

# Survey on Data Mining

CHARUPALLI CHANDISH KUMAR REDDY, O.PRUDHVI, V. HARSHAVARDHAN

**Abstract**— This paper provides an introduction to the basic concept of data mining. Which gives overview of Data mining is used to extract meaningful information and to develop significant relationships among variables stored in large data set/data warehouse. In the case study reported in this paper, a data mining approach is applied to extract knowledge from a data set. Data mining is the process of discovering potentially useful, interesting, and previously unknown patterns from a large collection of data. Data mining is a multidisciplinary field, drawing work from areas including database technology, machine learning, statistics, pattern recognition, information retrieval, neural networks, knowledge-based systems, artificial intelligence, high-performance computing, and data visualization. We present techniques for the discovery of patterns hidden in large data sets, focusing on issues relating to their feasibility, usefulness, effectiveness, and scalability. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems.

**Index Terms**— Introduction , Data mining: Convergence of three technologies , Applications of Data mining , Future enhancement , conclusion

## 1 INTRODUCTION

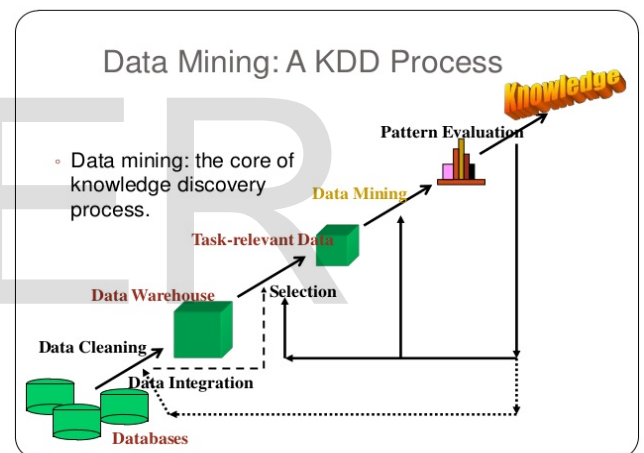
Data mining is a process to extract the implicit information and knowledge which is potentially useful and people do not know in advance, and this extraction is from the mass, incomplete, noisy, fuzzy and random data.

The essential difference between the data mining and the traditional data analysis (such as query, reporting and on-line application of analysis) is that the data mining is to mine information and discover knowledge on the premise of no clear assumption.

In addition to industry driven demand for standards and interoperability, professional and academic activity have also made considerable contributions to the evolution of the methods and models; an article published in a 2008 issue of the *International Journal of Information Technology and Decision Making* summarizes the results of a literature survey which traces and analyzes this evolution.

Data mining is the use of automated data analysis techniques to uncover previously undetected relationships among data items. Data mining often involves the analysis of data stored in a data warehouse. Three of the major data mining techniques are regression, classification and clustering.

Data Mining, also popularly known as *Knowledge Discovery in Databases* (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process. The following figure shows data mining as a step in an iterative knowledge discovery process.



**Fig.1: Data mining is the core of Knowledge Discovery Process**

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge.

- Charupalli Chandish Kumar Reddy is currently pursuing Master of Computer Applications in KMM Institute of PG studies in S.V University, Andhra pradesh, PH-8500433017. E-mail: charupallichandish@gmail.com
- O. Prudhvi is currently pursuing Master of Computer Applications in KMM Institute of PG studies in S.V University, Andhra pradesh, PH-9700665119. E-mail: prudhvionteri@gmail.com.
- Vemuri Harsha vardhan is currently working as Assistant Professor in KMM Institute of PG studies in S.V University, Andhra pradesh, PH-9959974091. E-mail: vemuriharsha@gmail.com

**The iterative process consists of the following steps:-**

**Data cleaning:** also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.

**Data integration:** at this stage, multiple data sources, often heterogeneous, may be combined in a common source.

**Data selection:** at this step, the data relevant to the analysis is decided on and retrieved from the data collection.

**Data transformation:** also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.


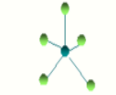



**Data mining:** it is the crucial step in which clever techniques are applied to extract patterns potentially useful.

**Pattern evaluation:** in this step, strictly interesting patterns representing knowledge are identified based on given measures.

