CHAPTER - II

LITERATURE SURVEY
CHAPTER-II

2.1 INTRODUCTION

Data Mining is the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. [3, 33]

2.2 LITERATURE REVIEW

Knowledge Discovery in Database (KDD) was formalized in 1989, with reference to the general concept of being broad and high level in the pursuit of seeking knowledge from data. The term data mining was then coined; this high-level application technique is used to present and analyze data for decision-makers. [3, 33]

Data Mining is only one of the many steps involved in Knowledge Discovery in Databases. The various steps in the knowledge discovery process include data selection, data cleaning and preprocessing, data transformation and reduction, Data Mining Algorithm selection and finally post-processing and interpretation of the discovered knowledge. The KDD process tends to be highly iterative and interactive. Data Mining Analysis tends to work up from the data and the best techniques are developed with an orientation towards large volumes of data, making use of as much data as possible to arrive at reliable conclusions and decisions. The analysis process starts with a set of data, and uses a methodology to develop an optimal representation of the structure of data, during which knowledge is acquired. Once knowledge is acquired, this can be extended to large sets of data on the assumption that the large data set has a structure similar to the simple data set. [33]

Knowledge Discovery in Databases is the process of identifying a valid, potentially useful and ultimately understandable structure in data. This process involves selecting a sampling data from a data warehouse, cleaning or preprocessing it, transforming or reducing it (if needed), applying a Data Mining component to produce a structure, and then evaluating the derived structure.
Data Mining is a step in the KDD process concerned with the algorithmic means by which patterns or structures are enumerated from the data under acceptable computational efficiency limitations.

Thus, the structures that are the outcome of the Data Mining process must meet certain conditions so that these can be considered to be Knowledge. These conditions are: validity, understandability, utility, novelty and being interesting.

DEFINITIONS

The term ‘Data Mining’ refers to the finding of relevant and useful information from databases. Data mining and knowledge discovery in the databases is a new interdisciplinary field, merging ideas from statistics, machine learning, databases and parallel computing. Researchers have defined the term ‘Data Mining’ in many ways.[3,33]

1. Data Mining or Knowledge Discovery in databases, as it is also known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes, and detecting anomalies.

2. Data Mining is the search for the relationships and global patterns that exist in large databases but are hidden among vast amounts of data, such as the relationship between patient data and their medical diagnosis. The relationship represents valuable knowledge about the database, and the objects in the database, if the database is a faithful mirror of the real world registered by the database.

3. Data Mining refers to using a variety of techniques to identify nuggets of information or decision-making knowledge in the database and extracting these in such a way that they can be put to use in areas such as decision support, prediction, forecasting and estimation. The data is often voluminous, but it has low value and no direct use can be made of it. It is the hidden information in the data that is useful.
4. Data Mining is the process of finding value from volume. In any enterprise, the amount of transactional data generated during its day-to-day operations is massive in volume. Although these transactions record every instance of an activity, it is of little use in decision making. Data Mining attempts to extract smaller pieces of valuable information from this massive database.

5. Discovering relations that connect variables in a database is the subject of Data Mining. The Data Mining system self-learns from the previous history of the investigated system, formulating and testing hypothesis about rules which system obey. When concise and valuable knowledge about the system of interest is discovered, it can and should be interpreted into some decisions support systems, which helps the manager to make wise and informed business decisions.

6. Data Mining is the process of discovering meaningful, new correlation patterns and trends by sifting through large amount of data stored in repositories, using pattern recognition techniques as well as statistical and mathematical techniques.

One important aspect of Data Mining is that it scans through a large volume of data to discover patterns and correlations between attributes. Thus, though there are techniques like clustering, decision trees, etc. existing in different disciplines, these are not readily applicable to Data Mining as they are not designed to handle large amounts of data. Thus, in order to apply statistical and mathematical tools, we have to modify these techniques to be able to efficiently sift through large amounts of data stored in the secondary memory.
Two kinds of Data Mining are Structured and Unstructured Data Mining.

**Unstructured Data Mining:**

- Text mining
- It involves the process of structuring the input text, deriving patterns within the structured data, and finally evaluating and interpreting the output.
- Image mining
- It is meant for extraction of low level color, shape and texture characteristics and their conversion into high-level semantic features using fuzzy production rules, derived with the help of an image mining technique.

