining helps end users extract useful bussiness information from large databases". If the databases were small then there is no need of any new technology to discover useful information. Traditionally bussiness organization have small amount of data but nowdays days there are huge amount of data. To keep these data records up to date data warehouses are introduced.

The difference between "Data warehousing" and "Data Mining" can be confusion. The idea that holds them together is that a large data amount of data warwhouse is the "Data mountain" presented to data mining tools. Data warehouse allows you to built that mountain and Data Mining allows you to shift that mountain down to usefull information that is important for our bussiness.

Data mining is the process used to find new, hidden, and unexpected patterns in data to predict the future of the business. It is decision support process that enables an organization to access directly large amount of historical data to predict the future trend of the business
Data mining is the process of extracting previously unknown, valid and actionable information from large data (as transaction data or database or data warehouse) and then using the information so derived to make crucial business and strategic decisions.

It refers to "*science of extracting" or "mining" knowledge from large amounts of data*. It should be noticed that the mining of gold from sand or rocks is termed as *gold* mining instead of rock or sand mining. Therefore, data mining must have been more appropriately named as "knowledge mining from data," which is unfortunately somewhat long. There are many other terms carry a similar meaning of data mining, such as **knowledge extraction, knowledge mining from data, data/pattern analysis, data dredging,** and **data archaeology**. Many people consider data mining as a single for another popularly term used known as Knowledge Discovery from Data.

The data mining involves sorting through large amounts of data and picking out the useful information. The term data mining is used to the two separate processes of knowledge discovery and prediction. Knowledge discovery provides explicit information that has a readable form and can be understood by a user. Predictive modeling provides predictions of future events that might be occur.

It is the process of comparing data from several perspectives and then making the summary of it so that useful information can be obtained that can decrease the cost of the product used in it and increase the total benefits overall. It allows clients to analyze data from several dimensions or degrees, categorize it, and make the summary of the relationships identified. Data mining is the technique of finding correlations or patterns between so many fields in large relational databases. "Data mining finds hidden relationships present in business data to make predictions for future use. Data mining has enhanced as a key business intelligence technology".

Its aim is to extract implicit, previously unknown and potentially useful patterns from data. It consists of many techniques such as classification, clustering, association. The mined

information may be terms of any associations or any relations between the data items in the data. But it is important that such associations or relationships identified by mining process be verified, actionable i.e. Can be causing some positive action, and previously unknown.

1.1 Why Data Mining is Important? Evolution of Data Mining?

Data Mining has focused a huge attention in the information industry and in the information society in the recent years, because of its necessity of large amount of data and the imminent need of converting that data into useful information and knowledge. This useful knowledge and information gained can be used for various applications like fraud detection, market analysis, medical sector, customer relation etc.

Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature:

- Massive data collection
- Powerful multiprocessor computers
- Data mining algorithms

1.1.1 Motivation for Data Mining-

Motivation behind data mining is getting some new innovatives ideas for the improvement of existing bussiness. These days various storage devices are availablein low rates, they can also help to manage and analyze the various data base and warehouses for extraction of usefull patterns, knowledge which can be used for decesion making.

Data mining is viewed as a consequence of the natural evaluation of information technology. It basically extracts or “mines” knowledge from large amount of data.

Following points motivates the industries for using data mining technology-

- Availability of large Databases and Data Warehousing
- Price Drop in Data Storage and Efficient Computer Processing
- New Advancements in Analytical Methodology
- Benefits of Data Mining

1.1.2 Steps in the Evolution of Data Mining-

Evolutionary Step	Business Question	Enabling Technologies	Product Providers	Characteristics
Data Collection (1960s)	"What was my total revenue in the last five years?"	Computers, tapes, disks	IBM, CDC	Retrospective Static data delivery
Data Access (1980s)	"What Were unit sales in New England last March?"	Relational databases (RDBMS), structured Query Languge (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in England last March? Drill down to Boson."	On-line analytic processing (OLAP), multidimensional data bases, data warehouses	Pilot, Comshare, Arbor, Cognos, Microstrategy	Retrospective, dynamic data delivery at multiple levels
Data Mining (Emerging Today)	"What's likely to happen to Boston unit sales next month? Why?"	Advanced algorithms, multiprocessor computers, massive database	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Table- 1-1 **Evolution of Data Mining**

1.2 What is Data Mining?

Data Mining is the process of analyzing and exploring that data to discover patterns and trends. In easy words it is referred as an extracting or "mining" knowledge from the large amount of data. It refers to the process of analyzing large amount of data from different sources and summarizing it into useful information. More appropriately it can be referring as "knowledge mining from data" and in short "knowledge mining".

"The process of extracting valid, previously unknown, comprehensible, and actionable information from large databases and using it to make crucial business decisions"

1.2.1 Data mining by definition-

- Data mining is the process used to find new, hidden, or unexpected pattern in data to predict the future of the business. It is a decision support process that enables an organization to access directly large amount of historical data to predict the future trend of the business.

- Data mining is used to describe those applications of either statistical analysis or data discovery products which analyze large populations of unknown data, to identify hidden patterns that might be useful.

- Data mining is that it automates the detection of relevant patterns in a database. It uses well established statistical and machine learning techniques to build models that predict customer behavior.

- Data mining is a collection of techniques that aim to find useful but undiscovered patterns in collected data. Its goal is to create models for decision-making that predict future behavior based on analysis of past activity.

- It is a process of extracting previously unknown, valid and actionable information from large data and then using that information so derived to make crucial business and strategic decisions.

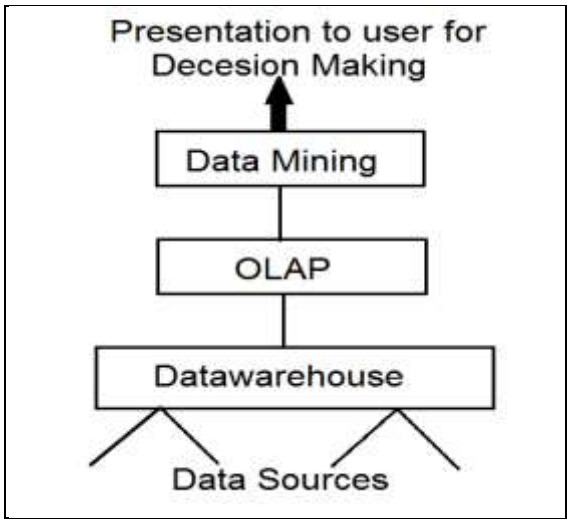


Fig.1-2**the Data Mining Hierarchy in business, decision supports**

1.2.2 Steps of Data Mining-

- Data cleaning: To remove noise and irrelevant data.
- Data integration: It is the place where several data source gathered.
- Data selection: Where data which is relevant for the analysis are retrained from the database and the selected.
- Data transformation: Where the data is consolidated or transformed according to the demand for mining by doing summary or aggregation operation.
- Data mining: An important essential process where intelligent methods are applied for the information extracting from relevant data patterns.
- Pattern evaluation: To identify the truly used interesting patterns representing knowledge based else some required measures.

- Knowledge presentation: Where knowledge and visualization representation techniques are applied to present the mined knowledge to the users.

The data mining steps might interact with the user or a client. The interesting patterns are built to the user and may be stores as a new knowledge in the knowledge base system. It is a knowledge discovery process. In simple words a modified view of data mining functionality: data mining is a process of discovering interesting meaningful knowledge from various large amounts of data stored either in databases, data warehouses or other information systems.

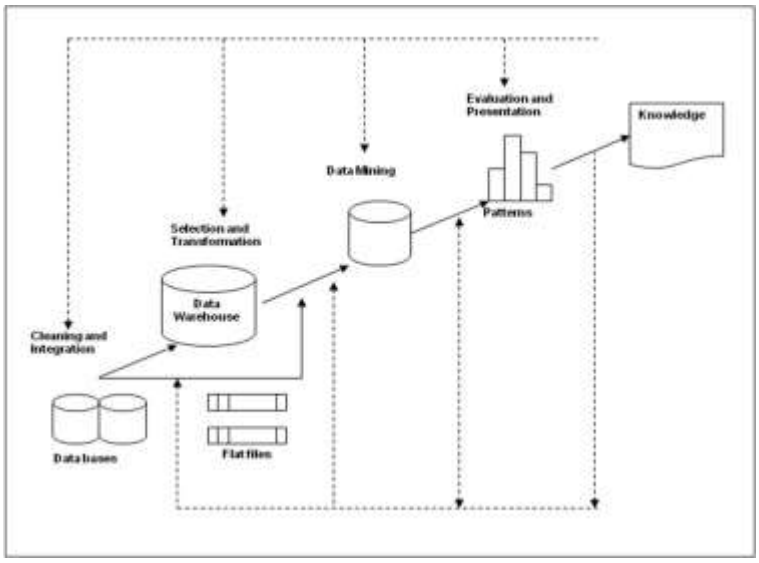


Fig.1-3 **Data mining as a step in the process of knowledge discovery**

Based on the data mining the functionality the architecture of a typical data mining system have six major components which is depicted infig.1-4

- **Database, world wide web, data warehouse, or other information repository:**
  These are the sets of database or information sets available. Data cleaning and data integration techniques can be applied.

- **Database or data warehouse server:**

It basically fetches relevant data which is based on the user's data mining request.

- **Knowledge base:**
  It helps in searching interesting patterns. It uses technique of concept hierarchy to organize data at different levels of abstraction.

- **Data mining engine:**
  It consists of a set of interesting modules for tasks.

- **Pattern evaluation module:**
  It interacts with data mining modules so as to focus towards interesting patterns.

- **User Interface:**
  This component communicates between the users and the data mining system, and allows the user to interact and communicate with the system by mentioning a data mining task or query.

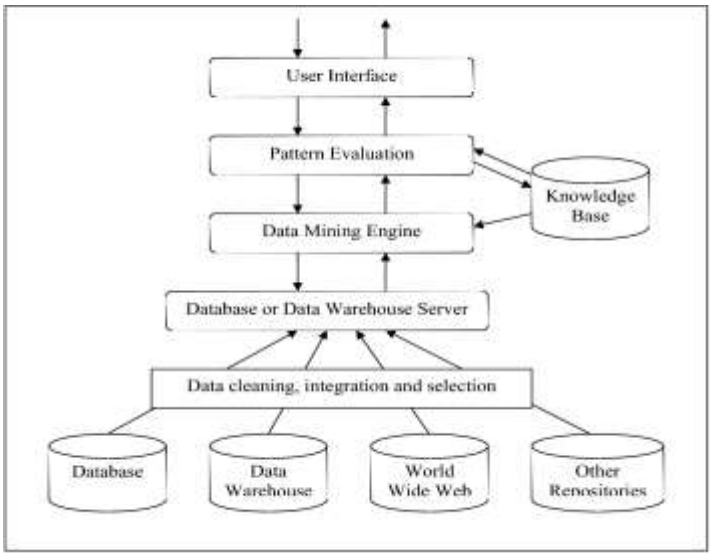


Fig.1-4 **Architecture of a typical Data mining system**

Step 1 Deciding business objectives for data mining

This is the very first step of data mining method. Business objectives should be decided before data mining process. This can be only done by the joint effort of the business analyst with domain knowledge and the data analyst who can translate the objective identified by the analyst into a precise problem.

Step 2 Data Preparation

After deciding the objectives the data is prepared for the mining process. It is a process that uses a computer application to input data and analyze and convert it into usable information. It involves recording, sorting, analyzing, calculating and summarizing. Data is most helpful when its well presented and informative, data processing system can be also referred as information system. Information system takes raw data as an input and produce information as an output.
Raw data is not ready for input in data warehouse because at the time of recording process of data there are some unwanted data or error in recording process that add noise in data. So we have to filter data and remove unwanted signals from them. Raw data may includes noise, missing and inconsistent data due to their huge size. So it has to be processed in order to improve the quality of the data and even the mining results.

Data preprocessing includes the following process

- Data cleaning
- Data integration
- Data transformation
- Data reduction

These techniques can improve the overall mining results.

Step 3 Data Mining

In this we have to apply various data mining algorithms to the data which is already preprocessed. It involves the actual mining process by the analyst who is available with pre-processed data.

Step 4 Analysis of Result

This is the most important step of data mining process. Analysis of data can be done by a data analyst with the help of business analyst using visualization aids and tools. Association rules should be developed properly in data mining.

Step 5 Assimilation of knowledge

It is the last step of data mining process**.** It includes steps to implement the knowledge gained by data mining.

Data Mining Architecture

The architecture of a complicated data mining system should have the following major components (Figure 1-4):

- **Database, World Wide Web, data warehouse, or other information repository:** This can be one or a set of databases, available information data sets, data warehouses, spreadsheets, or any other kinds of information systems. The data can be performed using Data cleaning and data integration techniques.

- **Database or data warehouse server**: It is responsible for catching the relevant data, based on the client data mining request.

1.2.3 Different types Of Data Mining-

- Business Data Mining
- Scientific Data Mining
- Internet Data Mining

1.2.4 Data Mining consists of Five major Elements-

- Extract, transformed, and load transaction data on the data warehouse system.
- Stores and manages the data in a multidimensional database system of data mining.
- Provide data accessibility to business tycoons and information technology professionals.
- Analyze the data by application software which is well suited for it.
- Present the data in a useful manner, such as a graph or tabular.

1.3 Data mining is applied on various kinds of data -

Data mining should be applied on any type of information system. This includes relational databases, data warehouses, advanced database systems, transitional databases, flat files and the World Wide Web. The techniques and challenges of mining may differ for each of the repository system.

1.3.1 Relational Database-

A database system, also called a database management system (DBMS), it is consisted of a collection of interrelated data, known as a database, and it is a set of software programs required to manage and access the data system.

Relational database is a combination of tables, in each of which relational database is assigned a unique name or Id. Every table represents a set of attributes and usually stores a large number of tupples. Each table is represented by a unique key by an object identified and mentioned by a set of attribute values.

It can be run by visualizing database queries written in relational query language, such as SQL or with help of GUI. A query is transformed first into a relational operation, such as join, selection and projection and then is optimized for efficient processing.

1.3.2 Data Warehouse-

It is a repository of information collected from various sources stored under single scheme and basically resides at a single site. Data warehouses are always constructed of a periodic data patterns.

Data warehouse are organized into major subjects. It is usually a multidimensional structure, where every dimension relates to an attribute and all stores the value of some aggregate data such as sales or count. The physical structure of data warehouse might be relational data store or multidimensional view of data and allows pre computation and fast accessing of data.

