International Journal of Advance Research in Science and Engineering
Vol. No.10, Issue No. 04, April 2021
www.ijarse.com
IJARSE
ISSN 2319 - 8354

REVISION OF MODELS AND FUNCTIONALITIES IN DATA MINING PROCESS

S.Sivasakthi
Assistant Professor in Computer Science Department,
G.Venkataswamy Naidu College, Kovilpatti
sivasakthi@gvncollege.edu.in

ABSTRACT:

Data Mining is defined as the system of extract information from enormous sets of data. In new words, we can say that data mining is mining knowledge from data. The information or knowledge extracted so can be used for Customer maintenance, Fraud recognition, and ProductionManagement, MarketStatus, and Science investigation. In Data Mining System there are a number of commercial system available today and however there are many challenges in this fieldwhich are implement through data warehouses and Online Analytical Processing along with different data mining models. In this paper I have paying attention on data mining works with revere to current research approach in mixture of fields.

Keywords: *Data Mining Models, Data Mining Functionalities, Future Trends of Data Mining, Data mining Applications, Challenges.*

I. INTRODUCTION

Data mining is the system of organizing through large amounts of data and selection out significant information. Data Mining is properlydistinct as the significant process of identifying valid, novel, potentially useful, and eventually understandable patterns in data.Mining of information is not the only process we must to process it; data mining also involves other processes such as Data Cleaning, Data Integration, Data Selection, Data Transformation, Data Mining, Pattern Evaluation and Data Presentation.

Data Cleaning:

Data Cleaning or Data Scrubbing is the process of amending or removing data that is incomplete and incorrect

Data Integration:

Data Integration is a **data** preprocessing technique that merges the **data** from various **data** sources into a consistent **data** store. **Data integration** may involve unpredictable **data** and therefore needs **data** cleaning.

Data selection:

Data Selection is the process where *data* relevant to the analysis task are retrieved from the databas*e.*

Data Transformation:

Data transformation is the process of converting data or information from one format to another, usually from the format of a source system into the required format of a new destination system.

Data Mining:

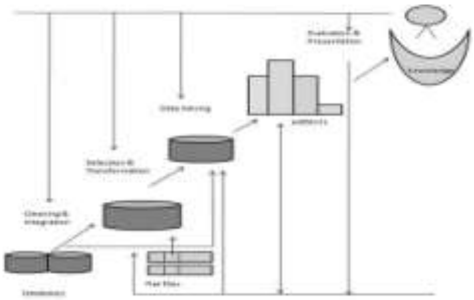
Afundamental process where intelligent methods are applied in order to extract data patterns.

Pattern Evaluation:

The process of evaluating the patterns signifying knowledge based on some interestingness measures.

Knowledge presentation:

Knowledge representation is the presentation of knowledge to the user for visualizationtechniques are used to present the mined knowledge.



The basic functionalities of data mining includes applying various methods and algorithms in order to preprocess the data,organize it, grouping and to discover useful patterns of stored data.

II. DATA MINING MODELS

In Data Mining there are two types of models are prescribed.

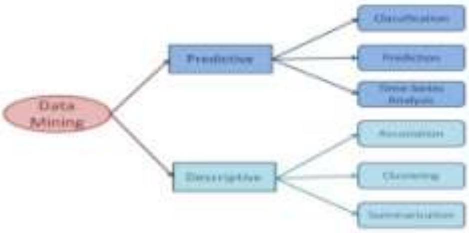
1. Predictive

2. **Descriptive.**

There are a number of data mining tasks

- classifications,
- Prediction,
- Time-Series Analysis,
- Association,
- Clustering,
- Summarization Etc.

All these tasks are either one predictive or descriptive.