**Knowledge representation:** is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

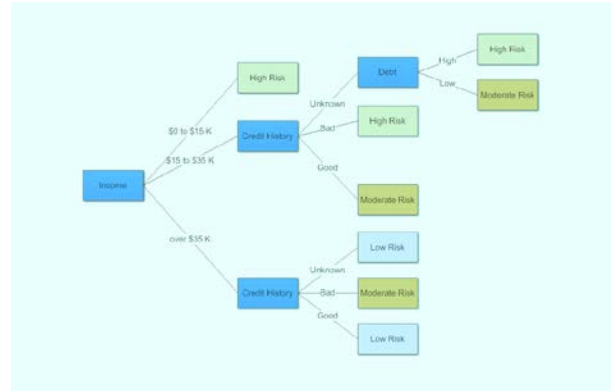
It is common to combine some of these steps together. For instance, *data cleaning* and *data integration* can be performed together as a pre-processing phase to generate a data warehouse. *Data selection* and *data transformation* can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

**Data Mining is.....**

-  • Decision Trees
-  • Nearest Neighbor Classification
-  • Neural Networks
-  • Rule Induction
-  • K-means Clustering

**Decision Trees:-**

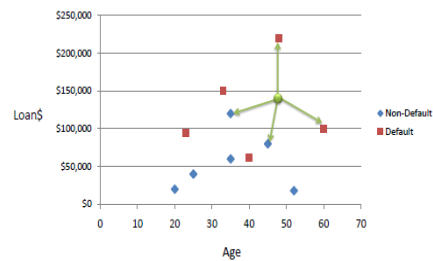
A **decision tree** is a **decision** support tool that uses a **tree-like** graph or model of **decisions** and their possible consequences, including chance event outcomes, resource costs, and utility. It is one way to display an algorithm.



**Fig.1: Decision Trees**

**Nearest Neighbor Classification:-**

KNN has been used in statistical estimation and **pattern recognition** already in the beginning of 1970's as a non-parametric technique. A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.



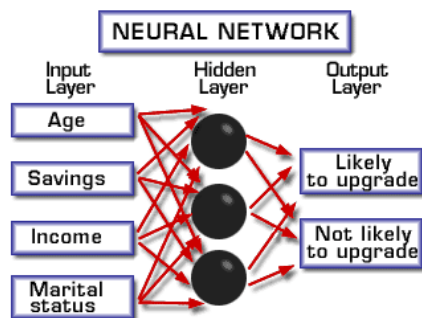
**Nearest Neighbor Classification**

**Neural Networks:-**

The backpropagation algorithm performs learning on a multi-layer feed-forward neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an input layer, one or more hidden layers, and an output layer.

Fundamental processing elements of a neural network is a neuron

1. Receives inputs from other source
2. Combines them in some way
3. Performs a generally nonlinear operation on the result
4. Outputs the final result



Neural Networks

**Rule Induction:-**

**Rule induction.** Rule induction is an area of machine learning in which formal rules are extracted from a set of observations. The rules extracted may represent a full scientific model of the data, or merely represent local patterns in the data.

**Rule Extraction from a Decision Tree**

- Rules are easier to understand than large trees
- One rule is created for each path from the root to a leaf
- Each attribute-value pair along a path forms a conjunction: the leaf holds the class prediction
- Rules are mutually exclusive and exhaustive
- Example: Rule extraction from our buys\_computer decision-tree



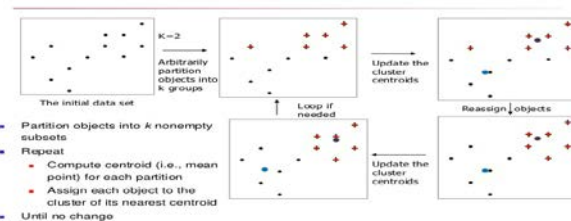
IF age = young AND student = no THEN buys\_computer = no  
 IF age = young AND student = yes THEN buys\_computer = yes  
 IF age = mid-age THEN buys\_computer = yes  
 IF age = old AND credit\_rating = excellent THEN buys\_computer = no  
 IF age = old AND credit\_rating = fair THEN buys\_computer = yes

**Rule Induction**

**K-means Clustering:-**

k-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics, and related fields.

**K-Means Clustering**



K-

**Means Clustering**

Data mining commonly involve's four classes of tasks:

**Clustering** - is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data. Clustering is a data mining (machine learning) technique used to place data elements into related groups without advance knowledge of the group definitions. Popular clustering techniques include k-means clustering and expectation maximization (EM) clustering.