Data warehouse systems are always suited for On-Line Analytical Processing or OLAP. Its tools helps to support data analysis, additional tools are needed for data mining to allow extraction in depth and automated analysis.

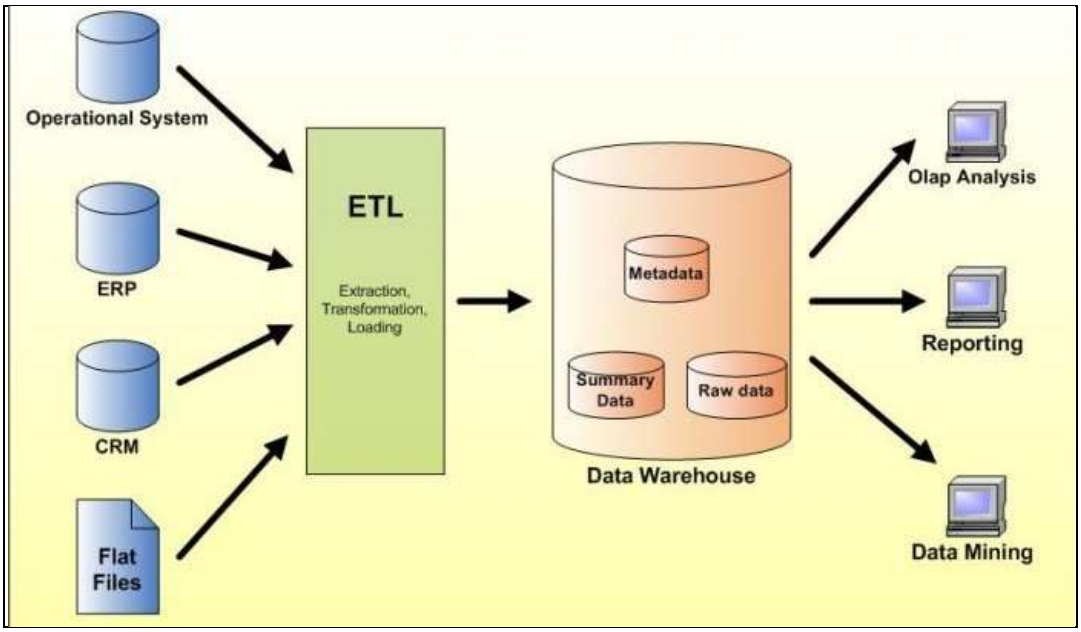


Fig.1-5 **Data Warehouse**

1.3.3 Transactional Database-

Transactional database consists of a file where the record in kept in a tabular form and each record represents a transaction. A transaction database consistsof a unique transaction identity number and a list of the relevant items that makes up the transaction.

The transactional database have additional tables associated with it, which contain other information regarding the sale, such as the date of transaction, the customer ID number, the ID number of the sales person and the branch where the sale occurred and loan.

1.3.4 Advanced Data and Advanced Applications and Information Systems -

Relational database is systems used in business applications. The new database application includes hypertext and multimedia data, time related data and the world wide data. These applications require efficient data structure, variable length reports, semi structured or unstructured data, text and multimedia data and database schemes with complex structures and dynamic changes.

While such database requires complex facilities to efficiently store retrieve and update large amount of complex data, they also provide smooth grounds and raise many challenging research issues for data mining.

Each of the advanced databases is listed below**:**

- Object- Relational Databases
- Temporal Databases, Sequence Databases, and Time-Series Databases
- Spatial Databases
- Text Databases and Multimedia Databases
- Heterogeneous Databases and Legacy Databases
- Data Streams
- World Wide Web

1.4 Data Mining Functionalities-

There are various kinds of patterns that can be mined and various types of databases and information repositories on which data mining can be performed.

The data mining can be broadly classified into two categories:

- Descriptive
- Predictive

Descriptive mining tasks approach the general properties of the data in database.
Predictive mining tasks approaches interfere on the current for making predictions.

In several cases, the users have no idea regarding what kinds of patterns they should use to make their data interesting, and hence they may search for various different kinds of patterns in parallel. Therefore it is needed to have a data mining system that can runs various multiple kinds of patterns to accomplish different user expectations or applications.
Data mining functionalities and there patterns, are described below:-

1.4.1 Concept or Class description: Characterization and Discrimination

Data can be associated with classes or concepts. Therefore we can describe data to individual classes and concepts in brief, are known called class concept descriptions. These descriptions can be derived vice-

- Data characterization, by briefing the data of the class
- Data discrimination, by comparing the target class
- Both data characterization and discrimination.

1.4.2 Association Analysis-

It is a set of rules showing attribute value conditions that occur simultaneously in a given set of data. Widely used for transaction data analysis ormarket basket. Association rules that contain to a single domain are known as single dimensional association rule and those contain more than one domain are known as multidimensional association rule.

1.4.3 Classification and Prediction-

It's the process of finding a set of models which distinguish ordescribe the data classes or concepts, for such being able to use the model for predicting the class of an object where the class label kept is unknown. The derived model is based on data object where the class object is known.
The derived model may be represented in various forms such as (IF-THEN) rules, decision tree, mathematical formulation, neural network.

Classification is used for prediction the class label of data objects. Although, prediction is distinct from classification prediction, emphasis on identification of distribution trends based on available data.

1.4.4 Clustering Analysis-

It studies data objects without consulting a known class label. In general, there is no class labels present in the sequential format because they are not known to being with. Such labels are generated via clustering. The objects are clustered based on the principle of maximizing the interclass similarity. Each cluster that is formed through this process may be viewed as a class of those particular objects, from which rules can be derived.

1.4.5 Outlier Analysis-

The data objects, which are different from the remaining set of data are simply known as outliers. They are caused by measurement or execution error.

Outlier mining has wide application:

- Fraud detection
- Useful for identifying the behavior of customers
- Medical analysis

1.4.6 Evolution and Deviation Analysis-

Data evolution analysis describes models functionalities or trends for those objects whose behavior changes during a certain time. It include characterization, discrimination, association, classification, clustering over time related data, various features of such analysis include time series data analysis, sequence or priority pattern matching and similarity based data analysis.

1.5 Data Mining Systems Classification-

Data mining is a process where data is interconnected, including database systems, machine learning, statistic, visualization, and information science. It also uses techniques from other fields, such as neural network, knowledge representation, and high- performance of computing. Depending upon the kinds of data that are to be mined, the data mining system uses the integrate techniques from spatial data analysis, pattern recognition,image analysis, information retrieval, signal processing, bioinformatics, web technology, economics, psychology, computer graphics, or business.

Because of the diversity of fields contributing to data mining, its research is to generate a large variety of data mining systems from the used sources. Data mining systems can be categorized into various types:

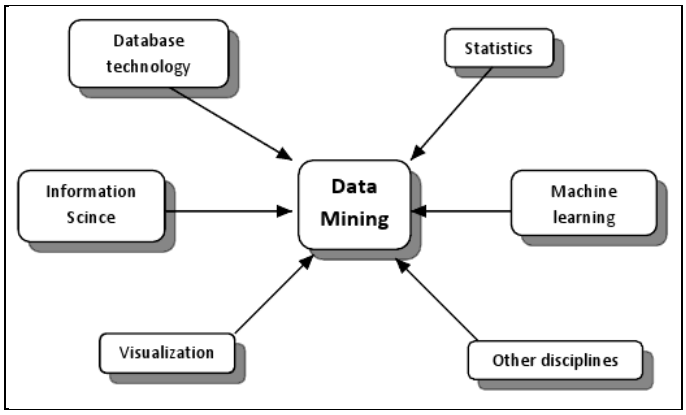


Fig.1-6 **Data mining as a confluence of multiple disciplines**

- Classification according to the Kind of databases mined.
- Classification according to the Kind of Knowledge mined.
- Classification according to the Kind of Techniques utilized.
- Classification according to the Applications Adapted.

1.6 Clustering Analysis-

Clustering analysis, also called “segmentation analysis” or “taxonomy analysis” aims to identify homogeneous objects into a set of groups, named clusters, by given criteria. Clustering is a very important technique of knowledge discovery for human beings. It has a long history and can be traced back to the times of Aristotle. These days, cluster analysis is mainly conducted on computers to deal with very large-scale and complex datasets. With the development of computer-based techniques, clustering has been widely used in data mining, ranging from web mining, image processing, machine learning, artificial intelligence, pattern recognition, social network analysis, bioinformatics, geography, geology, biology, psychology, sociology, customers behavior analysis, marketing to e-business and other fields.

Cluster analysis includes two major aspects: **clustering** and **cluster validation**. Clustering achieves to distinguish objects into groups according to certain criteria. The grouped objects are called clusters, where the similarity of objects is high within clusters and low between clusters. To achieve different application purposes, there are several number of clustering algorithms have

been developed. However, there are no general-purpose clustering algorithms that fit all kinds of applications, thus, the quality of clustering results plays the critical role of cluster analysis, i.e., cluster validation, which aims to assess the quality of clustering results and find a fit cluster scheme for a specific application.

However, in practice, it may not always be possible to cluster huge datasets by using clustering algorithms successfully, due to whichit weakens several commonly existing automated clustering algorithms are dealing with arbitrarily shaped data distribution of the datasets. As Abul et al pointed out “In high dimensional space, traditional clustering algorithms tend to break down in terms of efficiency as well as accuracy because data do not cluster well anymore”. In addition, the very high computational cost of statistics-based cluster validation methods directly impacts on the efficiency of cluster validation. The clustering of large sized datasets in data mining is an iterative process involving humans. Thus, the user’s initial estimation of the cluster number is important for choosing the parameters of clustering algorithms for the pre-processing stage of clustering. Also, the user’s clear understanding on cluster distribution is helpful for assessing the quality of clustering results in the post-processing of clustering. All these heavily rely on the user’s visual perception of data distribution. Clearly, visualization is a crucial aspect of cluster exploration and verification in cluster analysis. Visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps in data. Therefore, introducing visualization techniques to explore and understand high dimensional datasets is becoming an efficient way to combine human intelligence with the immense brute force computation power available nowadays.

Visualization used in cluster analysis maps the high-dimensional data to a 2D or 3D space and aids users having an intuitive and easily understood graph/image to reveal the grouping relationship among the data. As an indispensable revealing technique, visualization is almost involved into every step in data mining. Visual cluster analysis is a combination of visualization and cluster analysis. The data sets that clustering algorithms deals with, are normally in high dimensions (>3D). Thus, choosing a fit technique to visualize clusters of high dimensional data is the first task of visual cluster analysis. There have been many works on multidimensional data visualization, but those earlier techniques of multidimensional data visualization are not suitable

to visualize cluster structures in very high dimensional and very large datasets. With the increasing applications of clustering in data mining, in the last decade, more and more visualization techniques have been developed to study the structure of datasets in the applications of cluster analysis.

Several approaches have been proposed for visual cluster analysis but their arbitrary exploration of group information makes them inefficient and time consuming in the cluster exploration stage. On the other hand, the impreciseness of visualization limits its utilization in quantitative verification and validation of clustering results. Thus developing a visualization technique, with the features of purposeful cluster detection and precise contrast between clustering results is the motivation of this research.

To mitigate the above-mentioned problems, based on hypothesis testing, we propose a visual projection technique called **Hypothesis Oriented Verification and Validation by Visualization**, **(HOV3)**. HOV3 generalizes random adjustments of Star Coordinates based techniques as measure vectors. Thus, compared with the Star Coordinates based techniques, HOV3 has several superiorities.

- First, data miners can summarize their prior knowledge of the studied data as measure vectors, i.e., hypotheses of the data. Base on hypothesis testing, data miners can quantitatively analyze data distribution projected by HOV3 with hypothesis.

- Second, HOV3 avoids the arbitrariness and randomness of most existing visual techniques on cluster exploration, for example Star Coordinates and its implementations, such as VISTA/ iVIBRATE. As a consequence, HOV3 provides data miners a purposeful and effective visual method on cluster analysis.

Based on the quantified measurement feature of HOV3, we propose a “visual external cluster validation” model to verify the consistency of cluster structures. Compared with statistics based external cluster validation methods, we show that Hypothesis Oriented Verification and Validation by Visualization, (HOV3) based external cluster validation model is more intuitive

and effective. We also introduce a visual approach called M-HOV3/M Mapping to enhance the visual separation of clusters. With the above features of HOV3, a prediction-based visual approach is proposed to explore and verify clusters.

Clustering is an important technique that has been successfully used in data mining. The goal of clustering is to distinguish objects into groups (clusters) based on given criteria. In data mining, the datasets used in clustering are normally huge and are in high dimensions. Nowadays, clustering process is mainly performed by computers with automated clustering algorithms. However, those algorithms favor clustering spherical or regular shaped datasets, but are not very effective to deal with arbitrarily shaped clusters. This is because they are based on the assumption that datasets have a regular cluster distribution.

Visual Data Mining is mainly a combination of information visualization and data mining. In the data mining process, visualization can provide data miners with intuitive feedback on data analysis and support decision-making activities. In addition, visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps in data. Many visualization techniques have been employed to study the structure of datasets in the applications of cluster analysis. However, in practice, those visualization techniques take the problem of cluster visualization simply as a layout problem. Several visualization techniques have been developed for cluster discovery but they are more exploration oriented, i.e., stochastic and subjective in the cluster discovery process.

In this paper, we propose a novel approach, named HOV3, Hypothesis Oriented Verification and Validation by Visualization, which projects the data distribution based on given hypotheses by visualization in 2D space. Our approach adopts the user hypotheses (quantitative domain knowledge) as measures in the cluster discovery process to reveal the gaps of data distribution to the measures. It is more object/goal oriented and measurable.

2. LITREATURE REVIEW

CLUSTER ANALYSIS-

Cluster analysis is a research investigative process. It can be used to discover structures in data without any problem of interpretation/ explanation [JaD88]. Cluster analysis includes two major aspects: clustering and cluster validation. Clustering aims at dividing or separating objects into groups according to a certain criteria. To achieve different application purposes, various clustering algorithms have been developed[JaD88, KaR90, JMF99, Ber06]. While, there is not a single clustering algorithm that fits all kinds of applications, thus, it's required an evaluation process to analyze the quality of clustering result that produced by different algorithms or parameters, so that the client may find the best suitable cluster scheme for its specific requirement. The quality analysis process of clustering results is termed as cluster validation. Cluster analysis is a recurrent process of clustering and cluster verification by the user facilitated with clustering algorithms, cluster validation methods, visualization and domain knowledge to databases.