**Structured Data Mining:**

Structured Data Mining is the process of finding and extracting useful information from raw datasets. It describes mining for general data structures some of them are

- Concept mining - It aims at extracting concepts from documents.
- Database mining
  - Relational Data Mining - It look for patterns among multiple tables
  - Database
  - Document warehouse - It says why things have happened, instead of what has happened as in data warehousing.
  - Data warehouse - It's a main repository of an organization's historical data.
- Graph Mining
- Graph Mining refers to extracting knowledge from massive graphs.
- Molecule Mining
- Analysis of XML documents, citation networks, CAD circuits, web logs, and web searches are done by means of Molecule Mining. It is also applied in drug discovery and compound synthesis.
- Web Mining - It is to discover patterns from the Web
- Sequence Mining
**Sequence Mining** is concerned with finding statistically relevant patterns between data examples where the values are delivered in a sequence. It is usually presumed that the values are discrete. Sequence mining is a special case of Structured Data Mining.

There are two different kinds of sequence mining: **String Mining** and **Item Set Mining**. String Mining is widely used in biology, to examine Gene and Protein sequences, and is primarily concerned with sequences with a single member at each position. There exist a variety of prominent algorithms to perform alignment of a query sequence with those existing in databases. Item set mining is used more often in marketing and CRM applications, and is concerned with multiple symbols at each position.

- Software Mining - It involves understanding existing software artifacts.

### 2.3 HOW DOES DATA MINING WORKS?

While large-scale information technology has been evolving separate transaction and analytical systems, Data Mining provides the link between the two. Data Mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks. Generally, any of the four types of relationships are sought: [3, 33, and 25]

- **Classes**: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.
- **Clusters**: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.
- **Associations**: Data can be mined to identify associations. The beer-diaper example is an example of Associative Mining.
• **Sequential patterns**: Data is mined to anticipate behavior patterns and trends.

Data Mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data using application software.
- Present the data in a useful format, such as a graph or table.

### 2.4 DATA MINING METHODOLOGY:

Data Mining Methodology follows a sequence of steps or processes before analyzing the data. It is also known as KDD process and has six different stages. [3, 33]

They are

- Data Selection
- Data Cleaning
- Data Enrichment
- Data Transformation or Coding
- Data Mining
- Data Visualization

**Data Selection:**

Data is selected from the operational databases and is copied to a data warehouse during this process or stage.

**Data Cleaning:**

It is a process that removes polluted data from that data warehouse. It uses techniques like de-duplication, domain consistency and disambiguation. De-duplication removes replicated data that has occurred due to typing mistakes. Domain consistency checks for invalid or null data and are removed using this technique. Disambiguation also finds errors in the data and removes it. It is an important process in data mining methodology, because the data used for mining should not have any polluted data.
Data Enrichment:

Data can be added to the data warehouse at any time. Adding new information to the existing database is called Data Enrichment.

Data Transformation:

It is called data coding that performs transformation on the databases to make the data for easier process. This stage transforms or converts data in one format into another format that is suitable for that task of Data Mining

2.5 DATA MINING TECHNIQUES

Researchers identify two fundamental goals of Data Mining: Prediction and Description. Prediction makes use of existing variables in the database in order to predict unknown or future values of interest, and description focuses on finding patterns describing the data and the subsequent presentation for user interpretation. The relative emphasis of both prediction and description differ with respect to the underlying application and the technique. There are several Data Mining techniques fulfilling these objectives. Some of these are associations, classifications, sequential patterns and clustering. The basic premise of an association is to find all associations, such that the presence of one set of items in a transaction implies the other items. Classification develops profiles of different groups. Sequential patterns identify sequential patterns subject to a user-specified minimum constraint. Clustering segments of database into subsets or clusters. Another approach of the study of DM techniques is to classify the techniques as User –guided or verification-driven Data Mining, and Discovery-driven or automatic discovery of rules [3, 23, 25,]
VERIFICATION MODEL

In this process of Data Mining, the user makes a hypothesis and tests the hypothesis on the data to verify its validity. The emphasis is on the user who is responsible for formulating the hypothesis and issuing the query on the data to affirm or negate the hypothesis.