**Classification** - is the task of generalizing known structure to apply to new data. For example, an email program might attempt to classify an email as legitimate or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines. Working with categorical data or a mixture of continuous numeric and categorical data? Classification analysis might suit your needs well. This technique is capable of processing a wider variety of data than regression and is growing popularity.

**Regression** - Attempts to find a function which models the data with the least error. Regression is the oldest and most well known statistical technique that the data mining community utilizes. Basically, regression takes a numerical dataset and develops a mathematical formula that fits the data. When you're ready to use the results to predict future behavior, you simply take your new data, plug it into the developed formula and you've got a prediction! The major limitation of this technique is that it only works well with continuous quantitative data (like weight, speed or age). If you're working with categorical data where order is not significant (like color, name or gender) you're better off choosing another technique. Regression is a data mining (machine learning) technique used to fit an equation to a dataset. The simplest form of regression, linear regression, uses the formula of a straight line ( $y = mx + b$ ) and determines the appropriate values for m and b to predict the value of y based upon a given value of x. Advanced techniques, such as multiple regression, allow the use of more than one input variable and allow for the fitting of more complex models, such as a quadratic equation.

**Association rule learning** - Searches for relationships between variables. For example a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing pur-

poses. This is sometimes referred to as market basket analysis.

## II. DATA MINING : CONVERGENCE OF THREE TECHNOLOGIES

### Convergence of Three Technologies

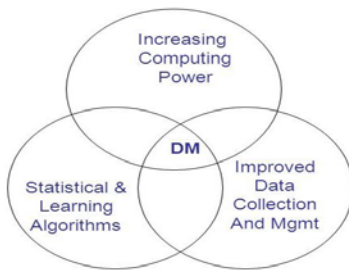


Fig.2: Convergence of three technologies

#### ❖ INCREASING COMPUTING POWER

Moore's law doubles computing power every 18 months

- Powerful workstations became common
- Cost effective servers (SMPs) provide parallel processing to the mass market
- Interesting tradeoff
- Small number of large analyses vs. large number of small analyses



#### ❖ Improved Data Collection

- Data Collection -> Access -> Navigation -> Mining
- The more data the better (usually)

#### ❖ Improved Algorithms

Techniques have often been waiting for computing technology to catch up

- Statisticians already doing "manual data mining"
- Good machine learning is just the intelligent application of statistical processes
- A lot of data mining research focused on tweaking existing techniques to get small percentage gains

## The Data Mining Process

Generally, data mining process is composed by data preparation, data mining, and information expression and analysis decision-making phases, the specific process.

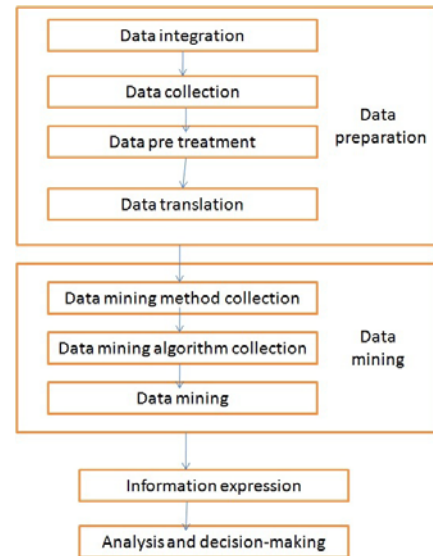


Fig.3: General process of Data Mining

### 1) Data preparation

Data preparation generally consists of two processes: data collection and data collation. Data collection is the first step of

data mining, and the data can come from the existing transaction processing systems, also can be obtained from the data warehouse; data collation is to eliminate noise or inconsistent data, it is the necessary link of data mining. The data obtained from the phase of the data collection may have a certain degree of "pollution", which refers to that in the data may be its own inconsistency, or some missing data, so the collation of the data is essential. At the same time, through data collation the data can be done on a simple generalization processing, thus on the basis of the original data more rich data information will be obtained, which will facilitate the next data mining step.

- **Data analysis** – The data is audited for errors and anomalies to be corrected. For large datasets, data preparation applications prove helpful in producing metadata and uncovering problems.
- **Creating an intuitive workflow** – A workflow consisting of a sequence of data prep operations for addressing the data errors is then formulated.

- **Validation** – The correctness of the workflow is next evaluated against a representative sample of the data-set. This process may call for adjustments to the workflow as previously undetected errors are found.
- **Transformation** – Once convinced of the effectiveness of the workflow, transformation may now be carried out, and the actual data prep process takes place.
- **Backflow of cleaned data** – Finally, steps must also be taken for the clean data to replace the original dirty data sources.