In this chapter, we give a review of cluster analysis as the background of the thesis. First we introduce clustering, clustering algorithms and their features, and there drawbacks. This is further followed by the introduction of cluster validation, exists the cluster validation methods, and the problems with the existing cluster validation approaches.

2.1 Clustering and Clustering Algorithms

Clustering is considered as an unsupervised classification process[JMF99]. The clustering problem is partition a dataset into groups (clusters) so that the data elements within a cluster are more similar to each other than data elements in different clusters by given criteria. Several various numbers of clustering algorithms have been developed to meet different requirements [JaD88, KaR90, XuW05, and Ber06]. Based on the tactics of how data objects are distinguished, clustering techniques can be divided into two classes: hierarchical clustering techniques and partitioning clustering techniques [Ber02]. However there is no clear boundary between these

two classes. Some efforts have been done on the combination of different clustering methods for dealing with specific applications. Beyond the two hierarchical and partitioning classes, there are several clustering techniques that are categorized into independent classes, for example, density-based methods, Model-based, and Grid-based methods clustering methods [Hak01, Ber06, and Pei]. A short description of these methods is described below.

2.1.1 Partitioning methods-

Partitioning clustering algorithms, such as K-means [Mac67], K-medoids PAM[KaR87], CLARA [KaR90] and CLARANS [NgH94] assign objects into K (predefined cluster number) clusters, and recurrence reallocate objects to improve the quality of clustering results.

K-means is the most easy and popular clustering algorithm [Mac67]. The main idea of K-means is as follows:

- Arbitrarily choose K objects to be the initial cluster centers/centroids;
- Assign each object to the cluster associated with the closest centroid;
- By the mean value of the objects compute the new position of each centroidin a cluster;
- Repeat Steps 2 and 3 until the means are fixed.

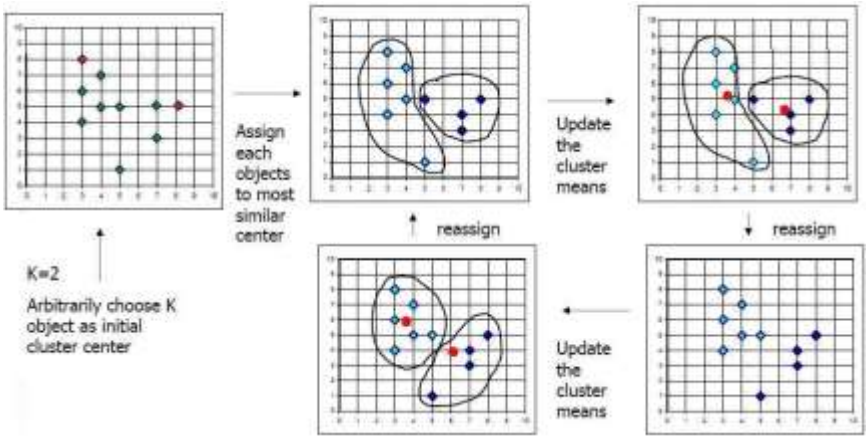


Figure 2-1 an example of clustering procures of K-means [HaK01]

Therefore, K-mean algorithm is highly sensitive in the selection of the initial centroids, else the different centroids produce different cluster results. Main drawback of K-means is that, there is no theoretical solution to find the valid the valid numbers of clusters for a data set.
In this a simple solution would help in comparing the results of multiple runs with different K numbers and choose the best one according to that criteria, an exception is there that when the data is huge or large, it takes a long time in other words it is very time consuming to have multiple runs of K-means and then comparing the clusters after each run.

A variation of K-mean, K-mediods calculates the medoid of the object in each cluster instead of using the mean value of data objects as the center of cluster. The K-mediod process is also similar to K-mean process. K-mediod clustering algorithm is very sensitive to outliers. Therefore, outliers could seriously affect the clustering results.

To solve this problem some techniques have been made on K-mediods algorithm. For eg: PAM (Partitioning Around Mediods) was proposed by Kaufman and Rousseeuw [KaR87]. PAM inherits the features of K-mediods clustering algorithm and for better PAM equips a medoids swap mechanism. PAM is more robust than k-means in terms of handling noise and outliers, since the medoids in PAM are less influenced by outliers. With the O(k(n-k)2) computational cost for each iteration of swap (where k is the cluster number, n is the items of the data set), it is clear that PAM only performs well on small-sized datasets, but does not scale well to large datasets.

In practice, PAM is embedded in the statistical analysis systems, such as SAS, R, S+ and etc. to deal with the applications of large sized datasets, i.e., CLARA (Clustering LARge Applications) [KaR90]. By applying PAM to multiple sampled subsets of a dataset, for each sample, CLARA can produce the better clustering results than PAM in larger data sets. But the efficiency of CLARA depends on the sample size. On the other hand, a local optimum clustering of samples may not the global optimum of the whole data set.

Ng and Han [NgH94] abstracts the mediods searching in PAM or CLARA as searching k sub graphs from n points graph, and based on this understanding, they propose a PAM- like

clustering algorithm called CLARANS (Clustering Large Applications based upon RANdomized Search). While PAM searches the whole graph and CLARA searches some random sub-graphs, CLARANS randomly samples a set and selects k medoids in climbing Sub-graph Mountains. CLARANS selects the neighboring objects of medoids as candidates of new medoids. It samples subsets to verify medoids in multiple times to avoid bad samples. Obviously, multiple times sampling of medoids verification is time consuming. This limits CLARANS from clustering very large datasets in an acceptable time period.

2.1.2 Hierarchical methods-

Hierarchical clustering algorithms assign objects in tree-structured clusters, i.e., a cluster can have data points or representatives of low level clusters [HaK01]. Hierarchical clustering algorithms can be classified into categories according their clustering process: agglomerative and divisive. The process of agglomerative and divisive clustering is exhibited in Figure 2-2.

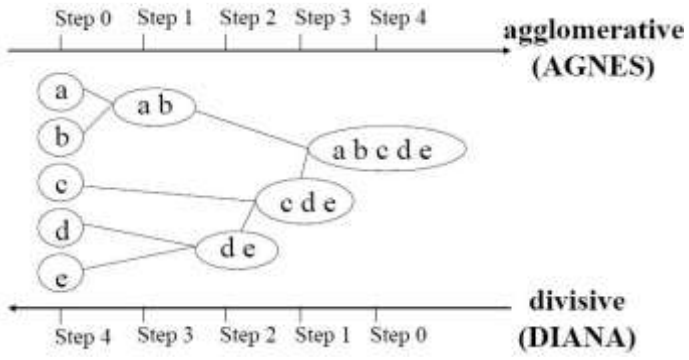


Figure 2-2 Hierarchical Clustering Process [HaK01]

- Agglomerative: one starts with each of the units in a separate cluster and ends up with a single cluster that contains all units.
- Divisive: to start with a single cluster of all units and then form new clusters by dividing those that had been determined at previous stages until one ends up with clusters containing individual units.

AGNES (Agglomerative Nesting) adopts agglomerative strategy to merge clusters [KaR90].AGNET arranges each object as a cluster at the beginning, then merges them as upper level clusters by given agglomerative criteria step-by-step until all objects form a cluster, as shown in Figure 2-2. The similarity between the two clusters is measured by the similarity function of the closest pair of data points in the two clusters, i.e., single link. DIANA (Divisive ANAlysis)
adopts an opposite merging strategy, it initially puts all objects in one cluster, then splits them into several level clusters until each cluster contains only one object [KaR90].

The merging/splitting decisions are critical in AGNES and DIANA. On the other hand, with O(n2) computational cost, their application is not scalable to very large datasets.

Zhang et al [ZRL96] proposed an effective hierarchical clustering method to deal with the above problems, BIRCH (Balanced and Iterative Reducing and Clustering using Hierarchies).
BIRCH summarizes an entire dataset into a CF-tree and then runs a hierarchical clustering algorithm on a multi-level compression technique, CF-tree, to get the clustering result. Its linear scalability is good at clustering with a single scan and its quality can be further improved by a few additional scans. It is an efficient clustering method on arbitrarily shaped clusters. But BIRCH is sensitive to the input order of data objects, and can also only deal with numeric data. This limits its stability of clustering and scalability in real world applications.

CURE uses a set of representative points to describe the boundary of a cluster in its hierarchical algorithm [GRS98]. But with the increase of the complexity of cluster shapes, the number of representative points increases dramatically in order to maintain the precision.

CHAMELEON [KHK99] employs a multilevel graph partitioning algorithm on the k-Nearest Neighbor graph, which may produce better results than CURE on complex cluster shapes for spatial datasets. But the high complexity of the algorithm prevents its application on higher dimensional datasets.

2.1.3 Density-based methods-

The primary idea of density-based methods is that for every point of a cluster the neighboring contains a given unit distance with at least a minimum number of points, i.e. the weight in the neighborhood should reach some threshold [EKS+96]. However, this idea is based on the assumption of that the clusters are in the spherical or regular shapes. DBSCAN (Density-Based Spatial Clustering of Applications with Noise) was proposed to adopt density reachability and density-connectivity for handling the arbitrarily shaped clusters and noise [EKS+96]. But DBSCAN is very sensitive to the parameter Eps (unit distance or radius) and Min Pts (threshold density), because before doing cluster exploration, the user is expected to estimate Eps and Min Pts. DENCLUE (DENsity-based CLUstEring) is a distribution-based algorithm [HiK98], which performs well on clustering large datasets with high noise. Also, it is significantly faster than existing density-based algorithms, but DENCLUE needs a large number of parameters. OPTICS is good at investigating the arbitrarily shaped clusters, but its non-linear complexity often makes it only applicable to small or medium datasets [ABK+99].

2.1.4 Grid-based methods-

The idea of grid-based clustering methods is based on the clustering-oriented query answering in multilevel grid structures. The upper level stores the summary of the information of its next level, thus the grids make cells between the connected levels, as illustrated in Figure 2-3.

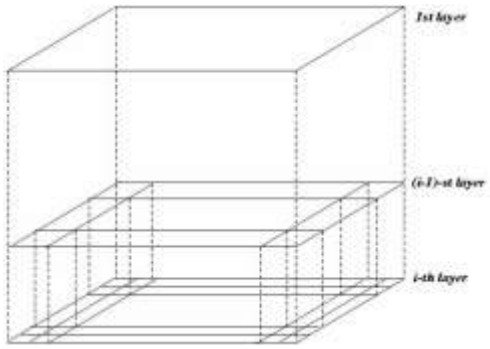


Figure 2-3 The grid-cell structure of gird-based clustering methods

Many grid-based methods have been proposed, such as STING (Statistical Information Grid Approach) [WYM97], CLIQUE [AGG+98], and the combination of grid-density based technique Wave Cluster [SCZ98]. The grid-based methods are efficient on clustering data with the complexity of O(N). However the primary issue of grid-based techniques is how to decide the size of grids. This quite depends on the user's experience.

2.1.5 Model-based clustering methods-

Model-based clustering methods are based on the assumption that data are generated by a mixture of underlying probability distributions, and they optimize the fit between the data and some mathematical model, for example statistical approach, neural network approach and other AI approaches. The typical techniques in this category are Auto clas [CKS+88], DENCLUE [HiK98] and COBWEB [Fis87].When facing an unknown data distribution, choosing a suitable one from the model based candidates is still a major challenge. On the other hand, clustering based on probability suffers from high computational cost, especially when the scale of data is very large.

Based on the above review, we can conclude that, the application of clustering algorithms to detect grouping information in real world applications in data mining is still a challenge, primarily due to the inefficiency of most existing clustering algorithms on coping with arbitrarily shaped distribution of data of extremely large and high-dimensional datasets.

Extensive survey papers on clustering techniques can be found in the literature [JaD88, KaR90, JMF99, EML01, XuW05, Ber06].

2.2 Cluster Validation-

There are several clustering algorithms to deal with applications. But the main problem arises which clustering algorithm is best suitable for specific application? How many clusters arise in

the data? Is there any better cluster scheme then this? These questions are related with the quality of results. i.e clustering validation.
Clustering validation is a procedure of improving the quality of clustering results and then finding the best application algorithm for the specific clusters. It aims in finding the best suited solution.

Cluster validation is an essential process of cluster analysis, because only this method can gives the simple clusters from main datasets then those different clustering algorithms usually super impose several cluster structures on a data set even if there is no cluster structure present in it [Gor98] [Mil96]. Cluster validation is required in data mining to solve different problems [HCN01]:

1. To measure all the generated partitions of a real data set generated by a clustering algorithm.
2. To find the well suited clusters for data mining from the partition.
3. To distinguish between the clusters.

2.2.1 Internal criteria-

It is a method of evaluating and comparing the quality of clusters when the records are generated to concentrate the quality of the processed clusters using the available data objects only [VSA05]. It excludes any kind information beyond all the clustering data, and only attention on assessing clusters.

2.2.2 Relative criteria-

It compares two structures and then measures their relative merit. Its aim is to run and optimize the clustering algorithm for a several number of times and identify the best clustering scheme that will suit it [ALA+03],i.e., later they works on the clustering results by applying an algorithm with several different parameters on a data set and finding the feasible optimal solution. Relative cluster validity is also known as cluster stability. [KeC00, LeD01, BEG02, RBL+02, BeG03].

2.2.3 External criteria-

The results are based on a pre-specified structure of a clustering algorithm, which gives a brief idea to the user about the clustering structure. [HKK05].
External cluster validation is a hypothesis procedure test in which several given sets of class labels are produced by a cluster scheme, and then later it is compared with all the clustering results by applying the same cluster scheme to the parts of a database, as shown in the Figure 2-4.

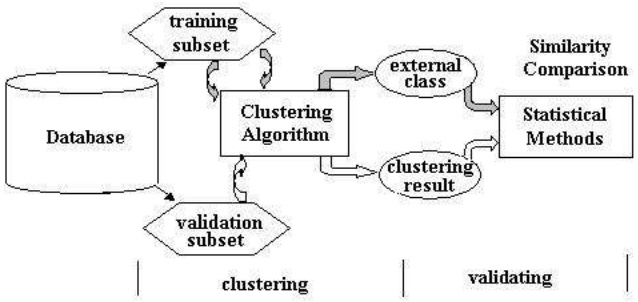


Figure 2-4 External criteria based validation [ZOZ07a]

It is based on the assumption that an output of the clustering algorithm obtained by finding a similarity or resemblance of the clusters with existing classes [Dom01], [KDN+96], [Ran71], [FoM83], [MSS83].