DISCOVERY MODEL

The discovery model differs in its emphasis, in that it is the system that automatically discovers important information hidden in the data. The data is sifted in search of frequently occurring patterns, trends and generalizations about it, without intervention or guidance from the user.

The typical discovery driven tasks are

- Discovery of association rules
- Discovery of classification rules
- Clustering
- Discovery of frequent episodes
- Deviation detection [3,33]

DISCOVERY OF ASSOCIATION RULES

An association rule is an expression of the form $X \Rightarrow Y$, where $X$ and $Y$ are the sets of items. The intuitive meaning of such a rule is that the transaction of the database, which contains $X$, tends to contain $Y$. Given a database, the goal is to discover all the rules that have the support and confidence greater than or equal to the minimum support and confidence, respectively. [33]
CLUSTERING

Clustering is a method of grouping data into different groups, so that the data in each group share similar trends and patterns. Clustering constitutes a major class of Data Mining Algorithms. The goal of the process is to identify all sets of similar examples in the data, in some optimal fashion.

Clustering according to similarity is a concept which appears in many disciplines. If a measure of similarity is available, then there are a number of techniques for forming clusters. [3, 33]

The objectives of clustering are:

- To uncover natural groupings
- To initiate hypothesis about the data
- To find consistent and valid organization of the data.

DISCOVERY OF CLASSIFICATION RULES

Classification involves finding rules that partition the data into disjoint groups. The input for the classification is the training data set, whose class labels are already known. Classification analyzes the training data set and constructs a model based on the class label, and aims to assign a class label to the future unlabelled records. Since the class field is known, this type of classification is known as supervised learning. A set of classification rules are generated by such a classification process, which can be used to classify future data and develop a better understanding of each class in the data base. [33]

There are several classification discovery models. They are: decision trees, neutral networks, genetic algorithms and statistical models like linear geometric discriminates.
FREQUENT EPISODES

Frequent Episodes are the sequence of events that occur frequently, close to each other and are extracted from the time sequences. How close it has to be to consider it frequent, is domain dependent. This is given by the user as the input and the output are the prediction rules for the times sequences.

The applications include telecommunications, and share market analysis, and these are mainly used for temporal data.

DEVIATION DETECTION

Deviation detection is to identify outlying points in a particular data set, and explain whether they are due to noise or other impurities being present in the data or due to trivial reasons. It is usually applied with the database segmentation, and is the source of true discovery, since the outliers express deviation from some previously known expectation and norm. By calculating the values of measures of current data and comparing those with previous data as well as with the normative data can obtain the deviations. They can be applied in forecasting, fraud detection, customer retention, etc.

NEURAL NETWORKS

Neural Networks are a new paradigm in computing, which involves developing mathematical structures with the ability to learn. The methods are the result of academic attempts to model the nervous system learning. Neural networks have the remarkable ability to derive meaning from complicated or imprecise data and can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Neutral networks have broad applicability to real world business problems and have already been successfully applied in many industries. Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs. [3,33]

Neural Networks use a set of processing element analogues to neurons in the brain. These processing elements are inter connected in a network that can then identify patterns in data once it is exposed to the data, i.e., the networks learn from experience just as people do.
GENETIC ALGORITHMS

Genetic Algorithms are a relatively new computing paradigm, inspired by Darwin’s theory of evolution. A population of individuals, each representing a possible solution to a problem, is initially created at random. Then pairs of individuals combine (crossover) to produce off springs for the next generation. A mutation process is also used to randomly modify the genetic structure of some members of each new generation. The algorithm runs to generate solutions for successive generations. The probability of an individual reproducing is proportional to the goodness of the solution it represents. Hence, the quality of the solutions in successive generations improves. The process is terminated when an acceptable or optimum solution is found, or after some fixed time limit. Genetic algorithms are appropriate for problems, which require optimization, with respect to some computable criterion.[3,33]

SUPPORT VECTOR MACHINES

Support Vector Machines (SVM) is based on statistical learning theory and is increasingly becoming useful in Data Mining. The main idea is to non-linearly map the data set into a high dimensional feature space and use a linear discriminator to classify the data. Its success has been demonstrated in the areas of regression, classification and decision-tree construction. [33]

2.6 DATA MINING ALGORITHMS:

This is the discovery stage of KDD process and pattern and, hidden information is found at this stage. If any polluted data is found at this stage, the data is to be cleaned first then the above processes are repeated. Several Data Mining Algorithms are

- Associations Rules
- Clustering
- Decision trees
- K- Nearest Neighbor Algorithm
- Neural Networks
- Genetic Algorithms
2.7 DATA ANALYSIS METHODS:

Analyzed data are given as input for this. It is always better to display the analyzed data in a graphical manner after mining it, so that the end user can easily understand the visualized format of data rather than the data that is not visualized.