## 2) Data mining

Data mining is the core stage of the entire process, it mainly uses the collected mining tools and techniques to deal with the data, thus the rules, patterns and trends will be found.

## 3) Information expression

Information expression is to use visualization and knowledge information expression technology to provide the mined knowledge information for users, is an important means to show the data mining results. Clear and effective mining result information expression will greatly facilitate the accuracy and efficiency of the decision-making.

## 4) Analysis and decision-making

The ultimate goal of data mining is to assist the decision making. Decision-makers can analyze the results of data mining and adjust the decision-making strategies combining with the actual situation.

### Data mining architecture:

There are three tiers in the tight-coupling data mining architecture:

1. Data layer: as mentioned above, data layer can be database and/or data warehouse systems. This layer is an interface for all data sources. Data mining results are stored in data layer so it can be presented to end user in form of reports or other kind of visualization.
2. Data mining application layer is used to retrieve data from database. Some transformation routine can be performed here to transform data into desired format. Then data is processed using various data mining algorithms.
3. Front-end layer provides intuitive and friendly user interface for end-user to interact with data mining system. Data mining result presented in visualization form to the user in the front-end layer.

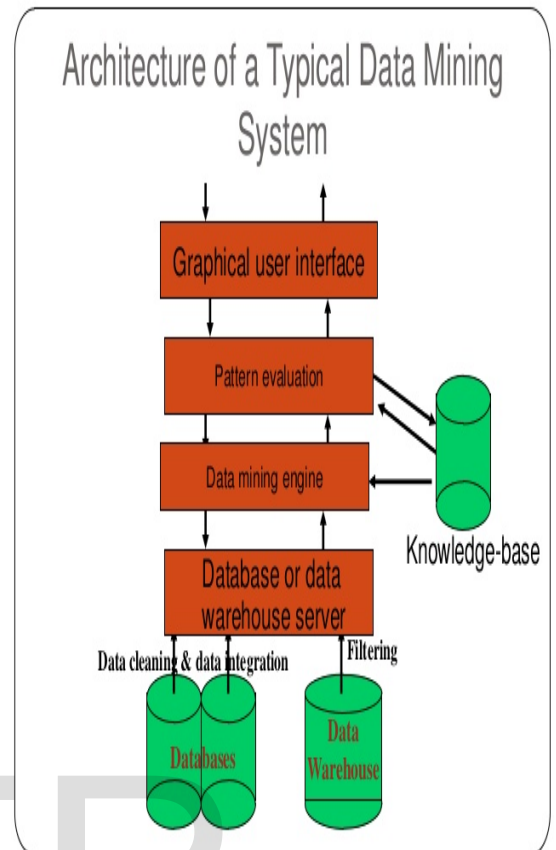


Fig.4: Architecture of Data Mining

In this article, we've discussed various *data mining architectures*, its advantages and disadvantages. And then we looked into a tight-couple **data mining architecture** – the most desired, high performance, high scalable data mining architecture.