2.3 The issues of cluster analysis-

Now we can say it clearly that there are two major drawbacks or problems that influence the accuracy of cluster analysis in data mining.

- The first is the inferiority of mainly all clustering algorithms that are dealingwith arbitrarily shaped elements of the datasets.

- The second issue is mentioned as, the quality of clustering results is time consuming when the statistics-based methods is used and when the database is large, and its cost is very high. [HBV02].

Moreover, due to the inefficiency of clustering algorithms on handling the arbitrarily shaped clusters in extremely large datasets, the data mining directly affects the effect of cluster validation, because it is based on the clustering results.
In addition, most of the existing clustering algorithms deal with the whole clustering algorithm automatically, i.e., once the user by default sets all the parameters of algorithms, the clustering result generated after the whole process, which excludes the user until the end. So, it is very hard to handle user domain knowledge into the clustering process.

Visualization techniques have proven to be of high value in exploratory data analysis and data mining [Shn01]. Therefore, the introduction of domain experts' knowledge supported by visualization techniques is a good remedy to solve those problems. A detailed review of visualization techniques used in cluster analysis is presented in the next chapter.

3. METHOLOGY

VISUAL CLUSTER ANALYSIS-

As described in the last chapter, most of the existing automated clustering algorithms suffer in terms of efficiency and effectiveness on dealing with arbitrarily shaped cluster distributions of extremely large and multidimensional datasets [1]. Baumgartner et al [BPR+04] concluded that "In high dimensional space, traditional clustering algorithms tend to breakdown in terms of efficiency as well as accuracy because data do not cluster well anymore". Another obstacle to the application of cluster analysis in data mining is that, the high computational cost of statistics-based cluster validation methods [HCN01]. These drawbacks limit the usability of clustering algorithms in real-world data mining applications. To mitigate the above problems, visualization has been introduced into cluster analysis. As Card et al [CMS99] described, visualization is the use of computer-supported interactive, and visual representation of abstract data to amplify cognition. Visualization is considered as one of the most intuitive methods for cluster detection and validation, especially performing well on the representation of irregularly shaped clusters. Visual data mining uses the use of visualization techniques to allow all data miners and analysts to guide the inputs, monitor, and evaluate, process and products of data mining [GHK+96]. As a branch of visual data mining, visual cluster analysis is a combination of information visualization and cluster analysis techniques. In the cluster analysis process, visualization provides analysts with intuitive feedback on data distribution and support decision-making activities. As a consequence, visual presentations can be very powerful in revealing trends, highlighting outliers, showing clusters, and exposing gaps in data [Shn01].A large number of visualization techniques have been developed to map multidimensional datasets to two or three dimensional space [WTP+95, AhW95, HKW99, RSE99, RKJ+99,ABK+99, AEK00, SGB00, MRC02, FGW02, Shn01, PGW03, LKS+04]. In practice, most of them simply take information visualization as a layout problem; therefore, they are not suitable to visualize clusters of very large datasets.

3.1 Multidimensional Data Visualization-

Many efforts have been performed on multidimensional (d>3) data visualization [OlL03].However, most of those visual approaches have difficulty in dealing with high dimensional and very large datasets. We give a more detailed discussion of them as follows.

3.1.1 Icon-based techniques-

Icon-based presentations are relatively older techniques for visual data mining. The idea of icon based techniques is to map each multidimensional data item as an icon, for example [Pic70, Che73, Bed90, Lev91, KeK94, and Hea95]. We explain several popular techniques below.

- **Chernoff Faces**

A well-known iconic approach is Chernoff faces [Che73]. The Chernoff face uses the two dimensions of multidimensional data to locate a face position in the two display dimensions. The remaining dimensions are mapped to the properties of the face icon, i.e., the shape of nose, mouth, eyes, and the shape of the face itself, as shown in Figure 3-1.
Chernoff face visualization capitalizes on the human sensitivity to faces and facial features. However, the number of data items that can be visualized using the Chernoff face technique is quite limited.

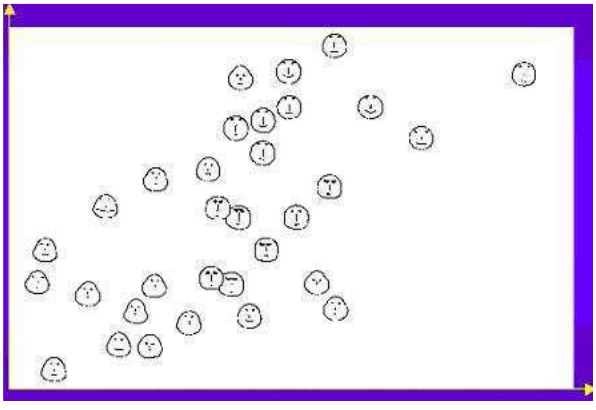
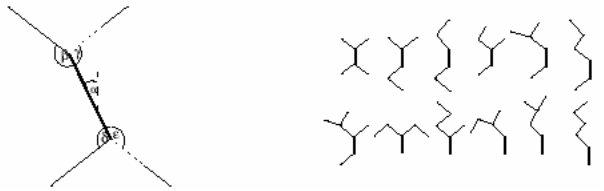


Figure 3-1. An example of Chernoff-Faces

- **Stick Figures**

Another famous icon-based technique is to use stick figures for visualizing a larger amounts of data, therefore, an adequate number of data items can be presented for data mining purposes [Pic70][PiG88]. The stick figures technique uses two dimensions as the display dimensions, and the other dimensions are mapped to the angles and lengths of the stick figure icon, as illustrated in Figure 3-2a. Different stick figure icons with variable dimensionality may be used, as shown in Figure 3-2b.



a. Stick Figure Icon b. A Family of Stick Figures

Figure 3-2. Stick Figure Visualization Technique

Figure 3-3 shows the census data of 1980 United States visualized by the stick figure visualization technique, and the census data have five dimensions. In Figure 3-3, where income and age are used as the display space, and other the attributes: occupation, education level, marital status and sex are visualized by the stick figures. However, it can be observed that, in Figure 4-3, the user cannot easily understand and interpret the graph of stick figures. The user has to have a good training in advance.

Figure 3-3. Stick Figure Visualization of the Census Data

Many other icon-based systems have also been proposed, such as Shape-Coding [Bed90], Color Icons [Lev91, KeK94], and Tile Bars [Hea95]. Icon-based techniques can display multidimensional properties of data, however, with the amount of data increasing, the user hardly makes any sense of most properties of data intuitively, and this is because the user cannot focus on the details of each icon when the data scale is very large.

3.1.2 Pixel-oriented Techniques-

These visualization techniques define each attribute value of data into a single colored pixel, telling the display of the most possible information at a time [KeK94, KKA95, Kei97, Ank01]. With this technique, each data value is mapped to a colored pixel and present the data values belonging to one attribute in separate windows, as displayed in Figure 3-4.

Pixel-oriented techniques use various color mapping approaches, such as linear variation of brightness, maximum variation of hue (color) and constant maximum saturation to map each data value to a colored pixel and arrange them adequately in limited space. Pixel-oriented techniques are powerful to provide an overview of large amounts of data, and meanwhile they preserve the perception of small regions of interest. This feature makes them suitable for being used in a variety of data mining tasks of extremely large databases.

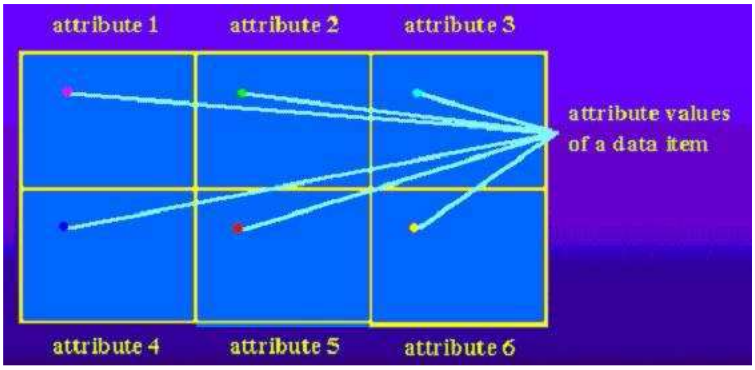


Figure 3-4. Displaying attribute windows for data with six attributes

Keim [KeK94] presented the first Pixel-oriented technique in the VisDB system, which has the capability to represent large amounts of multidimensional data with respect to a given query. As a result, users are able to refine their query based on the knowledge gathered from the visual representation of the data. Other pixel-oriented techniques have been developed, for example, Recursive Pattern Technique [KKA95], Circle Segments Technique [AKK96],Spiral [KeK94], Axes [Kei97], PBC [AEK00] and OPTICS [ABK+99]. They are successfully applied in data exploration for high dimensional databases.

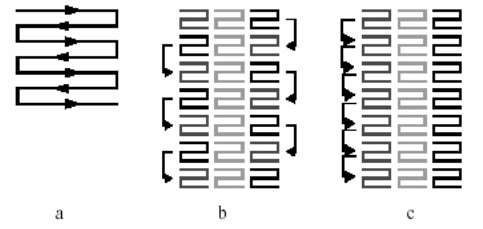


Figure 3-5. Illustration of the Recursive Pattern Technique

For having more expressive data in a limited area, the recursive pattern of pixel-oriented technique has been proposed based on a generic recursive schema [KeK94]. With changeable parameters of the recursive schema, the user can control the semantically meaningful substructures, which determine the arrangement of the attribute values, as presented in Figure3-5. A use of VisDB for visualizing financial information is illustrated in Figure 3-6.

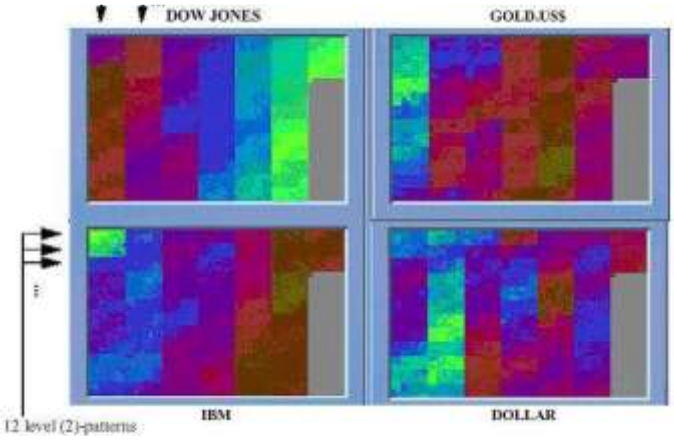


Figure 3-6. The Recursive Pattern Technique in VisDB [KeK94]

In particular, pixel-oriented techniques, aim at representing datasets in the input time order according to one attribute, because clustering arranges data items with similar values closer based on distance/density functions according to the similarity/dissimilarity measures.

However, the close data items are colored similarly, but distributed in time series order by the pixel-oriented techniques, which cannot visualize the insight of clusters very well.
Therefore, they are not suitable to be imposed as visual representation methods in cluster analysis very well.

3.1.3 Geometric Techniques-

The basic idea of geometric techniques is to visualize the geometric transformations and projections of the data to produce useful and insightful visualizations. Geometric projection techniques aim at finding "interesting" projections of multidimensional data sets [Hub85] [FrT74]. The typical systems used the geometric techniques are Scatterplot-Matrices [And72, Che73], Parallel Coordinates [Ins85, ID90], Star Plots [Fie79], Landscapes [WTP+95], Projection Pursuit Techniques [Hub85], Prosecution Views [FuB94, STD+95] and Hyper slice [WiL93]. Here we introduce several of them as follows.

- **Scatterplot-Matrices**

Plot-based data visualization approaches such as Scatterplot-Matrices [Cle93] and similar techniques [AlC91] visualize data in rows and columns of cells containing simple graphical depictions.

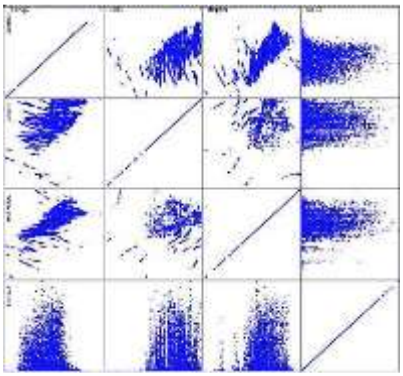


Figure 3-7. Scatterplot-Matrices [Cle93]

This category of techniques gives bi-attributes visual information. An example of Scatterplot-Matrices is shown in Figure 3-7. The user can clearly observe each bi-attributes' data distribution.

However, plot-based techniques do not provide the best overview of the whole dataset. As a result, they are not able to present clusters of datasets very well. On the other hand, Plot-based visual techniques do not perform well on the presentation of large number of dimensional databases, due to the physical size limitation of computer monitors.

- **Parallel Coordinates**

A famous multidimensional visualization technique, Parallel Coordinates, utilizes equidistant parallel axes to visualize each attribute of a given dataset and projects multiple dimensions ona two-dimensional surface [Ins97]. The axes which is corresponding to the dimensional data and is linearly scaled from the minimum value to the maximum value of the corresponding dimension. Each data item is presented as a polygonal line, intersecting, as presented in Figure 3-8.

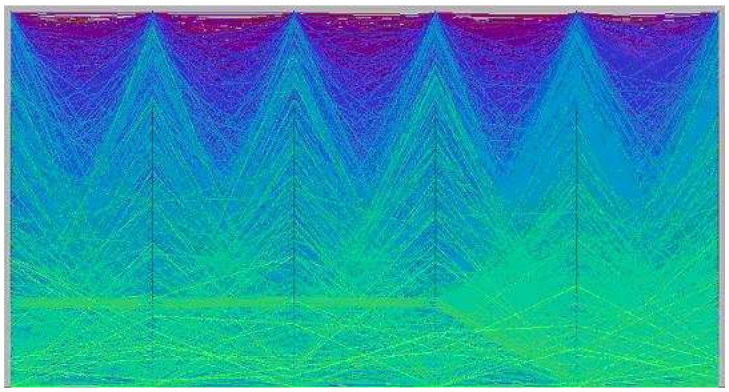


Figure 3-8. 15,000 colored data items in Parallel Coordinates

Star Plots arranges coordinate axes on a circle space with equal angles between neighboring axes from the centre of a circle and links data points on each axis by lines to form a star
[SGF71]. An example of Star Plots technique is presented in Figure 3-9.