1. **Predictive**

   Predictive Analytics, which use statistical models and estimation techniques to understand the future and the ability to "Predict" what might happen.Predictive analytics provide assessments about the likelihood of a prospect outcome. Most people are familiar with the use of predictive analytics to yield a credit score. These scores are used by economical services to define the possibility of customers making future credit payments on time. Usual business uses include, understanding how sales influence close at the end of the year, predicting what items customers will purchase together, or calculating inventory levels based upon a many of variables.

2. **Descriptive**

   Descriptive exploration does closely what the name suggests they "Describe", or review raw data and make it something that is interpretable by humans. Descriptive analytics are convenient because they allow us to learn from past behaviors, and understand how they might impactof future outcomes. Typically, the fundamental data is a count, or summative of a filtered column of data to which basic math is smeared. For all practical purposes, there are an infinite number of these statistics. Descriptive statistics are useful to show things like, total stock in portfolio, average amount spent per customer and turn over in sales. Common examples of descriptive are report that deliverhistoricperceptionsregarding the company's production, financial status, operations, sales, finance, inventory and customers.

III. DATA MINING FUNCTIONALITIES

Data mining functionalities are used to identify the kind of patterns to be found in data mining tasks.

Prediction − It is used to predict lost or unavailable numerical data values rather than class labels. Regression Analysis is commonly used for prediction. Prediction can also be used for identification of sharing trends based on available data.

Classification − It predicts the class of objects whose class label is unfamiliar. Its objective is to discover a derived model that describes and discriminates data classes or concepts. The Derived Model is based on the analysis set of training data i.e. the data object whose class label is well known.

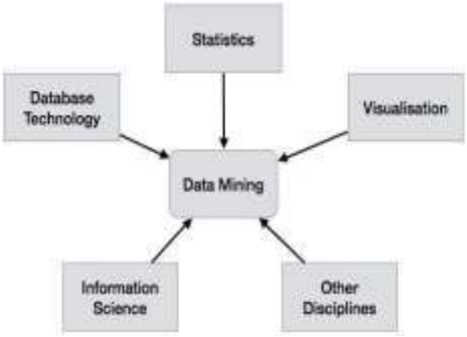
Clustering: Related to classification, clustering is the association of data in classes. In clustering, class labels are unknown and it is up to the clustering algorithm to discover suitable classes. Is also called*unsupervised classification*, because the classification is not read out by given class labels. There are many clustering methods all based on the standard of maximizing the resemblance between objects in a same class (*intra-class similarity*) and minimizing the similarity between objects of different classes (*inter-class similarity*).

Outlier Analysis − Outliers may be defined as the data objects that do not observe with the general performance or model of the data available.

Evolution Analysis − Evolution analysis discusses to the description and model predictabilities or trends for objects whose behavior changes over time.

IV. CLASSIFICATION OF DATA MINING SYSTEMS

A data mining system can be classified agreeing to the following criteria −



Classification of data mining systems according to the type of data sources Mined:

This classification is according to the type of data handled such as spatial data, multimedia data, time-series data, text data, World Wide Web, etc.

Classification of data mining systems according to the database involved:

This classification based on the data model involved such as relational database, object oriented database, data warehouse, transactional database, etc.

Classification of data mining systems according to the kind of knowledge discovered:

This classification based on the kind of knowledge discovered or data mining functionalities, such as characterization, discrimination, association, classification, clustering, etc. Some systems tend to be comprehensive systems offering several data mining functionalities together.

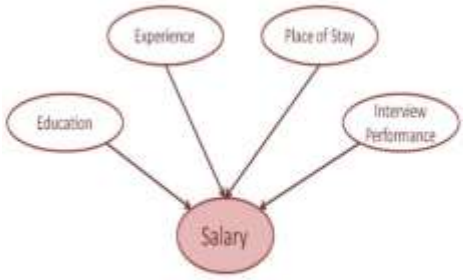
Classification of data mining systems according to mining techniques used:

This classification is according to the data analysis approach used such as machine learning, neural networks, genetic algorithms, statistics, visualization, database oriented or data warehouse-oriented, etc. The classification can also take into account the degree of user interaction involved in the data mining process such as query-driven systems, interactive exploratory systems, or autonomous systems.