Different levels of analysis are available:

- **Artificial neural networks**: Non-linear predictive models that learn through training and resemble biological neural networks in structure.

- **Genetic algorithms**: Optimization techniques that use process such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.

- **Decision trees**: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that can be applied to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.

- **Nearest neighbor method**: A technique that classifies each record in a dataset based on a combination of the classes of the $k$ record(s) most similar to it in a historical dataset (where $k \geq 1$). Sometimes is called the $k$-nearest neighbor technique.
2.8 DATA VISUALIZATION TOOLS

Visualization is an important tool in large Data Mining applications. Patterns and trends hidden in large voluminous data can easily be explored using appropriate visualization and statistical tools. Data visualization is an effective technique to graphically display large volumes of data by converting raw data into meaningful images for effortless human comprehension or communication. Visualization software display data in various forms, and allow a user to manipulate the display.

Visualization Categories

The same data can be presented in many forms. Summarization is a special form of presentation in which summary information is arranged in tabular or graphical form.

Tables

Two-dimensional tables are organized as row and columns. The intersection of row and a column is called a cell or element. The values stored in a cell are known as cell values or table entity. Dimension of a table is the actual number of data cell in it, and is expressed in the form of (r x c). Where r is the number of rows and c is the number of columns, although row and column labels and caption are used to describe the table data.

Graphics

Graphics are pictorial representations of data to visually highlight facts and relationships. They are a step ahead of tabular presentations, and are the preferred choice in Data Mining. An advantage of graphics is that no detailed technical knowledge is needed to comprehend them. Computer generated graphics can also be linked to the parent data source that is used to generate it in such a way that when the parent data changes, the graphics are automatically updated. Thus dynamic data updates are instantaneously reflected in graphics. A good graph should have clarity, brevity, lack of distortion, and instant comprehension.
TYPES OF GRAPH

- Bar Charts
- Line Charts
- Histogram
- Pictogram
- Time Charts
- Temporal Histogram
- Spatial Histogram
- Pareto Diagrams
- Pie-Charts
- Radar Charts
- Frequency Polygons and Frequency Curves
- Stem-and-Leaf Plots
- Overlay Charts
- Scatter Plot
- Bubble Charts
- Contour Plots
- Hierarchical Charts
- Polar Trees
- Cause-and-Effect Diagrams
- Q-Q Plots
- Cher off Plots
- Box and Whisker Plots
- Stem Plots
2.9 APPLICATION AREAS OF DATAMINING

- Business intelligence
- Business performance management
- Discovery science
- Loyalty card
- Intelligence services
- Chem. informatics
  - Quantitative structure-activity relationship
  - Bioinformatics

Bioinformatics:

Computer science and biology fuse in the relatively new discipline of Bioinformatics. It is one of the most important and vast areas of Data Mining application. Bioinformatics represents a new field of scientific inquiry about life. The key goal of Bioinformatics is to create database systems and software platforms capable of storing and analyzing large set of biological data. The exponential growth of biological data sets, and the desire to share data for open scientific exchange, the bioinformatics community is continually exploring new options for data storage, exchange and representation. This interdisciplinary work is driven by the need to analyze and make sense out of the vast amount of data that is produced, when biological systems are studied. Science, especially biology produces vast, complex and noisy data of unseen proportion. The sequence of the whole human DNA poses a new challenge for data mining and computer science.