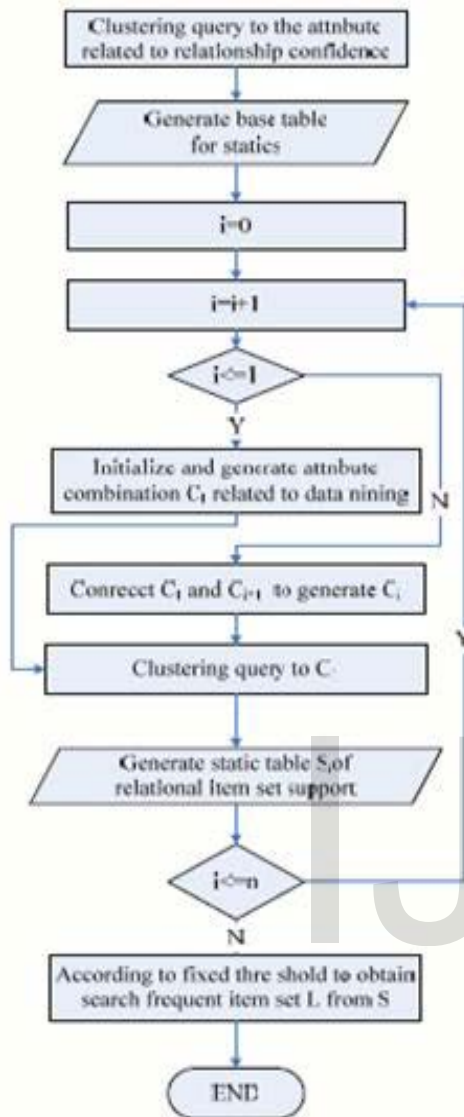
### Algorithm idea

The relation database contains complex multi-valued, multi-dimensional association rules, if analyzed from Boolean-based mining idea, the mining process is bound to be complex and cumbersome than the one-dimensional. Boolean rule in affair database; but if analyzed from the view of SQL-based operation technology, the mining algorithm of the association rules in relational database is more easily understood and realized. SQL language only needs to use its nine verbs to meet users' operation request on the database.

Each data mining algorithm can be decomposed into four components:

1. Model or pattern structure
2. Interestingness measure (score function)
3. Search method
4. Data management strategy

**Fig.5: Algorithm Process**



**Data mining based on decision tree**

**Decision tree learning**, used in statistics, data mining and machine learning, uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are **classification trees** or **regression trees**. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In decision analysis, a decision tree can be used to visually and explicitly represent decisions and decision making. In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. A decision support system (DSS) is a computer-based information system that supports business or organizational decision-making activities. DSSs serve the management, operations, and planning levels of an organiza-

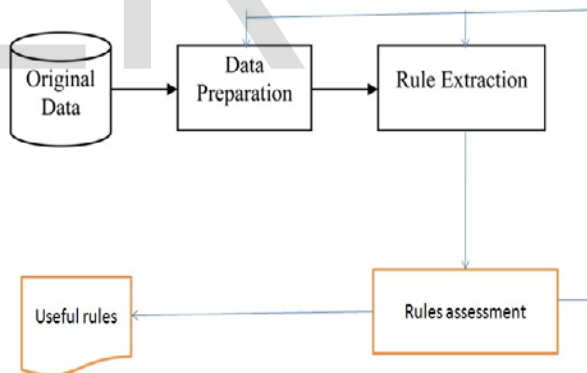
tion and help to make decisions, which may be rapidly changing and not easily specified in advance.

DSSs include knowledge-based systems. A properly designed DSS is an interactive software-based system intended to help decision makers compile useful information from a combination of raw data, documents, personal knowledge, or business models to identify and solve problems and make decisions.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation is when the data are accrued, possibly from various sources, and put together so that they can be analyzed. This is not data mining per se, but a result of the preparation of data before and for the purposes of the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when originally the data were anonymous.

**Data mining based on neural network:**

The data mining based on neural network is composed by data preparation, rules extracting and rules assessment three phases.



**Fig.6: Data mining process on neural network**

There are seven common methods and techniques of data mining which are the methods of statistical analysis, rough set, covering positive and rejecting inverse cases, formula found, fuzzy method, as well as visualization technology. Here, we focus on neural network method. Neural network method is used for classification, clustering, feature mining, prediction and pattern recognition. It imitates the neurons structure of animals, bases on the M-P model and Hebbien learning rule, so in essence it is a distributed matrix structure. Through training data mining, the neural network method gradually calculates (including repeated iteration or cumulative calculation) the weights the neural network connected.

**Data mining: K means clustering:**

K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

### The k-means Algorithm:

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

### III. APPLICATIONS OF DATA MINING

- Data Mining in Agriculture
- Surveillance / Mass surveillance
- National Security Agency
- Quantitative structure-activity relationship
- Customer analytics
- Police-enforced ANPR in the UK
- Stellar wind (code name)
- Educational Data Mining

#### Advantages of Data Mining

##### Marketing / Retail

Data mining helps marketing companies to build models based on historical data to predict who will respond to new marketing campaign such as direct mail, online marketing campaign and etc. Through this prediction, marketers can have appropriate approach to sell profitable products to targeted customers with high satisfaction.

Data mining brings a lot of benefits to retail company

in the same way as marketing. Through market basket analysis, the store can have an appropriate production arrangement in the way that customers can buy frequent buying products together with pleasant. In addition, it also help the retail company offers a certain discount for particular products what will attract customers.

##### Finance / Banking

Data mining gives financial institutions information about loan information and credit reporting. By building a model from previous customer's data with common characteristics, the bank and financial can estimate what are the good and/or bad loans and its risk level. In addition, data mining can help banks to detect fraudulent credit card transaction to help credit card's owner prevent their losses.