V. ROOTS OF DATA MINING

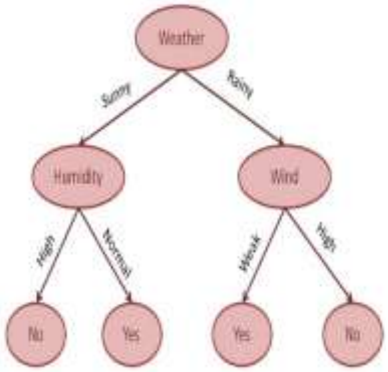
a) **Statistics**

Statistics are the foundation of most technologies on which data mining is built. Statistics hold concepts such as standard distribution, standard deviation,regression analysis, standard variance, discriminate analysis, confidence intervals and cluster analysis all of which are used to study data and data relationships. These are the very building blocks with which more traditional statistical analyses are supported. Today's data mining tools and techniques, classical statistical analysis plays a major role.



b)**Artificial Intelligence & Machine Learning**

AI is built upon heuristics as contrasting to statistics, and attempts to apply human-thought like treating to statistical problems. Because this approach requires vast computer processing power, it was not practical until the early 1980s, when computers began to offer useful power at sensible prices. AI found a few applications at stealing high end scientific/government markets, but the required supercomputers of the era priced AI out of the reach of effectively everyone else. Machine Learning could be considered as an evolution of AI, because it balances AI heuristics with advanced statistical methods. It let computer programs study about the data and then apply learned knowledge to data.



c) **Databases**

Huge amount of data needs to be stored in anorigin, and that too needs to be managed. So, comes in light the databases. Previously data was managed in records and fields, then in various models like network,hierarchicaletc. Relational model helped the needs of data storage for long while. Other advanced system that developed is object relational databases. But in data mining, capacity of data is too high, so we need specialized servers for it. We call the word as Data Warehousing. Data warehousing also supports OLAP operations to be applied on it, to support decision making.

d) Other Technologies

Apart from these, data mining trains various other areas e.g. visualization, pattern discovery, business intelligence etc. The table summarizes the evolution data mining on the grounds of development in databases.

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 Language (SQL), ODBC	Oracle, Sybase, Informix, IBM, Microsoft	Retrospective, dynamic data delivery at record level
Data Warehousing & Decision Support (1990s)	"What were unit sales in New England last March? Drill down to Boston."	On-line analytic processing (OLAP), multidimensional databases, data warehouses	Pilot, Comshare, Arbor, Cognos, Micro strategy	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 databases	Pilot, Lockheed, IBM, SGI, numerous startups (nascent industry)	Prospective, proactive information delivery

Table 1: Steps in Evolution of Data Mining

VI. HISTORICAL TRENDS OF DATA MINING

Data mining is useful in various disciplines, which includes database management systems (DBMS), Statistics, Artificial Intelligence (AI), and Machine Learning (ML). The era of data mining applications was conceived in the year1980 primarily by research-driven tools focused on single tasks. The early day's data mining trends are as under.

a) Data Trends

In earlier days, data mining algorithms work best for numerical data together from a single data base, and various data mining techniques have progressed for flat files, traditional and relational databases where the data is stored in tabular representation. Later on, with the confluence of Statistics and Machine Learning techniques, various algorithms evolved to mine the non-numerical data and relational databases.

b) Computing Trends

The field of data mining has been greatly inclined by the development of fourth generation programming languages and various related computing techniques. In, initial days of data mining most of the algorithms working only statistical techniques. Later on they developed with various computing techniques like AI, ML and Pattern Reorganization. Various data mining techniques (Induction, Compression and Approximation) and algorithms developed to mine the large volumes of various data stored in the data warehouses.