The technologies in data mining have been successfully applied to bioinformatics research in the past few years, but more research in this field is necessary. While tremendous progress has been made over the years, many of the fundamental challenges in bioinformatics are still open. Data mining plays an essential role in understanding the emerging problems in genomics, proteomics, and systems biology. Advanced Data Mining Technologies in Bioinformatics covers important research topics of Data Mining on bioinformatics.
2.10 RECENT RESEARCH ACHIEVEMENTS:

The opportunities today in Data Mining rest solidly on a variety of research achievements, the majority of which are the result of government sponsored research. In this section, a few of the more important ones are mentioned. Note that several of them are interdisciplinary in nature, resting on discoveries made by researches from different disciplines working together collaboratively.

Neural Networks: Neural networks are systems inspired by the human brain. A basic example is provided by a back propagation network which consists of input nodes, output nodes, and intermediate node also called hidden nodes. Initially, the nodes are connected with random weights. During the training, a gradient descent algorithm is used to adjust the weights so that the output nodes correctly classify data present to the input nodes. The algorithm was invented independently by several groups of researchers.

Graphical Models and Hierarchical Probabilistic Representations: A directed graph is a good means of organizing information about qualitative knowledge about conditional independence and casually gleamed from domain experts. Graphical models generalize Marko models and hidden Marko models, which have proved themselves to be a powerful modelling tool. Graphical models were independently invented by computational probability and artificial intelligence researchers studying uncertainty.

Ensemble Learning: Rather than using Data Mining to build a single predictive model, it is often better to build a collection or ensemble of models and to combine them, say with a simple, efficient voting strategy. This simple idea has now been applied in a wide variety of contexts and applications. In some circumstances, this technique is known to reduce variance of the predictions and therefore to decrease the overall error of the model.

Linear Algebra: Scaling Data Mining Algorithms often depends critically upon scaling underlying computations in linear algebra. Recent work in parallel algorithms for solving linear system and algorithms for solving sparse linear systems in high dimensions are important for a variety of data mining applications, ranging from text mining to detecting network instructions.
**Large Scale Optimization:** Some Data Mining Algorithms can be expressed as large-scale, often non-convex, optimization problems. Recent work has provided parallel and distributed methods for large-scale continuous and discrete optimization problems, including heuristic search methods for problems too large to be solved exactly.

**High Performance Computing and Communication:** Data Mining requires statistically intensive operation on large data sets. These types of computations would not be practical without the emergence of powerful SMP workstations and high performance clusters of workstations supporting protocols for high performance computing such as MPI and MPIO. Distributed data mining can require moving large amounts of data between geographically separated sites, something which is now possible with the emergence of wide area high performance networks.

**Databases, Data Warehouses, and Digital Libraries:** The most time consuming part of the data mining process is preparing data for data mining. This step can be stream-lined in part if the data is already in a database, data warehouse, or digital library, although mining data across different databases, for example, is still a challenge. Some algorithms, such as association algorithms, are closely connected to databases, while some of the primitive operations being built into tomorrow’s data warehouses should prove useful for some data mining applications.

**Visualization of Massive Data Sets:** Massive data sets, often generated by complex simulation programs, required graphical visualization methods for best comprehension. Recent advances in multi-scale visualization allow the rendering to be done far more quickly and in parallel, making these visualization tasks practical.
2.11 NEW APPLICATIONS

The discipline of data mining is driven in part by new applications which require new capabilities not currently being supplied by today’s technology. These new applications can be naturally divided into three broad categories.


b. Scientific, Engineering & Health Care Data. Scientific data and meta-data tend to be more complex in structure than business data. In addition, scientists and engineers are making increasing use of simulation and of systems with application domain knowledge.

c. Web Data: The data on the web is growing not only in volume but also in complexity. Web data now includes not only text and image, but also streaming data and numerical data.

In this section, I describe several such applications from each category.

- **Business Transactions:** Today, businesses are consolidating and more and more business has millions of customers and billions of their transactions. They need to understand risks and opportunities.

- **Electronic Commerce:** Not only does electronic commerce produce large data sets in which the analysis of marketing patterns and risk patterns is critical, but unlike some of the applications above, it is also important to do this real or near real time, in order to meet the demands of on-line transactions.

- **Genomic Data:** Genomic sequencing and mapping efforts have produced a number of databases which are accessible over the web. In addition, there are also a wide variety of other on-line data bases, including those containing information about diseases, cellular function, and drugs. Finding relationships between these data
sources, which are largely unexplored, is another fundamental data mining challenge. Recently, scalable techniques have been developed for comparing whole genomes.