In principle, these two techniques can provide visual presentations of any number of attributes. However, neither Parallel Coordinates nor Star Plots is adequate to give the user a clear overall

insight of data distribution when the dataset is huge, primarily due to the unavoidably high overlapping between data points. And another drawback of these two techniques is that, though they can supply a more intuitive visual relationship between the neighboring axes, for the non-neighboring axes, the visual presentation may confuse the users' perception. These obstacles make them not properly to visualize cluster structure in very large and high dimensional databases.

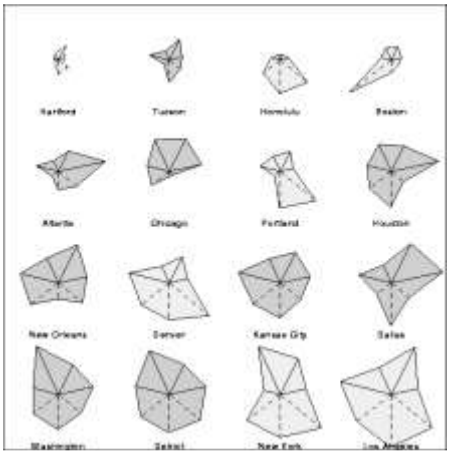


Figure 3-9. Star plots of data items [SGF71]

An integrated multidimensional data visualization system, Xmdv Tool has been proposed, which includes such as scatter plot matrix, parallel coordinates, star plots, and dimensional stacking by linking alternative views using brushing [War95]. Many other techniques have been introduced in multidimensional data visualization [OlL03]. However, most of them either suffer from the weakness on visualizing large amount data items and higher dimensional data or hardly provide clearly clustered perception in visual form to the user. The multidimensional data visualization techniques developed earlier can be found in the literature [BeR78, Fie79, HoG01, Che07, FrD07].

In the last decade, many efforts have been made on using visualization techniques to assist data miners for finding cluster patterns in data. A survey of these visualization techniques is presented below.

3.2 Visual Cluster Analysis-

Visual cluster analysis, as the term implies, is a discipline of information visualization and cluster analysis techniques. With wide applications of cluster analysis in data mining, many Visualization techniques have been employed to study the structure of datasets in the applications of cluster analysis. Reviews of these works can be found in the literature [AnK01, HoG01, Kei02, OlL03, MiG04]. Several representative visualization techniques that are especially important in cluster analysis are discussed below.

3.2.1 MDS and PCA-

Multidimensional scaling (MDS) maps multidimensional data as points in a 2D Euclidean space, where the distances between data points reflect similarity/dissimilarity of them [KrW78] as illustrated in Figure 3-10. However, the relative high computational cost of MDS(polynomial time O(N2)) limits its usability in very large datasets.

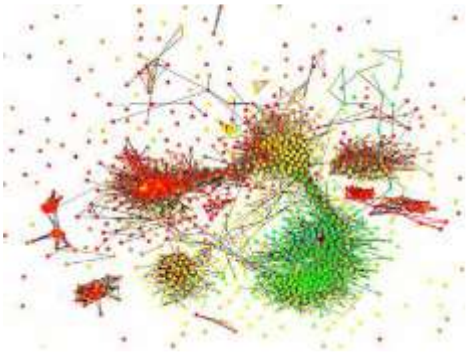


Figure 3-10. Clustering of 1352genes in MDS by [Bes]

Principal Component Analysis (PCA) involves a mathematical procedure that transforms a number of (possibly) correlated variables (of higher dimensions) into a number of uncorrelated variables (of smaller, lower dimensions) called principal components [Jol02].
PCA first has to find the correlated variables for reducing the dimensionality, which restricts its performance in the exploration of unknown data.

3.2.2 HD-Eye-

HD-Eye is an interactive visual clustering system based on density-plots of any two interesting dimensions [HKW03]. It projects any two dimensions of the multi dimensional data based on density-plots to investigate interested grouping clues. HD-Eye employs icons to represent the clusters and the relationship between the clusters, as shown in Figure 3-11.

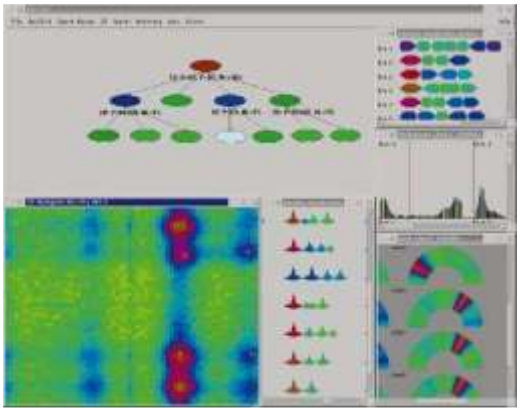


Figure 3-11. The framework of HD-Eye system and its different visualization projections

HD Eye provides the user a rough data structure of a high dimensional data in visual representations. Whereas, the user hardly synthesizes all of the interesting 2D projections to find the general pattern of the clusters.

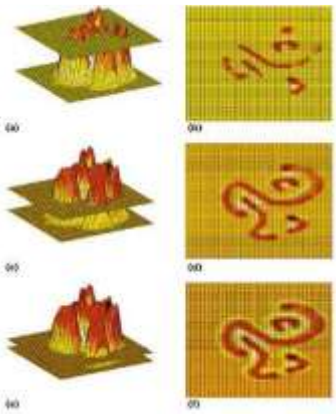


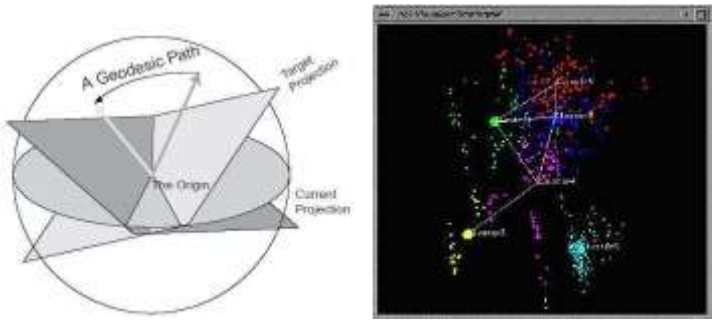
Figure 3-12. The 3D data structures in HD-Eye and their intersection trails on the plane

HD-Eye is also addressed to use 3D techniques to visualize data in mountain-like structures, and use the intersected planes of the 3D graphs for presenting the trails of the graphs on the planes in different level in 2D forms [HKW99], as illustrated in Figure 3-12. But the kernel density-based 3D graphs formation in HD-Eye limits it to be employed in the interactive cluster detection of large datasets.

3.2.3 Grand Tour-

The Grand Tour technique uses a series of variable projections to map multidimensional data onto a two orthogonal 2D space in order to obtain different perspectives of data [Asi85]. For effectively reducing huge search space of data, Projection Pursuit is introduced to help the user for the purpose of investigating the interesting projections [CBC+95]. The projection of Grand Tour with Projection Pursuit is illustrated in Figure 3-13a. However, due to Grand Tour systems have several times projections and complicated computation; their visualization models are not intuitive to users.

Based on the Grand Tour technique, several extensions have been proposed. For example, Yang implemented a 3D version Grand Tour technique to project data in animations [Yan03], but the complex 3D graph formation of this technique limits its use in large scale data visualization. An example of Yang's Grand Tour based visualization is presented in Figure 3-13b. Dhillon, et al. proposed a technique to visualize cluster structure [DMS98], but their technique visualizes 3 clusters. It requires more assistance with a sophisticated Grand Tour technique to deal with more than 3 clusters.



a. The projections of the Grand Tour technique b. Grand Tour based 3D animation by [Yan03

Figure 3-13. The Grand Tour Technique and its 3D example

3.2.4 Hierarchical BLOB-

Based on the hierarchical clustering and visualization algorithm H-BLOB, Sprenger et al presented a technique for visualizing hierarchical clusters in the nested blobs [SBG00], as shown in Figure 3-14.

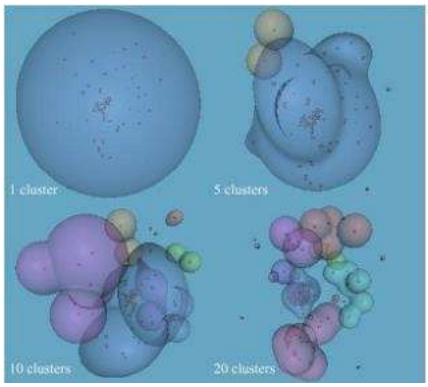


Figure 3-14. Cluster hierarchies are shown for 1, 5, 10 and 20 clusters [SBG00]

The most significant feature of their technique is that, H-BLOB not only provides the overview manner of whole dataset in blobs, but also gives the detailed visual representation of lower leveled clusters. Exhibiting clusters in the form of blobs H-BLOB results in a very intuitive and easily understood visual presentation. However the high visual complexity of the two stages of blob graphs' formation makes them unsuitable to be applied in cluster visualization of very large sized datasets.

3.2.5 SOM-

Kaski el. al employs Self-organizing maps (SOM) technique [Koh97] to project multidimensional data sets to 2D space for matching visual models [KSP01]. Technically, in their method, a sample data is mapped into a bar graph, then the graph is compared with all existing vector models in bar graphs to find the most matched one, as shown in Figure 3-15, where the bar graphs in the rectangular region are existing models, Xk is the sample bar graph.

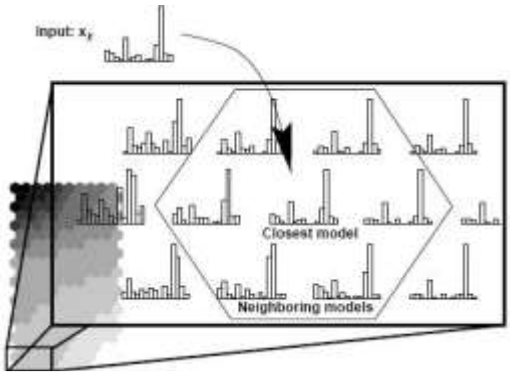


Figure 3-15. Model matching with SOM by [KSP01]

However, the traversal matching process is time-consuming. On the other hand, the SOM technique is based on a single projection strategy. It is not powerful enough to discover all the interesting features from the original data. Another drawback of this technique is that, with an increasing number of dimensions, the bar graphs would be wider. As a result, the user cannot easily observe the matched cluster model in intuition.

3.2.6FastMap-

Huang et.al proposed the approaches based on Fast Map [FaL95] to assist users in identifying and verifying the validity of clusters in visual form [HCN01, HuL00]. Their techniques work well in cluster identification, but are unable to evaluate the cluster quality very well. On the other hand, these techniques visualize clusters statically and do not always present the genuine cluster structure. As a consequence, they do not provide enough information for either clustering or cluster validation.

3.2.7 OPTICS-

OPTICS uses a density-based technique to detect cluster structure and visualizes them in "Gaussian bumps" [ABK+99]. It is an intuitive method to assist the user to observe cluster structures. But its non-linear time complexity makes it neither suitable to deal with very large data sets, nor suitable to provide the contrast between clustering results, as shown in Figure3-16.

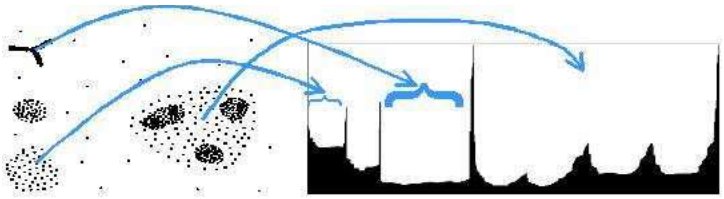


Figure 3-16. Data structure mapped in Gaussian bumps by OPTICS [ABK+99]

OPTICS also visualizes clusters in 1D visualization manner [ABK+99]. It works well in finding the basic arbitrarily shaped clusters, as presented in Figure 3-17. However it lacks the ability in helping the user understand inter-cluster relationships.

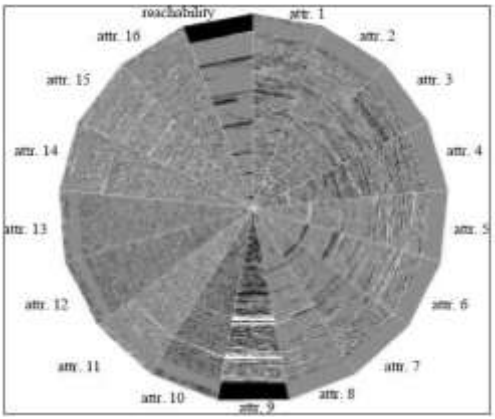


Figure 3-17. Clustering structure of 30,000 16-Dimensional data items Visualized by OPTICS [ABK+99]

3.2.8 Star Coordinates and VISTA-

The most relevant approach to this thesis is the Star Coordinates technique [Kan01]. The idea of Star Coordinates technique is intuitive, which extends the perspective of traditional orthogonal 2D X-Y and 3D X-Y-Z coordinate's technique to a higher dimensional space. Star coordinates plots a 2D plane into n equal sectors with n coordinate axes, where each axis represents a dimension and all axes share the initials at the centre of a circle on the 2D space [Kan01].

First, data in each dimension are normalized into [0, 1] interval. Then the values of all axes are mapped to orthogonal X-Y coordinates which share the initial point with Star Coordinates on the 2D space. Thus, an n-dimensional data item is expressed as a point in the X-Y 2D plane. Figure 3-18 illustrates the mapping from 8 Star Coordinates to X-Y coordinates.

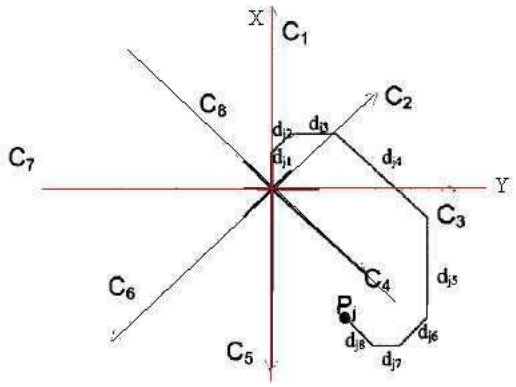


Figure 3-18. Positioning a point by an 8-attribute vector in Star Coordinates [Kan01]

Star Coordinates provides users the ability to apply various transformations dynamically, integrate and separate dimensions of interest, analyze correlations of multiple dimensions, view clusters, trends, and outliers in the distribution of data.

Formula (4-1) states the mathematical description of Star Coordinates.

