# Data Mining

Author: Pushpendar Sharma, Rajesh Sharma

Co-Author: Dr. Priya Swaminarayan,

Department of MCA, Parul University, Vadodara, India.

*Abstract:* Data mining may be a field of intersection of computing and statistics wont to discover patterns within the information bank. The main aim of the info mining process is to extract the useful information from the dossier of knowledge and mold it into a clear structure for future use. There are different processes and techniques used to carry out data mining successfully.

In terms of knowledge processing, classical statistical models are restrictive; it requires hypotheses, the knowledge and knowledge of specialists, equations, effective knowledge of probabilities distribution and therefore the data must have a top quality, being subject to prior processing and transformations. Because of these disadvantages the concept of knowledge mining has emerged, implementing knowledge extraction algorithms from the massive data collections. Data mining may be a process consisting in collecting knowledge from databases or data warehouses and therefore the information collected that had never been known before, it's valid and operational. Nowadays data mining is a modern and powerful IT&C tool, automatizing the process of discovering relationships and combinations in raw data and using the results in an automatic decision support.

Keywords: Data Mining, Data Mining Algorithms, Data Mining Application Area, Data mining methodology, Data mining Future Technology, Data mining tools, weka data mining tool.

## I. INTRODUCTION

In this information era, a huge amount of data is collected daily. Analyzing that huge amount of data and extract meaningful information from that data is a necessity to achieve goals. Now we are living in the world where a lot of data (scientific data, medical data, banking data, marketing data & Financial data etc., related to different fields are available but nobody have time to retrieve meaningful information from this data manually. To retrieve this information in easy way, we find shortcut methods to automatically classify it, to automatically summarize it, to automatically discover & characterize trends in it. Data Mining discover large datasets to dig out the unknown and earlier weird pattern, relationships and knowledge that are not easy to detect with the algorithms & traditional statistical methods. Data mining has effectively been used in many fields such as marketing, banking, medical, business, fraud detection, weather forecasting etc. Data Mining or KDD(knowledge discovery in database) is the process to find the helpful knowledge from a collection of data. This is mostly used data mining technique in this process that includes data preparation and selection, data cleansing, incorporating earlier knowledge on data sets and interpreting perfect solutions from the pragmatic results.

## II. APPLICATION AREAS OF DATA MINING

### 1) FUTURE HEALTHCARE

Data mining holds great potential to enhance health systems. It uses data and analytics to spot best practices that improve care and reduce costs. Researchers use data processing approaches like multi-dimensional databases, machine learning, soft computing, data visualization and statistics. Mining are often wont to predict the quantity of patients in every category. Processes are developed that confirm that the patients receive appropriate care at the proper place and at the proper time. data processing also can help healthcare insurers to detect fraud and abuse.

### 2) EDUCATION

There is a replacement emerging field, called Educational data processing, concerns with developing methods that discover knowledge from data originating from educational Environments. The goals of EDM are identified as predicting students' future learning behavior, studying the consequences of educational support, and advancing knowledge domain about learning. data processing are often employed by an establishment to require accurate decisions and also to predict the results of the scholar . With the results the institution can specialize in what to show and the way to show. Learning pattern of the scholars are often captured and wont to develop techniques to show them.

### 3) CRM

Customer Relationship Management is all about acquiring and retaining customers, also improving customers' loyalty and implementing customer focused strategies. To maintain a correct relationship with a customer a business got to collect data and analyze the knowledge. This is where data mining plays its part. With data processing technologies the collected data are often used for analysis. Instead of rambling where to focus to retain customer, the seekers for the answer get filtered results.

### 4) FRAUD DETECTION

Billions of dollars are lost to the action of frauds. Traditional methods of fraud detection are time consuming and sophisticated . Data mining aids in providing meaningful patterns and turning data into information. Any information that's valid and useful is knowledge. A perfect fraud detection system should protect information of all the users. A supervised method includes collection of sample records. These records are classified fraudulent or non-fraudulent. A model is formed using this data and therefore the algorithm is made to spot whether the record is fraudulent or not.

## III. ALGORITHM

1) Page rank algorithm
2) AdaBoost

## IV. TOOLS AND TECHNOLOGY

### 1) Orange

Orange is an open source data mining tool written in python language. It is a component based & machine learning tool which is used for data visualization. In this tool data mining can be done through visual programming & python scripting.

### 2) Rapid Miner

Rapid Miner is written in the Java Programming language, this tool offers advanced analytics through template-based frameworks. Rapid Miner also provides functionality like data preprocessing, visualization, predictive analytics, statistical modeling, evaluation and deployment. Rapid Miner is used in business. Industrial application, research, education, etc,.

### 3) Weka Tool

The original non-Java version of WEKA primarily was developed for analyzing data from the agricultural domain. With the Java-based version, the tool is very sophisticated and used in many different applications including visualization and algorithms for data analysis and predictive modelling. Its free under the GNU General Public License, which is a big plus compared to Rapid Miner, because users can customize it however they please. WEKA supports several standard data mining tasks, including data pre-processing, clustering, classification, regression, visualization and feature selection. WEKA would be more powerful with the addition of sequence modelling, which currently is not included.

## V. FUTURE TECHNOLOGIES

### 1) Multimedia Data Mining

This is one among the newest methods which is catching up due to the growing ability to capture useful data accurately. It involves the extraction of data from differing types of multimedia sources like audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in several formats. These methods are often utilized in clustering and classifications, performing similarity checks, and also to spot associations.

### 2) Ubiquitous Data Mining

This method involves the mining of data from mobile devices to urge information about individuals. In spite of getting several challenges during this sort like complexity, privacy, cost, etc. this method features a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.

### 3) Distributed Data Mining

This type of knowledge mining is gaining popularity because it involves the mining of giant amount of data stored in several company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.

### 4) Spatial and Geographic Data Mining

This method involves the mining of data from mobile devices to urge information about individuals. In spite of getting several challenges during this sort like complexity, privacy, cost, etc. this method features a lot of opportunities to be enormous in various industries especially in studying human-computer interactions.