- **Sensor Data**: Satellites, buoys, balloons and a variety of other sensors produce voluminous amounts of data about the earth’s atmosphere, oceans and lands. A fundamental challenge is to understand the relationships, including casual relationships amongst this data. For example, do industrial pollutants affect global warming? There are also large terabyte to petabyte data sets being produced by sensors and instruments in other disciplines, such as astronomy, high energy physics, and nuclear physics.

- **Simulation Data**: Simulation is now accepted as a third mode of science, supplementing theory and experiment. Today, not only do experiments produce huge data sets, but so do simulations. Data mining, and more generally data intensive computing, is proving to be a critical link between theory, simulations, and experiment.

- **Health Care Data**: Health care is the most rapidly growing segment of the nation’s GDP for some time. Hospitals, health care organizations, insurance companies, and the federal government have large collections of data about patients, their health care problems, the clinical procedures used, their costs, and the outcomes. Understanding relationships in this data is critical for a wide variety of problems, ranging from determining what procedures and clinical protocols are most effective to how best to deliver health care to the most people in an era of determining resources.

- **Multi-media Documents**: Few people are satisfied with today’s technology for retrieving documents on the web, yet the number of documents and the number of people accessing these documents is growing explosively. In addition, it is becoming easier and easier to archive multi-media data, including audio, images and video data, but harder and harder to extract meaning full information from the archives as the volume grows.
• **The Data Web:** Today the web is primarily oriented toward documents and their multimedia extensions. HTML has proved itself to be a simple, yet powerful language for supporting this. Tomorrow the potential exists for the web to prove equally important for working with data. The Extensible Markup Language (XML) is an emerging language for working with data in networked environments. As this infrastructure grows, data mining is expected to be a critical enabling technology for the emerging data web.

### 2.12 RESEARCH CHALLENGES IN DATA MINING

In this section, we describe some of the major research challenges identified by the three workshops. The research challenges are arranged into five broad areas: A) Improving the scalability of Data Mining Algorithms, B) Mining non-vector data, C) Mining distributed data, D) Improving the ease of use of Data Mining Systems and environments, and E) Privacy and Security issues for Data Mining.

**A. Scaling Data Mining Algorithms.** Most Data Mining Algorithms today assume that the data fits into memory. Although success on large data sets is often claimed, usually this is the result of sampling large data sets until they fit into memory. A fundamental challenge is to scale data mining algorithms as

1. The number of records or observations increases;
2. The number of attributes per observation increases;
3. The number of predictive models or rules sets used to analyze a collection of observations increases;
4. And, as the demand for interactivity and real-time response increases.

Parallel and out-of-memory versions of current Data Mining algorithms are developed, but genuinely new algorithms are required. For example, association algorithms today can analyze out-of-memory data with one or two passes, while only some auxiliary data is kept in memory.
B. Extending Data Mining Algorithms to new data types. Today, most Data Mining Algorithms work with vector values data. It is an important change to extend data mining algorithms to work with other data types, including 1) time series and process data, 2) unstructured data, such as text, 4) Semi-structured data, such as HTML and XML documents, 4) multi-media and collaborative data, 5) hierarchical and multi-scale data, and 6) collection-valued data.

C. Developing distributed Data Mining Algorithm. Today most Data Mining Algorithms require bringing all together data to be mined in a single, centralized data warehouse. A fundamental challenge is to develop distributed versions of Data Mining Algorithms so that Data Mining can be done while leaving some of the data in place. In addition, appropriate protocols, languages, and networks services are required for mining distributed data to handle the meta-data and mappings required for mining distributed data. As wireless and pervasive computing environments become more common, algorithms and systems for mining the data produced by these types of systems must also be developed.

D. Ease of Use. Data Mining today is at best a semi-automated process and perhaps destined to always remain so. On the other hand, a fundamental challenge is to develop data mining systems which are easier to use, even by casual users. Relevant techniques include improving user interface, supporting casual browsing and visualization of massive and distributed data sets, developing techniques and systems to manage the meta-data required for Data Mining, and developing appropriate languages and protocols for providing casual to data. In addition, the development of Data Mining and knowledge discovery environments which address the process of collecting, processing, mining, and visualizing data, as well as the collaborative and reporting aspects necessary when working with data and information derived from it, is another important fundamental challenge.
E. Privacy and Security: Data Mining can be a powerful means of extracting useful information from data. As more and more digital data becomes available, the potential for misuse of data mining grows. A fundamental challenge is to develop privacy and security models and protocols appropriate for Data Mining and to ensure that next generation Data Mining systems are designed from the ground up to employ these models and protocols.