VII. CURRENT TRENDS & APPLICATIONS OF DATA MINING

A number of data mining applications have been applied in various fields like telecommunication, aviation, banking and finance, Astronomy climate, retail, health care, fraud detection, finance, telecommunication, and risk analysis...etc.. The ever growingcomplications in various fields and improvements in technology have modelled new challenges to data mining; the various challenges include different data formats, data from different locations, advances in computation and networking resources, research and scientific fields, ever growing business challenges etc. In data mining with various methods and techniques have designed the present data mining applications to switch the various challenges, the current trends of data mining applications are:

1. Fight against Terrorism

After 9-11 attacks, many countries forced new laws against struggling terrorism. These laws allow intelligence agencies to efficiently fight against terrorist organizations.USA launched Total Information Awareness program with the aim of creating a massive database of that secure all the information on population. Related projects were also launched in European countries and rest of the world. This program faced several problems,

a) The heterogeneity of database, the target database had to deal with text, audio, image and multimedia data.

b) Second problem was scalability of algorithms. The execution time increases as size of data (which is huge).

For example, 230 cameras were placed in London, to read number plates of vehicles. An estimated 40,000 vehicles pass camera every hour, in this way the camera must recognize 10 vehicles per second, which poses heavy loads on both hardware and software.

2. Web and Semantic Web

Web is the most recent and hottest trend now, but it is unstructured. Data mining is helping web to be structured, which is called Semantic web. The underlying technology is Resource Description Framework (RDF) which is used to describe resources. FOAF is also a supporting technology, heavily used in Face book and Orkut for cataloging. But still there are some issues like combining all RDF statements and dealing with invalid RDF statements. Data mining technologies are helping a lot to make the web, a semantic web.

3. Bio-informatics and Cure for Diseases

The second most important application trend, deals with mining and analysis of biological sequences and structures. Data mining tools are quickly being used in finding genes regarding cure of diseases like Cancer and AIDS.

4. Business Trends

Today’s business background is more dynamic, so businesses must be capable to react quicker, must be more beneficial, and offer high quality services that ever before. Here, data mining serves as a fundamental technology in enabling customer’s transactions more exactly faster and meaningfully. Data mining techniques of classification, regression, and cluster analysis are used for in current business trends. Almost all of the current business data mining applications are based on the classification and prediction techniques for supporting business decisions, thus creating strong Business Intelligence (BI) system.

APPLICATIONS

As data mining matures, new and progressively more innovative applications for it emerge. Although a wide range of data mining scenarios can be described. The applications of data mining are divided in the following categories:

- Banking &Finance
- Healthcare
- Telecommunication
- Retail industry
- Higher Education
- Text Mining & Web Mining

i) Banking &Finance

Most banks and financial institutions offer a wide diversity of services i.e. credit (such as business, mortgage, and automobile loans),banking services (such as checking, saving, and business and individual customer transactions), and investment services (such as mutual funds and also offer insurance services and stock services. Financial data collected in the banking and financial industry is often moderately complete, reliable and high quality, which facilitates systematic data analysis and data mining. For example it can also help in fraud detection by detecting a group of people who stage accidents to collect on insurance money.

ii) Healthcare

The past decade has seen an unstable growth in biomedical research, ranging from the development of new pharmaceuticals and in cancer therapies to the recognition and study of human genome by discovering large scale sequencing patterns and gene functions. Recent research in DNA analysis has led to the discovery of genetic causes for many diseases and disabilities as well as approaches for disease diagnosis, prevention and treatment.

iii) Telecommunication

The telecommunication industry has quickly evolved from offering local and long distance telephone services to provide many other comprehensive communication services including voice, fax, pager, cellular phone, images, e-mail, computer and web data transmission and other data traffic. The integration of telecommunication, computer network, Internet and numerous other means of communication and computing are underway. Moreover, with the deregulation of the telecommunication industry in many countries and the development of new computer and communication technologies, the telecommunication market is rapidly expanding and highly competitive. This creates a great demand from data mining in order to help understand business involved, identify telecommunication patterns, catch fraudulent activities, make better use of resources ,and improve the quality of service.