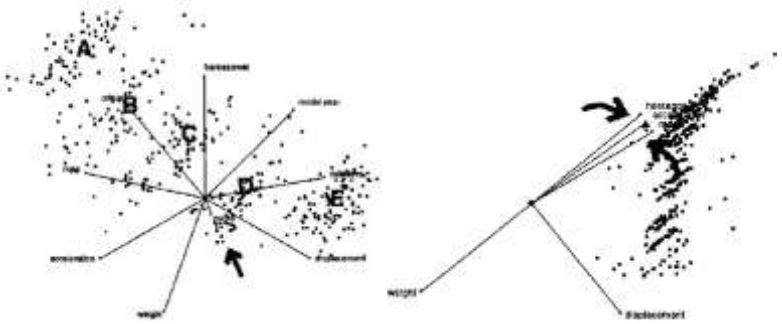
$$p_j(x,y) = \left(\sum_{i=1}^{n} \bar{u}_{xi}(d_{ji} - \min_i), \sum_{i=1}^{n} \bar{u}_{yi}(d_{ji} - \min_i)\right) \quad (4\text{-}1)$$

where pj (x, y) is the normalized location of Dj=(dj1, dj2, …., djm), and dji is the value of the jth record of a data set on the ith coordinate Ci in Star Coordinates space; ūxi·(dji-mini) and ūyi·(dji-mini) are unit vectors of dji mapping to X direction and Y direction, mini=min(dji,0≤j<m) and maxi=max (dji, 0≤ j<m) are the minimum and maximum values of the ith dimension respectively; and m is the number of records in the data set.

As presented in the formula (4-1), the computational complexity of Star Coordinates projection is linear time. Therefore, the Star Coordinates based techniques are powerful for interactive visualization and analysis of clusters.

- **Interactive Functions of Star Coordinates**

Star Coordinates provides various interactive functions to stimulate visual thinking in early stages of the knowledge discovery process. Those functions include scaling axes, see Figure 4-20-a; rotating angle between axes Figure 4-20-b; marking data points in a certain area by coloring; selecting data value ranges on one or more axes and marking the corresponding data points in the visualization; presenting histograms of selected clusters; foot print for tracing the foot points for data points, see Figure 3-29-c and etc [Kan01].



a. Axis scaling of "name" attribute of auto-mpg data

b. Angle rotation of the attributes of auto-mpg data

c. Foot prints of axes scaling of "weight" and "mpg" attributes of auto-mpg data

Figure 3-19 axis scaling, angle rotation and foot print functions of Star Coordinates [Kan01]

- **The Star Coordinates based techniques**

Based on the idea of the Star Coordinates, instead of normalizing data in each dimension into [0, 1] interval by Star Coordinates, Chen and Liu proposed an approach, named a-mapping, which normalizes data into [-1, 1] interval of each dimension for having more expressive space of axis scaling in their VISTA and/or iVIBRATE systems [ChL04, ChL06]. Moreover, Chen and Liu also discussed using their approach to refine and verify clusters by VISTA/iVIBRATE [ChL06]. Shaik and Yeasin addressed a 3D manner of the Star Coordinates to provide users a more intuitive 3D environment for the observation of cluster structure [ShY06]. Ma and Teoh employed the Star Coordinates technique into their Star Class system for data classification by visualization [MaT03].

A very similar technique of Star Coordinate, RadViz were presented by [HGM97]. But its non-linear mapping is an obstacle for RadViz to be employed as an interactive tool for cluster analysis of very large-sized databases.

- **The issues of the existing Star Coordinates based Visualization Techniques**

The existing Star Coordinates based visualization techniques employed in cluster analysis tend to be used as information rendering tools, but do not perform well on verifying the validation of clustering results. On the other hand, the exploration-oriented characteristics of these techniques, inevitably lead them to be random and imprecise in the process of cluster detection and validation.

Chen and Liu combined clustering algorithms and their visualization technique VISTA/iVIBRATE to observe cluster structures of datasets, refine the quality of clusters produced by clustering algorithms, and validate clusters [ChL04, ChL05]. However, the data observation based on a-mapping (a-adjustment) of their approach is still an randomly exploratory process, which inevitably suffers from subjectivity and randomness. In addition, VISTA adopts “landmark” points as representatives from a clustered subset and re samples them to deal with cluster validation [ChL04]. But its experience-based “landmark” point selection does not always handle the scalability of data very well, due to the well representatively and mark points selected in a subset may fail in other subsets of a database.

3.3 Major Challenges-

Visualization is considered as a collection of transformations from the “problem domain” to the “representation domain” [GMH+94]. A more practical and effective approach of cluster visualization is to incorporate all available clustering information, for example algorithmic clustering results and the domain knowledge, into visual cluster exploration.

3.3.1 Requirements of Visualization in Cluster Analysis

By the above analysis, we can summarize that the visualization techniques to be used in cluster analysis should be able to handle several important aspects of visual perception:

1. Visualizing large and multidimensional datasets;
2. Providing a clear overview and detailed insight of cluster structure;
3. Having linear time complexity on data mapping from higher dimensional space to lower dimensional space;
4. Supporting interactive cluster visual representation dynamically;
5. Involving knowledge of domain experts into the cluster exploration;
6. Giving data miners purposeful and precise guidance of cluster investigation and cluster validation rather than simply random cluster exploration.

As discussed above, most existing cluster visualization techniques work well on visualizing multidimensional data sets. However, as the size and dimensionality of data sets increase, these techniques do not perform well on very large data visualization, they can hardly deal with visual representation of higher dimensional data, they cannot provide an intuitive overview of cluster structure, etc. In short, they satisfy not all of the above requirements.

3.3.2 Motivation

A question arises: which visualization technique can provide a genuine representation of cluster structure of data? In practice, a few visualization techniques can achieve the above requirements.

As Seo and Shneiderman pointed out, “A large number of clustering algorithms have been developed, but only a small number of cluster visualization tools are available to facilitate researchers’ understanding of the clustering results” [SeS05]. How to preserve the identity of “problem domain” and “representation domain” by visualization is the critical challenge of cluster visualization.

Star Coordinates based techniques are a good choice for cluster visualization, because they almost meet all the considerations above, except the last one. Simple static visualization is not sufficient in visualizing clusters [Kei01, Shn02], and it has been shown that clusters can hardly be satisfactorily preserved in a static visualization [CBC+95, DMS98]. With the feature of linear time transformation/projection, Star Coordinates based techniques are powerful for large scale data visualization, especially for interactive and dynamic cluster visualization. But the random and subjective characteristics of these techniques hinder their effectiveness and efficiency in real-world applications. The main motivation of this thesis is to provide an effective and purposeful visual guidance to data miners in cluster analysis.

3.4 Our Approach-

In this subsection, we briefly describe a novel approach called HOV3 for addressing the challenge presented above. As a publication-based thesis, the detailed discussion of the works in the thesis can be found in the cited papers.

3.4.1 HOV3 Model

Visualization is typically employed as an observational mechanism to assist users with intuitive comparisons and better understanding of the studied data. Instead of precisely contrasting clustering results, most of the existing visualization techniques employed in cluster analysis focus on providing the user with an easy and intuitive understanding of the cluster structure, or explore clusters randomly.

In general, it is not easy to visualize multidimensional data sets on 2D space and give a "genuine" visual interpretation. This is because mapping multidimensional data onto 2Dspace inevitably introduces overlapping and bias. For mitigating the problem, Star Coordinates based techniques provide some visual adjustment mechanisms [Kan01, ChL04,and ChL06]. However, the stochastic adjustment of Star Coordinates and VISTA limits their usability in cluster analysis.

To overcome the arbitrary and random adjustments of Star Coordinates and its extensions, Zhang et al proposed a hypothesis-oriented visual approach (Hypothesis Oriented Verification and Validation by Visualization) HOV3 in short, to detect clusters [ZOZ+06,ZOZ06]. The idea of HOV3 is that, in analytical geometry, the difference of a data set (a matrix) Dj and a measure vector M with the same number of variables as Dj can be represented by their inner product, Dj·M. HOV3 uses a measure vector M to represent the corresponding axes' weight values. Then given a non-zero measure vector M in Ñn, and a family of vectors Pj, the projection of Pj against M in the complex number system, the HOV3model is presented as:

$$P_j(z_0) = \sum_{k=1}^{n} [(d_{jk} - \min_k(d_k)) / (\max_k(d_k) - \min_k(d_k)) \cdot z_0^k \cdot m_k] \quad (3\text{-}1)$$

where (dk) and (dk) represent the minimal and maximal values of kth dimension respectively; and mk is the kth attribute of measure M.

The aim of interactive adjustments of Star Coordinates and its extensions is to have some separated groups or full-separated clustering result of data by tuning the weight value of each axis (axis scaling in Star Coordinates, a-adjustment in VISTA/ iVIBRATE), but their arbitrary and random adjustments limit their applicability. As shown in formula (3-2), HOV3summarizes these adjustments as a coefficient/measure vector. Compared the formulas (3-1) and (3-2), it can be observed that HOV3 subsumes the Star Coordinates model [ZOZ06]. Thus the HOV3 model provides the user a mechanism to quantify a hypothesis/prediction about a data set as a measure vector of HOV3 for precisely exploring grouping information.

4. CONCLUSION AND FUTURE WORK

4.1 Conclusion-

This thesis has proposed a novel visual approach called HOV3, Hypothesis Oriented Verification and Validation by Visualization, to assist data miners in cluster analysis of high dimensional datasets. HOV3 provides data miners an effective mechanism to introduce their quantified domain knowledge as predictions in the cluster exploration process for revealing the gaps of data distribution against the predictions. As a result, it is more efficient and purposeful by using HOV3 to investigate cluster clues in very large and high-dimensional datasets.

This thesis has also proposed a visual cluster validation approach based on distribution matching supported by the projection mechanism of HOV3. This approach is based on the assumption that by using a measure vector to project the data sets in the similar cluster structure, the similarity of the changes of their behavior of data distribution should be high. By comparing the data distributions of a clustered subset and non-clustered subsets projected by HOV3 with measures, the data miners can intuitively have a visual assessment, and also have a precise evaluation of the consistency of the cluster structure by performing geometrical computation on their data distributions. Compared with existing visual techniques involved in cluster validation, it has been observed that this approach is not only efficient in performance, but also effective in real-world applications.

Based on the projection technique of HOV3, a visual approach called M-HOV3/M-mapping has also been introduced to enhance the visual separation of clusters. The visual severability of clusters is significant for cluster analysis. Fully geometrical separation of clusters is not only beneficial in revealing the membership formation of clusters, but also beneficial in verifying the validity of clustering results. With M-HOV3/M-mapping, data miners can both explore cluster distribution intuitively and verify clustering results easily by matching the similarity between the geometrical distributions of clustered and non-clustered subsets simultaneously produced by MHOV3/M-mapping.

Experiments show that HOV3 technique can improve the effectiveness of cluster analysis by visualization. HOV3 can be seen as a bridging process between qualitative analysis and quantitative analysis. It not only supports quantified domain knowledge verification and validation, but also directly utilizes the abundant statistical measurements of the studied data as predictions in order to give data miners an effective guidance for having more precise cluster information in data mining.

As a consequence, with the advantage of the quantified measurement feature of HOV3 data miners can identify the cluster number in the pre-processing stage of clustering efficiently, and also verify and refine the membership of data points among the clusters effectively in the post-processing stage of clustering. We believe the application of HOV3 will be fruitful.

4.2 Future Work-

This thesis has addressed the challenges of introducing visualization techniques to cluster analysis in data mining, and proposed a visual technique called HOV3 to mitigate the problems in visual cluster analysis. However, there are still some open research issues worth future efforts.

4.2.1 Three Dimensional HOV3

This thesis has introduced the quantified measures as predictions with HOV3 to detect cluster clues and verify clustering results in clustering large datasets that automated clustering algorithms cannot effectively handle. So far HOV3 projects high dimensional data onto 2Dspace [ZOZ06]. 3D visualization can provide more intuitions and also more information on the studied data [Rei95]. However, most of the existing 3D visual techniques involved in cluster analysis are density-based or metaphor-based [KOC+04]. They suffer from the high computational cost on composing 3D graphs of clusters. This drawback limits them to be applied for 3D cluster investigation in very large databases, especially for 3D interactive cluster exploration. [Yan03]. Recently, Shaik and Yeasin proposed a 3D visualization model based on the Star Coordinates technique [ShY06]. However, the relatively complex projection of 3D formation of their approach is a drawback on 3D visualization of large datasets.

In fact, with the known two orthogonal vectors in HOV3 to compose the third dimensional vector is not hard. Then the 3D visual presentation of data in HOV3 can be produced by linear combinations of the three vectors. Based on the advantage of linear time complexity of HOV3 projection, the 3D HOV3 projection is also in linear time. Thus data miners may more effectively grasp the cluster clues from the studied datasets by the interaction of 3D HOV3exploration.

4.2.2 Dynamic Visual Cluster Analysis

Dynamic clustering is also called stepwise clustering, which is a kind of iterative clustering method based on distance [ZHW+03]. Dynamic clustering intends to study the groups' behavior changes and revise clusters dynamically along with the cluster exploration process, even revise the criteria of clustering. It deals with data grouping as a cluster analysis in time series [BeC96, AbM98, CHS04]. In each clustering iteration, clustering algorithms sample at the time series points and revise the formation of clusters dynamically by given criteria.

However, the existing clustering algorithms do not perform well with arbitrarily shaped data distribution of the datasets, and the very high computational cost of the statistics-based cluster validation methods limits their usability in real time applications. Based on the HOV3 model, We have proposed a cluster validation method based on distribution matching in this thesis [ZOZ07a]. This approach can provide a solution to the above problem, because the approach only calculates the overlapping rate between the classifier (a clustered subset of a dataset) and its geometrically covered data points. It is much quicker than the existing statistics based cluster validation methods [ZOZ07a]. For revising clustering criteria, the newly produced clustering criteria can be generated automatically by the density function of the data points of overlapped area.

4.2.3 Quasi-Cluster Data Points Collection

Based on the quantified measurement feature of HOV3, an external cluster validation based on distribution matching has been proposed in this thesis to verify the consistency of cluster

structures between a clustered subset and non-clustered subsets of a dataset [ZOZ07a]. But, so far, the quasi-cluster point is picked up manually by a geometrical intuition. The Newton method of data analysis can be introduced into this quasi-cluster point collection. The Newton method is an efficient approach to find the neighboring points of a given point [Smi86]. This would improve the accuracy and effectiveness of quasi-cluster point collection in HOV3.