2.13 DATA MINING SUCCESS STORIES

In this section, we briefly describe some success stories involving Data Mining and knowledge discovery.

- **Association Rules:** Suppose we have a collection of items. The data for many applications consists of multiple transactions, where each transaction consists of one or more items. A basic example is provided by a supermarket, where the items are the products offered for sale and the transactions are purchases, consisting of one or more products purchased by an individual at a given time. A fundamental problem is to uncover associations: which products tend to be purchased together. There has been a lot of recent work on this problem and a variety of algorithms have been developed which can discover associations, even in very large data sets, with just a few passes over the data. A variety of commercial data mining systems support association rules and they are now routinely applied to a range of problems from database marketing to product placement for supermarkets. In addition, association rules algorithms have spurred new research in a variety of areas from databases to complexity theory.

- **Fraud Detection:** Although relatively few credit card transactions are fraudulent, the sheer volume of transactions means that over $500m are lost each year in this way. A variety of Data Mining techniques have been used to develop fraud systems which can detect fraudulent credit card transactions in near-real time. This problem is challenging due to the size of the data sets, the rarity of the events of interest, and the
performance requirements for near-real time detection. Data Mining has also improved fraud detection in other application areas, including telecom fraud and insurance fraud.

- **Astronomical Data**: Traditionally, the search for new galaxies, stars, and quasars has primarily been done by astronomers visually examining individual photographic plates. Classification algorithms from data mining have recently been used to automate this process yielding new astronomical discoveries. The classification algorithms are applied to derived attributes produced by image processing, such as the brightness, area, and morphology of sky objects. The approach has also proved useful for detecting new objects too faint to be observed by a manual analysis or traditional computational techniques. For the 2nd Palomar Oberservatory Sky Survey, this approach resulted in over a three-fold increase in the size of the catalog.

- **Genomic Data**: Genomic data is stored all over the world, in a variety of formats and managed by a variety of applications and systems. Recently, systems have been developed which allow discoveries to be made involving information distributed over several systems. In particular, the new systems have enabled for the first time whole genome comparison, gene identification, and whole genome functional interpretation and analysis. The technique developed for analyzing genomic and other types of scientific data can be expected to play a role on analyzing a broad range of biological data.

- **Distributed Data Mining**: Traditionally, Data Mining has required that the relevant data be warehoused in a single location. Recently, distributed data mining systems have expected wide area, high performance next networks, such as the NSF v BNS network, to mine large amounts of distributed scientific and health care data. Recently, these systems have set records for the sustained movement of very large amounts of data over wide area networks. Separately, a prototype has been developed exploiting distributed data mining to improve the detection of credit card fraud.
• **Text Mining:** Recently, Data Mining has been combined with algorithms from information retrieval to improve the precision and recall of queries on very large collections of documents. In particular, some of these algorithms have proved useful on multi-lingual collections and others have shown their worth on querying using concepts instead of key words.

### 2.14 CONCLUSIONS

Data Mining and Knowledge Discovery are emerging as a new discipline with important applications to science, engineering, health care, education, and business. Data Mining rests firmly on 1) research advances obtained during the past two decades in a variety of areas and 2) more recent technological advances in computing, networking and sensors. Data Mining is driven by the explosion of digital data and the scarcity of scientists, engineers, and domain experts available to analyze it.

Data Mining is beginning to contribute research advances of its own, by providing scalable extensions and advances to work in associations, ensemble learning, graphical models, techniques for on-line discovery, and algorithms for the exploration of massive and distributed data sets.

An advance in Data Mining requires a) Supporting single investigators working in Data Mining and the underlying research domains supporting data mining. b) Supporting interdisciplinary and multi-disciplinary research groups working on important basic and applied Data Mining problems and c) Supporting the appropriate test beds for mining large, massive and distributed data sets.

Appropriate privacy and security models for Data Mining must be developed and implemented.