iv) Retail Industry

Retail industry collects huge amount of data on sales, goods transportation,and customer shopping history and consumption and service records and so on. The quantity of data collected continues to develop rapidly, particularly due to the increasing ease, availability and popularity of the business conducted on web, or e-commerce. Retail data mining can help recognize customer behavior, discover customer shopping patterns and trends, improve the quality of customer service, achieve better customer maintenance and satisfaction, enhance goods consumption ratios design more effective goods transportation and distribution policies and reduce the cost of business.

v) Higher Education

Higher education faces today is predicting paths of students and alumni. Which student will enroll in particular course programs? Who will need additional assistance in order to graduate? For now additional issues, enrolment management and time-to degree, continue to apply stress on colleges to search for new and faster solutions. Institutions can better address these students and alumni through the analysis and presentation of data.

Data mining has rapidly emerged as a highly popular tool for using current reporting capabilities to reveal and understand hidden patterns in huge databases.

vi) **Text Mining and Web Mining**

Text mining is the procedure of searching large volumes of documents from certain keywords or key phrases. By searching literally thousands of documents various relationships between the documents can be recognized. Using text mining we can easily derive certain patterns in the comments that may help classify a commonest of customer perceptions not captured by the other survey questions. Extension of text mining is web mining. Web mining is astimulating new field that integrates data and text mining within a website. It enhances the web site with intellectual behavior, such as suggesting related links or recommending new products to the consumer. Web mining is enables tasks that were previously difficult to implement. They can be configured to monitor and gather data from a wide variety of locations and can analyze the data across one or multiple sites. For example the search engines work on the standard of data mining.

VIII. FUTURE TRENDS IN DATA MINING

Businesses which have been slow in assuming the process of data mining are now gathering up with the others. Extracting important information through the process of data mining is generally used to make critical business verdicts. In the coming era, we can expect data mining to become as pervasive as some of the more widespread technologies used today. Some of the key data mining trends for the future include

1. **Multimedia Data Mining**

   This is one of the most recent methods which is catching up because of the increasing ability to capture useful data accurately. It contains the extraction of data from different kinds of multimedia sources such as text, audio,hypertext, video, images, etc. and the data is converted into a numerical representation in altered formats. This method can be used in classificationsclustering and performing similarity checks, and also to identify associations.

2. **Ubiquitous Data Mining**

   This method involves the mining of data from portable devices to get information about individuals. In spite of having several challenges in this type such as privacy, cost, complexity, etc. this method has a lot of chances to be huge in various businesses especially in studying human-computer interactions.

3. **Distributed Data Mining**

   This method is gaining reputation as it involves the mining of huge amount of information stored in different company locations or at different establishments. Highly sophisticated algorithms are used to extract data from different locations and deliver proper visions and reports based upon them.

4. **Spatial and Geographic Data Mining**

   This type of data mining which includes astronomical, extracting information from environmental, and geographical data which also includes images taken from outer space. This type of data mining can expose various facets such as distance and topology which is mainly used in geographic information systems and other navigation applications.

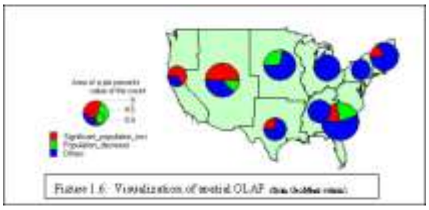


Figure 1.6: Visualization of spatial OLAP

5. **Time Series and Sequence Data Mining**

This type of data mining is study of cyclical and seasonal trends. This practice is also helpful in examining even random events which happen outside the normal series of events. This method is mainly being use by retail companies to access customer's buying forms and their behaviors.