4.2.4 Combination of Fuzzy Logical approaches and HOV3

Fuzzy clustering is an active branch of cluster analysis [OlP07]. Instead of data points being only exactly assigned into one cluster, in fuzzy (soft) clustering, data points can belong to more than one cluster [Sim93]. The data points can be associated with different grades with clusters. The grades of data points indicate the nearness degree of relationship to clusters.

However, when fuzzy clustering algorithms deal with dynamic clustering applications, the recompilation of the grades of membership associated with clusters is very high [BLO+03]. In fuzzy clustering proposed in [BaB99], each data point has a vector V(1...k) associated with the K clusters [Bez81]. In this thesis, HOV3 model has been proposed to assist data miners in cluster investigation and verification [ZOZ+06, ZOZ06]. The HOV3 model is presented in the formula (8) in [ZOZ06]. There, the measure coefficient mk of the k^{th} dimension can be combined with the associated grade of each data point. Thus with the color mapping function[Fai98], HOV3 could provide very intuitive visual presentation of the membership of each data point, due to the closest data points being colored similarly. This approach would be very helpful to the data miners to identify the membership formation of clusters during interactive cluster exploration.

BIBLIOGRAPHY

[AAP+003] A. L. Abul, R. Alhajj, F. Polat, and K. Barker, “Cluster Validity Analysis Using Sub sampling”, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics, IEEE Press, Vol. 2: pp. 1435-1440 (2003)

[ABK+99] M. Ankerst, M. Breunig, H.-P. Kriegel, J. Sander “OPTICS: Ordering Points To Identify the Clustering Structure”, Proceedings of ACM SIGMOD ‘99, International Conference on Management of Data, Philadelphia, PA. pp. 49-60 (1999)

[AbM98] A. J. Abrantesy , J. S. Marques, “A Method for Dynamic Clustering of Data”, Proceedings of the British Machine Vision Conference 1998, BMVC 1998, Southampton, UK,1998. British Machine Vision Association, pp.154-163 (1998)

[AEK00] M.Ankerst, M. Ester M, H. P. Kriegel, “Towards an Effective Cooperation of the Computer and the User for Classification“, Proceedings of. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2000), Boston, MA, pp. 179-188 (2000)

[AGG+98] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan, “Automatic subspace clustering of high dimensional data for data mining applications”, Proceedings of the ACM SIGMOD Conference, Seattle, WA., pp.94-105 (1998)

[AhW95] C. Ahlberg, E. Wistrand, “IVEE: An Environment for Automatic Creation of bDynamic Queries Applications”, Proceedings of Human Factors in Computing Systems CHI ’95 Conference, Demo Program, Denver, CO (1995)

[AKK96] M. Ankerst, D. A. Keim, H.-P. Kriegel, “Circle Segments: A Technique for Visually Exploring Large Multidimensional Data Sets”, Proceedings of Visualization ‘96, Hot Topic Session, San Francisco, CA, 1996.

[ALA+03] O. Abult, A. Lo, R. Alhajjt, F. Polat, K. Barked, “Cluster Validity Analysis Using Sub sampling”, Proceedings of IEEE International Conference on Systems, Man and Cybernetics (IEEE-SMC), Vol.2 pp.1435- 1440 (2003)

[AlC91] B. Alpern, L. Carter. ”Hyperbox”, Proceedings of Visualization ‘91, San Diego, CA,pp.133-139 (1991)

[And72] D. F. Andrews, ”Plots of High-Dimensional Data”, Biometrics, Vol. 29, pp. 125-136(1972)

[And73] M. Anderberg, Cluster Analysis for Applications. New York: Academic (1973)

[AnK01] M. Ankerst, and D. Keim, "Visual Data Mining and Exploration of Large Databases", Proceedings of 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01), Freiburg, Germany, September 2001

[Asi85] D. Asimov, "The grand tour: A tool for viewing multidimensional," SIAM Journal of Scientific and Statistical Computing, Vol. 6 (1), pp. 128-143 (1985)

[BaB99]A. Baraldi, P. Blonda, "A Survey of Fuzzy Clustering Algorithms for Pattern Recognition—Part I", IEEE Transactions on Systems, Man, and Cybernetics—Part b:Cybernetics, vol. 29(6), pp.778-785 (1999)

[BeC96] D. J. Berndt and J. Clifford, "Finding patterns in time series: A dynamic programming approach," in Advances in Knowledge Discovery and Data Mining. Menlo Park, CA: AAAI/MIT Press, 1996, pp. 229–248.

[Bed90] J. Beddow, 'Shape Coding of Multidimensional Data on a Mirco computer Display", Proceedings of Visualization '90, San Francisco, CA, 1990, pp. 238-246.

[BeG03] A. Ben-Hur and I. Guyon, "Detecting stable clusters using principal component analysis," Methods in Molecular Biology, M.J. Brownstein and A. Kohodursky (eds.)Humana press, pp.159-182 (2003)

[BEG02] A. Ben-Hur, A. Elisseeff and I. Guyon, "A stability based method for discovering structure in clustered data," Proceedings of the Pacific Symposium on Bio computing (2002)

[Ber06] Berkhin, P: A Survey of Clustering Data Mining Techniques, Kogan, Jacob; Nicholas, Charles; Teboulle, Marc (Eds.) Grouping Multidimensional Data, Springer Presspp. 25-72 (2006)

[BeR78] J. R. Beniger and D. L. Robyn, "Quantitative graphics in statistics: A brief history",The American Statistician, Vol 32(1) pp. 1-9 (1978)

[Bes] C. Best, http://www.computationalgroup.com/tigertiger/cb/index.html[Bez81] J. C. Bezdek, Pattern Recognition with Fuzzy Objective Function, Alnon'thms.Plenum Press. New York. 1981

[BLO+03] M. Buerki, K.O. Lovblad, H. Oswald, A.C. Nirkko, P. Stein, C. Kiefer and G.Schroth, "Multiresolution fuzzy clustering of functional MRI data", Neuroradiology Vol.45,pp.691-699 (2003)

[BPR+04] C. Baumgartner, C. Plant, K. Railing, H-P. Kriegel, P. Kroger, "Subspace Selection for Clustering High-Dimensional Data", Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM'04), pp:11-18 (2004)

[CBC+95] D. Cook, A. Buja, J. Cabrera, and C. Hurley, “Grand tour and projection pursuit”, Journal of Computational and Graphical Statistics, vol. 23, pp.155-172 (1995)
[Che07] C. Chen, “A Brief History of Data Visualization”, W. Hardle and A. Unwin (eds.),Handbook of Computational Statistics: Data Visualization, Vol III, Springer, 2007.

[Che73] H. Chernoff, “The Use of Faces to Represent Points in k-Dimensional Space Graphically”, Journal Amer. Statistical Association, Vol. 68, pp.361-368 (1973)

[Chi00] E. Chi. “A taxonomy of visualization techniques using the data state reference model”, Proceedings of the Symposium on Information Visualization (InfoVis’2000), pp.69-75(2000)

[CHS04] W.-P. Chen, J. C. Hou, L. Sha, “Dynamic Clustering for Acoustic Target Tracking in Wireless Sensor Networks”, IEEE Transactions on Mobile Computing, Vol. 3 (3), July-September 2004, pp. 358-371

[CKS+88]] P. Cheeseman, J. Kelly, M. Self, J. Stutz, W. Taylor, and D. Freeman, “AutoClass: A bayesian classification system”, Proceedings of 5th International Conference on Machine Learning, Morgan Kaufmann, pp. 54-64 (1988)

[Cle93] W. S. Cleveland, Visualizing Data”, AT&T Bell Laboratories, Murray Hill, NJ,Hobart Press, Summit NJ, (1993)

[CMS99] S. K. Card, J. D. Mackinlay, and B. Shneiderman, editors. Readings in Information Visualization: Using Vision to Think. Morgan Kaufmann, San Francisco, 1999.

[DMS98] I. S. Dhillon, D. S. Modha, and W. S. Spangler, “Visualizing class structure of multi dimensional data,” the 30th Symposium on the Interface: Computing Science and Statistics,Vol. 30, pp.488–493 (1998)

[Dom01] B. Dom, "An information-theoretic external cluster-validity measure", Research Report, IBM T.J. Watson Research Center RJ 10219 (2001)

[DuJ79] R. Dubes and A. K. Jain, "Validity studies in clustering methodologies", Pattern Recognition, Vol. 1(1), pp.235-254 (1979)

[EKS+96] M. Ester, H-P Kriegel, J. Sander, X. Xu, “A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise”, Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), pp.226-231(1996)

[ELL01] B. Everitt, S. Landau , and M. Leese, Cluster Analysis. London: Arnold, 2001.

[Fai98] Mark D. Fairchild, Color Appearance Models, Addison-Wesley, Reading, MA (1998)

[FaL95] C. Faloutsos and K. Lin, Fastmap: a fast algorithm for indexing, data-mining and visualization of traditional and multimedia data sets" Proceedings of ACM-SIGMOD95,pp.163-174. (1995)

[FGW02] U. Fayyad, G. Grinstein and A. Wierse (eds.), "Information Visualization in Data Mining and Knowledge Discovery", Morgan Kaufmann Publishers, 2002

[Fie79] S. E. Fienberg, "Graphical methods in statistics", American Statisticians Vol.33 pp.165-178 (1979)

[Fis87] D. Fisher, "Improving Inference through Conceptual Clustering", Proceedings of1987 AAAI Conferences, Seattle Washington, pp.461-465 (1987)

[FoM83] E. Fowlkes and C. Mallows, "A method for comparing two hierarchical clusterings", "Journal of American Statistical Association", Vol. 78, pp.553–569 (1983)

[FrD01] J Fridlyand J. and Dudoit S., "Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method," University of California, Statistics Department Technical Report, No.600 (2001)

[FrD07] M. Friendly, D. J. Denis, "Milestones in the history of thematic cartography, statistical graphics and data visualization", York University, Canada (2007)

[FrT74] J. Friedman, J. Tukey, "A Projection Pursuit Algorithm for Exploratory DataAnalysis", IEEE Transactions on Computers, Vol. 23, pp. 881-890 (1974)

[FuB94] G. W. Furnas, A. Buja, "Prosections Views: Dimensional Inference through Sectionsand Projections", Journal of Computational and Graphical Statistics, Vol. 3(4), pp.323-353(1994)

[Fuk90] K. Fukunaga, "Introduction to Statistical Pattern Recognition", San Diego CA, Academic Press (1990)

[GMH+94] G. Grinstein T. Mihalisin, H. Hinterberger A. Inselberg, "Visualizing multidimensional (multivariate) data and relations", Proceedings of the conference on Visualization '94, IEEE Visualization, pp. 404-409 (1994)

[Gor98] A. D. Gordon, "Cluster validation", Data Science, Classification, and RelatedMethods, C. Hayashi, N. Ohsumi, K. Yajima, Y. Tanaka, H-H. Bock and Y. Baba Edited Springer, Tokyo, pp 22-39(1998)

[GRS98] S. Guha, R. Rastogh and K. Shim, "CURE: An efficient clustering algorithm for large databases," Proceedings of ACM SIGMOD Conference 98, pp.73–84 (1998)

[HaK01] J. Han and M. Kamber, “Data Mining: Concepts and Techniques,” Morgan Kaufmann Publishers (2001)

[HaJ97] P. Hansen and B. Jaumard, “Cluster analysis and mathematical programming,” Math. Program, Vol. 79, pp.191–215 (1997)

[HaV01] M. Halkidi and M. Vazirgiannis, “Clustering Validity Assessment: Finding theOptimal Partitioning of a Data Set”, Proceedings of ICDM 2001, pp. 187-194 (2001)

[Har75] J. Hartigan, Clustering Algorithms. New York: Wiley (1975)

[HBV01] M. Halkidi, Y. Batistakis and M. Vazirgiannis M., “On Clustering Validation Techniques,” Journal of Intelligent Infomation Systems, Vo1.7(2-3) (2001)

[HBV02] M. Halkidi, Y. Batistakis, M. Vazirgiannis: “Cluster Validity Methods: Part I&II”, SIGMOD Record,Vol. 31(2-3) (2002

[HCN01] Z. Huang, D. W. Cheung, M. K. Ng, “An Empirical Study on the Visual Cluster Validation Method with Fastmap”, Proceedings of the 7th International Conference on Database Systems for Advanced Applications, pp. 84-91 (2001)

[Hea95] M. Hearst, “Tile Bars: Visualization of Term Distribution Information in Full Text Information Access”, Proceedings of ACM Human Factors in Computing Systems Conference, (CHI'95), pp.59-66 (1995)

[HGM97] P. Hoffman, G. Grinstein,. K. Marx, I. Grosse, and E. Stanley, “Dna visual and analytic data mining,” IEEE Visualization, pp. 437-442 (1997)

[HiK98] A. Hinneburg and D. Keim, "An Efficient Approach to Clustering in Large Multimedia Databases with Noise”, Proceedings of KDD-98 (1998)

[HKK05] [J. Handl, J. Knowles and D. B. Kell, “Computational cluster validation in post genomicdata analysis”, Journal of Bioinformatics, Vol. 21(15), pp.3201-3212 (2005)

[HKW99] A. Hinneburg, D. A. Keim., M, Wawryniuk, “HD-Eye:Visual Mining of High-Dimensional Data”, IEEE Computer Graphics and Applications, Volume 19, Issue 5(September 1999), pp.22-31

[HKW03] A. Hinneburg, D. A. Keim., M, Wawryniuk, “HD-Eye-Visual Clustering of High dimensional Data”, Proceedings of the 19th International Conference on Data Engineering,pp.753-755 (2003)

[HoG01] Patrick E. Hoffman Georges G. Grinstein, “A survey of visualizations for multidimensional data mining”, Information visualization in data mining and knowledge discovery, Morgan Kaufmann Publishers Inc, pp. 47-82, 2001

[Hub85] P. J. Huber, “Projection Pursuit”, The Annals of Statistics, Vol. 13 (2), pp.435-474(1985)

[HuL00] Z. Huang and T. Lin, “A visual method of cluster validation with Fastmap”, Proceedings of PAKDD-2000, pp.153- 164 (2000)

[InD90] A. Inselberg, B. Dimsdale, “Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry”, Proceedings of Visualization ‘90, San Francisco, CA, pp. 361-370(1990)

[Ins85] A. Inselberg, “The Plane with Parallel Coordinates, Special Issue on Computational Geometry”, The Computer, Vol. 1, pp. 69-97 (1985)

[Ins97] A. Inselberg, “Multidimensional Detective”, Proceedings of IEEE Information Visualization '97 pp.100-107 (1997)

[JaD88] A. K. Jain and R. C. Dubes, “Algorithms for Clustering Data”, Prentice Hall Press (1988)

[Jac08] S. Jaccard, “Nouvellesrecherchessur la distribution florale”, Bull. Soc. Vaud. Sci.Nat., 44, pp.223-270 (1908)

[JMF99] A. Jain, M. N. Murty and P. J. Flynn, “Data Clustering: A Review”, ACM Computing Surveys, Vol. 31(3), pp. 264-323 (1999)

[Jol02] T. Ian Jolliffe. “Principal Component Analysis”, Springer Press (2002)[Kei01] D. A. Keim, “Visual exploration of large data sets,” ACM Communication, vol. 44 (8), pp.38–44, (2001)

[Kan00] E. Kandogan, “Star Coordinates: A Multi-dimensional Visualization Technique with Uniform”, Treatment of Dimensions”, IEEE Symposium on Information Visualization 2000.Salt Lake City, Utah. pp.4-8 (2000)

[KaR90] L. Kaufman and P. J. Rousseeuw, “Finding Groups in Data: an Introduction to Cluster Analysis”, John While & Sons. (1990)

[KDN+96] T. Kanungo, B. Dom, W. Niblack, and D. Steele, "A fast algorithm for mdl-based multi band image segmentation", in Image Technology, J. Sanz, Ed. Springer-Verlag, 1996.