##### Manufacturing

By applying data mining in operational engineering data, manufacturers can detect faulty equipments and determine optimal control parameters. For example semi-conductor manufacturers had a challenge that even the conditions of manufacturing environments at different wafer production plants are similar, the quality of wafer are lot the same and some for unknown reasons even contain defects. Data mining has been applied to determine the ranges of control parameters that lead to the production of golden wafer. Then those optimal control parameters are used to manufacture wafers with desired quality.

##### Governments

Data mining helps government agency by digging and analyzing records of financial transaction to build patterns that can detect money laundering or criminal activity.

##### Disadvantages of data mining

##### Privacy Issues

The concerns about the personal privacy have been increasing enormously recently especially when internet is booming with social networks, e-commerce, forums, blogs.... Because of privacy issues, people are afraid of their personal information is collected and used in unethical way that potentially causing them a lot of trouble. Businesses collect information about their customers in many ways for understanding their purchasing behaviors trends. However businesses don't last forever, some days they may be acquired by other or gone. At this time the personal information they own probably is sold to other or leak.

##### Security issues

Security is a big issue. Businesses own information about their employee and customers including social security number,

birthday, payroll and etc. However how properly this information is taken is still in questions. There have been a lot of cases that hackers were accesses and stole big data of customers from big corporation such as Ford Motor Credit Company, Sony... with so much personal and financial information available, the credit card stolen and identity theft become a big problem.

### Misuse of information/inaccurate information

Information collected through data mining intended for marketing or ethical purposes can be misused. This information is exploited by unethical people or business to take benefit of vulnerable people or discriminate against a group of people. In addition, data mining technique is not perfectly accurate therefore if inaccurate information is used for decision-making will cause serious consequence.

#### Challenges of Data Mining

- Scalability
- Dimensionality
- Complex and Heterogeneous Data
- Data Quality
- Data Ownership and Distribution
- Privacy Preservation
- Streaming Data

### Marketplace surveys

Several researchers and organizations have conducted reviews of data mining tools and surveys of data miners. These identify some of the strengths and weaknesses of the software packages. They also provide an overview of the behaviors, preferences and views of data miners.

## IV. FUTURE ENHANCEMENT

Over recent years data mining has been establishing itself as one of the major disciplines in computer science with growing industrial impact. Undoubtedly, research in data mining will continue and even increase over coming decades involve Mining complex objects of arbitrary type, fast, transparent and structured data preprocessing, Increasing usability. All aim at understanding consumer behavior, forecasting product demand, managing and building the brand, tracking performance of customers or products in the market and driving incremental revenue from transforming data into information and information into knowledge.

Although data mining is still in its infancy, companies in a wide range of industries - including retail, finance, health care, manufacturing transportation, and aerospace - are already using data mining tools and techniques to take advantage of historical data. By using pattern recognition technologies and statistical and mathematical techniques to sift through warehoused information, data mining helps analysts recognize significant facts, relationships, trends, patterns, exceptions and anomalies that might otherwise go unnoticed.

## V. CONCLUSION

Data mining is a hot topic of the computer science research in recent years, and it has a extensive applications in various fields. Data mining technology is an application oriented technology. It not only is a simple search, query and transfer on the particular database, but also analyzes, integrates and reasons these data to guide the solution of practical problems and find the relation between events, and even to predict future activities through using the existing data.

Data mining brings a lot of benefits to businesses, society, governments as well as individual. However privacy, security and misuse of information are the big problem if it is not address correctly.

## REFERENCES

- [1] Ming-Syan Chen, Jiawei Han, Philip S yu. Data Mining: An Overview from a Database Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6):866-883.
- [2] R Agrawal ,T 1 mielinski, A Swami. Database Mining: A Performance Perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1993,12:914-925.
- [3] Fayyad, Usama; Gregory Piatetsky-Shapiro, and Padhraic Smyth (1996). "From Data Mining to Knowledge Discovery in Databases".<http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf> Retrieved 2008-12-17.
- [4] Han, J. & M. Kamber, Data mining: concepts and techniques, San Francisco: Morgan Kaufman (2001).
- [5] "Data mining tools", by Ralf Mikut, Markus Reischl, Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2011.
- [6] "Data mining and ware housing". Electronics Computer Technology (ICECT), 2011 3rd International Conference on Volume:1, Publication Year: 2011 , Page(s): 1 – 5.