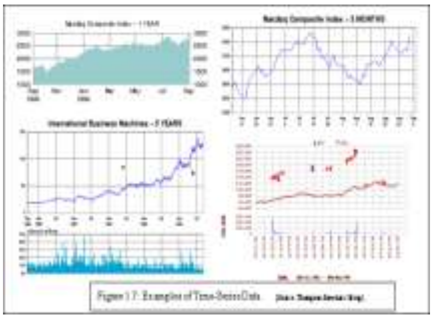


Figure 1.7: Examples of Time-Series Data

VIII. NEW TECHNIQUES IN DATA MINING

Not only do the application areas of data mining expand continuously, but also the utilized techniques keep up improving. In the rest of this article we take a closer look at four new methods:

- Multi-relational data mining,
- Support vector machines,
- Bayesian networks
- Ensemble methods.

Multi-relational data mining

Most data mining algorithms are propositional; this means that they were created to determine patterns in a single data table. However, larger databases generally contain several tables between which a number of associations have been defined. Propositional algorithms create classification rules of the following form:

IF (income > 108000) THEN important customer = YES

Notice that only the information from the first table was used for the construction of this rule. Relational algorithms on the other hand are able use the relationships that occur between the tables. An example of such a rule is:

IF (x is married with a person with income > 10800) THEN
important customer (x) = YES

Relational Database with two tables

Customer Table

ID	Gender	Age	Income	Expense	Important Customer
C1	Male	30	214000	18800	Yes
C2	Female	19	139000	15100	Yes

C3	Male	55	50000	8600	No
C4	Female	48	26000	8600	No

Married With Tablel

Partner 1	Partner 2
C1	C2
C3	C4

Support Vector Machines

Classification and regression are possibly the well-known applications of data mining. A multitude of techniques have been proposed for solving these responsibilities. Linear least-squares regression, discriminant analysis,decision trees and neural networks are only a few of them. When challenged with a classification or regression problem, the data mining practitioner must often make a trade-off between the intelligibility and presentation of the available techniques. Classifications by decision trees on the other are obviously motivated by a number of rules that are denoted by the tree.

Techniques that are capable to give the motivation behind their decisions are called white-box classifiers. In present years, lots of research has been accomplished to convert results from the difficult to understand black-boxes into white-boxes. Recently, a new black-box technique has been suggested that shows even better performance: support vector machines.

The basic idea behind SVMs is the following: the data is first being drawn into a high-dimensional space and subsequently a linear classifier is constructed in this high- dimensional space. The resulting models can be represented as constrained optimization problems which give anexclusive solution.

Bayesian Networks

Moreover support vector machines, we perceiveagrowing application of Bayesian networks within industrial applications. A Bayesian network is a graphical model where variables are presented by nodes and the edges between two nodes represent the requirements between the variables.

Ensemble Methods

The fundamentalnotion behind ensemble methods is very simple: several classifiers are qualified on the data and consequently these specificpredictions are combined to achieve one general forecast. The first classifier bases itself mainly on the language, the second classifier bases itself on the colors. If both classifiers are combined, we obtain a classifier that makes better predictions than each of the individual classifiers distinctly. Some ensemble methods are: bagging, boosting and stacking.

Bagging, the abbreviation of bootstrap accumulating, n random subsets from the original data are selected and a classifier is created for each of these subsets. A new surveillance is then classified by combining the forecasts of the n classifiers.

Boosting is same as bagging, but it selecting entirely random subsets, weights are given to the training observations. Sometimes Observations are misclassified to receive a larger weight and their chance of being incorporated in the subset will increase because of this. Boosting therefore gives more consideration to those observations which are tougher to predict.

Stacking, relates to several classifiers are trained on the available data and their forecasts are used as inputs for a so-called meta-learner. This meta-learner uses these individual forecasts to obtain one overall forecast.