[KeC00] M. K. Kerr and G. A. Churchill, "Bootstrapping Cluster Analysis: Assessing the Reliability of Conclusions from Microarray Experiments", Proceedings of the National Academy of Sciences (2000)

[Kei02] D. A. Keim, Information Visualization and Data Mining, IEEE Transactions on Visualization and Computer Graphics, Vol. 7(1), January-March 2002, pp.100-107 (2002)

[KeK94] D. A. Keim and H.-P. Kriegel, "VisDB: Database Exploration Using Multidimensional Visualization", IEEE Computer Graphics and Applications, 14(5) pp. 40-49 (1994)

[KHK99] G. Karypis, E.-H. S Han, and V. Kumar, "Chameleon: hierarchical clustering using dynamic modeling," IEEE Computer, vol. 32(8), pp.68–75 (1999)

[KrW78] J. B. Kruskal, M. Wish, "Multidimensional Scaling", SAGE university paper series on quantitive applications in the social sciences, Sage Publications, CA. pp. 07-011 (1978)

[KKA95] D.A. Keim, H.-P.Kriegel, M. Ankerst, "Recursive Pattern: A Technique for Visualizing Very Large Amounts of Data", Proceedings of Visualization '95, Atlanta, GA,pp. 279-286 (1995)

[KOC+04] S. Kabelac, S. Olbrich, K. Chmielewski, K. Meier, C. Holzknecht, "3DVisualization of Molecular Simulations in High-performance Parallel Computing Environments", Journal of Molecular Simulation, Volume 30(7), June 2004, Taylor and Francis Ltd. pp. 469-477 (2004)

[Koh97] T. Kohonen, "Self-Organizing Maps" Springer, Berlin, second extended edition (1997)

[KSP01] S. Kaski, J. Sinkkonen. and J. Peltonen, "Data Visualization and Analysis with Self-Organizing Maps in Learning Metrics", DaWaK 2001, LNCS 2114, pp.162-173 (2001)

[LeD01] E. Levine and E. Domany, "Resampling Method for Unsupervised Estimation of Cluster Validity", Neural Computation. 2001.

[Lev91] H. Levkowitz, "Color icons: Merging color and texture perception for integrated visualization of multiple parameters", Proceedings of the 2nd conference on Visualization '91,San Diego, CA, pp. 164-170 (1991)

[LKS+04] J. Lin, E. Keogh, S. Lonardi, J. Lankford and D. M. Nystrom, "Visually Mining and Monitoring Massive Time Series", KDD '04, August 22-25, 2004, Seattle, Washington, U.S.A (2004)

[Mac67] J. B. Mac Queen, "Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, University of California Press, pp.281-297 (1967)

[MaT03] K.-L. Ma, S. T. Teoh, “Star Class: Interactive Visual Classification Using Star Coordinates”, Proceedings of the 3rd SIAM International Conference on Data Mining, pp.178-185 (2003)

[MaW] The Math Works, Inc. textbook online, http://www.mathworks.com/[MiG04] James R Miller, E A Gustavo; "The Immersive Visualization Probe for Exploring n-Dimensional Spaces", Proceedings of IEEE Computer Graphics and Applications 2004, pp.76-85 (2004)

[MiI80] G. W. Milligan, and P. D. Isaac, “The validation of four ultra metric clustering algorithms”, Pattern Recognition, Vol. 12, pp.41-50 (1980)

[Mil81] G. W. Milligan, “A Monte Carlo study of thirty internal criterion measures for cluster analysis”, Psycho metrika, Vol. 46 (2), pp. 187-199 (1981)

[Mil96] G. W. Milligan, “Clustering validation: results and implications for applied analysis. in Clustering and Classification” ed. P. Arabie, L. J. Hubert and G. (1996)De Soete, World Scientific, pp.34 1-375.

[MRC02] A. Morrison, G. Ross and M. Chalmers, “Combining and comparing clustering and layout algorithms”, University of Glasgow (2002)

[MSS83] G.W. Milligan, L.M. Sokol, and S.C. Soon “The effect of cluster size, dimensionality and the number of clusters on recovery of true cluster structure”, IEEE Trans PAMI, Vol. 5(1), pp. 40-47 (1983)

[OlL03] F. Oliveira, H. Levkowitz, “From Visual Data Exploration to Visual Data Mining: A Survey”, IEEE Trans.Vis.Comput. Graph, Volume 9(3), pp.378-394 (2003)

[OlP07] J. V. de Oliveira and W. Pedrycz (Editor), Advances in Fuzzy Clustering and its Applications, Wiley, June (2007)[Pei] http://www.cs.sfu.ca/~jpei/

[PGW03] E. Pampalk, W. Goebl, and G. Widmer, “Visualizing Changes in the Structure of Data for Exploratory Feature Selection”, Proceedings of the ninth ACM SIGKD Dinternational conference on Knowledge discovery and data mining (SIGKDD ’03), August24-27, 2003, Washington, DC, USA pp.157-166 (2003)

[Pic70] R. M. Pickett, ”Visual Analyses of Texture in the Detection and Recognition of Objects”, in: Picture Processing and Psycho-Pictorics, Lipkin B. S., Rosenfeld A. (eds.),Academic Press, New York (1970)

[Ran71] W. M. Rand, “Objective Criteria for the Evaluation of Clustering Methods”, Journal of the American Statistical Association, Vol 66, pp. 846-850 (1971)

[Rei95] S. P. Reiss, “An Engine for the 3D Visualization of Program Information”, Journal of Visual Languages and Computing, Vol. 6, pp. 299-323 (1995)

[RBL+02] V. Roth, M. L. Braun, T. Lange and J. M. Buhmann "Stability-Based Model Order Selection in Clustering with Applications to Gene Expression Data", Lecture Notes In Computer Science; Vol. 2415, Proceedings of the International Conference on Artificial Neural Networks, pp.607-612 (2002)

[RKJ+99] W. Ribarsky, J. Katz, F. Jiang, A. Holland, “Discovery Visualization using Fast Clustering“, IEEE Computer Graphics and Applications, Vol. 19(5) 1999.

[RSE99] R.M. Rohrer, J.L. Sibert, D.S. Ebert, “Shape-based Visual Interface for Text Retrieval“, IEEE Computer Graphics and Applications, Vol. 19(5) 1999.

[SBG00] T.C. Sprenger, R. Brunella, M. H.Gross, “H-BLOB: a hierarchical visual clustering method using implicit surfaces”, Proceedings of Visualization 2000, pp. 61-68 (2000)

[SCZ98] G. Sheikholeslami, S. Chatterjee, A. Zhang, “Wavecluster: A multi-resolution clustering approach for very large spatial databases”, Proceedings of Very Large Databases Conference (VLDB98), pp.428-439 (1998)

[SDT+95] H. Su, H. Dawkes, L. Tweedie, R. Spence, ”An Interactive Visualization Tool for Tolerance Design”, Technical Report, Imperial College, London, (1995)

[SeS05] J. Seo and B. Shneiderman, “From Integrated Publication and Information Systemsto Virtual Information and Knowledge Environments”, Essays Dedicated to Erich J. Neuholdon the Occasion of His 65th Birthday. Lecture Notes in Computer Science Vol.3379, Springer (2005)

[SGF71] J.H. Siegel, R. M. Goldwyn and H. P. Friedman, “Irregular polygon to represent multivariate data (with vertices of equal intervals, distanced from the centre proportionally to the value of the variable)”, USA (1971) October

[Sha96] S. Sharma, “Applied multivariate techniques”, John Wiley & Sons, Inc. (1996)[Shn01] B. Shneiderman, “Inventing Discovery Tools: Combining Information Visualization with Data Mining, Proceedings of Discovery Science 2001,Lecture Notes in Computer Science Vol 2226, pp.17-28 (2001)

[Shn02] B. Shneiderman,, “Inventing discovery tools: Combining information visualization with data mining,” Information Visualization, Vol. 1, pp.5–12 (2002)

[ShY06] J. Shaik and M. Yeasin, “Visualization of High Dimensional Data using an Automated 3D Star Co-ordinate System”, Proceedings of International Joint Conference on Neural Networks, 2006 (IJCNN '06), 16-21 July 2006,Vancouver, Canada, IEEE Press,pp.1339-1346

[Sim93] P. K. Simpson, “Fuzzy min-max neural network—Part II: Clustering,”, IEEE Trans. Fuzzy Syst., Vol. 1(1), pp. 32–45 (1993)

[Smi86] W. A. Smith, Elementary Numerical Analysis, Prentice-Hall, (1986)

[Thk99] S. Theodoridis and K. Koutroubas, “Pattern Recognition”, Academic Press. 1999.

[VSA05] R.Vilalta, T. Stepinski, M. Achari, “An Efficient Approach to External Cluster Assessment with an Application to Martian Topography”, Technical Report, No. UH-CS-05-08, Department of Computer Science, University of Houston (2005)

[War95] M. Ward, “High dimensional brushing for interactive exploration of multi variatedata”, Proceedings of Visualization'95, pp.271-278 (1995.)

[WiL93] J. J van Wijk., R. D. van Liere, “Hyperslice”, Proceedings of Visualization ’93Conference, San Jose, CA, pp.119-125 (1993)

[WTP+95] J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, V. Crow, “Visualizing the Non-Visual: Spatial Analysis and Interaction with Information from Text Document”, Proceedings of Symposium on Information Visualization1995, Atlanta, GA,pp.51-58 (1995)

[WYM97] W. Wang, J. Yang, and R. Muntz, “STING: A statistical information grid approach to spatial data mining”, Proceedings of the 23rd International Conference on Very Large Data Bases (VLDB97), pp.186-195 (1997)

[XEK+98] X. Xu, M. Ester, H-P. Kriegel, and J. Sander, “A distribution-based clustering algorithm for mining in large spatial databases”, Proceedings of IEEE International Conference on Data Engineering, (ICDE 98), pp.324-331 (1998)

[XuW05] R. Xu and D. C. Wunsch, “Survey of Clustering Algorithms”, IEEE Transactions on Neural Networks, Vol. 16(3), May 2005, pp.645-678 (2005)

[Yan03] L. Yang, ”Visual Exploration of Large Relational Data Sets through 3D Projections and Footprint Splatting”, IEEE Transactions on Knowledge and Data Engineering, Vol.15(6), pp.1460-1471, November/December (2003)

[ZHW+03] X. Zheng, P. He, F. Wan, Z. Wang, G. Wu, “Dynamic Clustering Analysis of Documents Based on Cluster Centroids”, Proceedings of the Second International Conference on Machine Learning and Cybernetics, XiAn, 2-5 Nov. 2003, IEEE Press. Vol.1, pp.194-198(2003)

[ZOZ+06] K-B, Zhang, M. A. Orgun, K. Zhang and Y. Zhang, “Hypothesis Oriented Cluster Analysis in Data Mining by Visualization”, Proceedings of the working conference on Advanced visual interfaces 2006 (AVI06), May 23-26, 2006, Venezia, Italy. ACM Press, pp. 254-257 (2006)

[ZOZ06] K-B, Zhang, M. A. Orgun, K. Zhang, “HOV3: An Approach for Visual Cluster Analysis”, Proceedings of The 2nd International Conference on Advanced Data Mining and Applications. (ADMA 2006), Xi'an, China, August 14-16, 2006, Lecture Notes in Computer Science, Volume 4093 Springer Press, pp.316-327 (2006)

[ZOZ07a] K-B. Zhang, M. A. Orgun and K. Zhang, “A Visual Approach for External Cluster Validation”, Proceedings of the first IEEE Symposium on Computational Intelligence and Data Mining (CIDM2007), Honolulu, Hawaii, USA, April 1-5, 2007, IEEE Press. pp. 576-582 (2007)

[ZOZ07b] K-B. Zhang, M. A. Orgun and K. Zhang, “Enhanced Visual Separation of Clusters by M-mapping to Facilitate Cluster Analysis”, Proceedings of 9th International Conference series on Visual Information Systems (VISUAL 2007), June 28-29, 2007,Shanghai, China, Lecture Notes in Computer, Volume 4781, Springer Press, pp. 288-300(2007)

[ZOZ07c] K-B. Zhang, M. A. Orgun and K. Zhang, “A Prediction-based Visual Approach for Cluster Exploration and Cluster Validation by HOV3”, Proceedings of 18th European Conference on Machine Learning/11th European Conference on Principles and Practice of Knowledge Discovery in Databases (ECML/PKDD 2007), Warsaw, Poland, September 17-21, 2007, Lecture Notes in Computer, LNAI 4702 Springer Press, pp. 336–349 (2007)

[ZOZ07d] K-B. Zhang, M. A. Orgun and K. Zhang, “Predictive Hypothesis Oriented Cluster Analysis by Visualization”, Journal of Data Mining and Knowledge Discovery(2007) (submitted)

[ZRL96] T. Zhang, R. Ramakrishana and M. Livny, "An Efficient Data Clustering Method for Very Large Database", Proceedings of ACM SIGMOD International Conference on Management of Data, pp.103-114 (1996)