X. CHALLENGES IN DATA MINING

Nowadays data mining research is “too”ad-hoc” and there are so many challenges to merge different data mining tasks. Some of the challenges in the area are as under:

Scalability:

The important challenge is mining data from large data bases. Computer data network and satellite data can easily be of this balance but to-days technology in data mining are too slow to switch data of this scale. If data mining algorithms are efficient adequate to control these huge data sets then they must be scalable. Future data mining should be a continuous, online process as an alternative of one time little process. The said scalability also warrants the implementation of novel data structure to access individual records in a soft manner.

Complex and Heterogeneous Data

Another challenge is appearance of more data complexity. A good system must extent the complexity from users. Previous analysis data mining method deal with the data set consisting attribute of similar type i.e. continuous or categorical due to increasing role of data mining in different areas, a need is arise to develop procedures which can handle mixed attributes. Such developed techniques for mining such complex objects ought to have taken care the relationships in data, like sequential and spatial auto-correction, graph connectivity and parent-child relationships between the components in semi-structures text and XML documents.

High Dimensional Data and High Data Streams

One challenge is to design classifiers to manage ultra-high dimensional classification problems for mining huge, massive and high dimensional data set out-of-memory, equivalent and distributed algorithms, algorithm is need to be developed. The conventional data analysis techniques developed for low-dimensional data do not work for dimensional data.

Data Ownership, Security and Privacy

It is a big challenge to discover data for an analysis at one location or to be owned by one location or to be owned by one unit. An automatic data mining in distributed atmosphere can develop serious issues in terms of data privacy or its security. These issues can be addressed by developing of acompetent algorithms and data structures to evaluate the knowledge reliability of a collection of data and further to measure the impact on the variation of data values on discovered pattern’s statistical importance.

Data Distribution

This challenge in data mining is very significant in linkage problems. This can be addressed by the development of distributed data mining techniques. The key challenges in distributed data mining are:

a) To reduce the amount of communication needed to implement the distributed computation.
b) To combine the data mining results obtains from multiple sources in aneffective manner.

c) To attack data security issues.

XI. CONCLUSION

It is difficult to imagine our civilization today without data mining. Both in scientific and industrial world, the applications have become too pervasive. In this paper, a short review was given about new domains in which data mining can cause huge changes. There are still many problems to overcome, from which privacy defense draws most responsiveness. And also reviewed some data mining trends and applications from its initiation to the future. This review puts attention on the hot and hopeful areas of data mining. Data mining is becoming increasingly shared in both the private and public sectors. Industries such as, insurance, medicine, and retailing commonly use data mining to reduce costs,banking, enhance research, and increase sales. So, data mining will be further and more useful in future.

REFERENCES

[1] Data Mining and Its Current Research Directions Amit Kumar Patnaik Pre Final Year B.Tech (CSE) Gandhi Institute for Technological Advancement, Dept. of CSE, Bhubaneswar, Odisha, India

[2] Data Mining Trend In Past, Current And Future ,SangeetaGoele, NishaChanana Research Scholar, University School of Management, Kurukshetra University, Kurukshetra

[3] Data Mining: Future Trends and Applications, Annan Naidu PaidiAsst.Prof of CSE Centurion University, Odisha, India.

[4] New Trends in Data Mining by J. HUYSMANS, B. BAESENS, D. MARTENS, K. DENYS and J. VANTHIENEN

[5] Ling Chen, MingqiLv, Qian Ye, Gencai Chen, John Woodward, A personal route prediction system based on trajectory data mining, Information Sciences, Volume 181, Issue 7, 1 April 2011, Pages 1264-1284.

[6] Jing He.2009. Advances in Data Mining: History and Future, Third international Symposium on Information Technology.

[7] Han, J. and M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann, 2001.

[8] Introduction to Data Mining and Knowledge Discovery, Third Edition ISBN: 1-892095-02-5, Dunham, M. H., Sridhar, S., “Data Mining: Introductory and Advanced Topics”, Pearson Education, New Delhi, 1st Edition, 2006, ISBN: 81-7758-785-4.