International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 6 Issue: 7 124 - 127

# Data Mining Concepts
## A survey paper

Parul Mahawar
Dept. of Computer Science
Arya college of Engineering and IT, RTU Jaipur, *Rajasthan*
*parul.18m@gmail.com*

Vishal Shrivastava
Professor, Dept. of Computer Science Arya College of Engineering and IT, RTU Jaipur, Rajasthan
*Vishal500371@yahoo.co.in*

**Abstract**—Data Mining is a significant field in today's data-driven world. Understanding and implementing its concepts can lead to discovery of useful insights. This paper discusses the main concepts of data mining, focusing on two main concepts namely Association Rule Mining and Time Series Analysis

***Keywords***- *Data Mining, Association Rule Mining, Time Series Analysis*

*****

## I. DATA MINING

### *A. The idea behind Data Mining*

The world is constantly experiencing an exponential data growth. Sensor data, social media streams , images, video and mobile data have contributed a lot to this massive increase. However, all this contributes to unstructured data. Proper mining of this unstructured data is important because insights, patterns and concepts are deep buried in this human language communication data. Good Mechanisms are needed to locate, extract, organize and store this data. Here is whereData mining comes into play.

Various definitions of Data Mining have been proposed like "The efficient discovery of previously unknown pattern in large databases" [1] or "the non-trivial extraction of implicit, previously unknown and potentially useful information from data in database" [2].

Data Mining is thus a process to extract useful information from large amount of data. This data can be stored in database, data warehouse or any other information repository. It is also referred as knowledge mining from database, data analysis, data archaeology and knowledge extraction.

### *B. Data Mining System Architecture*

There are various steps involved in extracting knowledge from large data sets.All these steps form the data mining system architecture and are explained below as: -

- **Data sources**- Data warehouse,databases,text files and world wide web are the key data sources. This data as coming from multiple sources so is in different formats. The necessary data is cleaned, integrated and passed to the server.
- **Database or Data Warehouse Server**- It contains the processed data from various data sources. The job of this server is to retrieve the relevant data based on data mining request of the user
- **Data Mining Engine**- It is the core component of Data Mining System It contains modules which perform mining task like clustering, classification, prediction and time series analysis.
- **Pattern Evaluation Modules**- Here the interestingness of patterns is measured with the help of a threshold value.
- **Graphical User Interface**– It acts as a communication link between user and Data Mining system.When a user specifies a query, this module interacts with the system and displays the result in user-friendly format.
- **Knowledge Base**- This module mainly benefits the Pattern Evaluation Module and Data Mining Engine. Pattern Evaluation interacts with the knowledge base to get inputs and update the existing ones if required. Search engine also use the knowledge base for getting more accurate and reliable results.

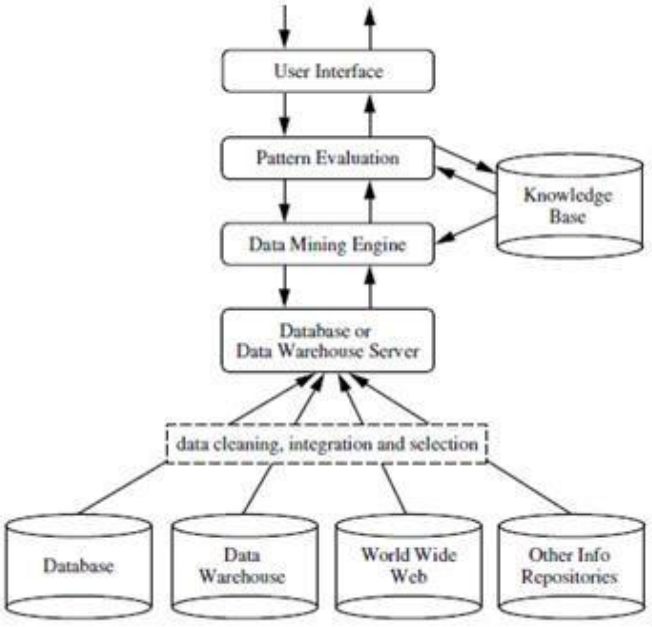


**Figure 1- Data Mining System Architecture**

### *C. Data Mining Functionalities*

Data Mining tasks can be broadly classified as-

- **Descriptive Tasks**
  These talk about the general properties of data stored in database. Mainly they work on discovering patterns like cluster, correlations, trends and anomalies in data. Being precise the main descriptive tasks are
  - Clustering- it is a process of collaborating similar objects together into groups called clusters.
  - Summarization – It describes individual classes or objects in precise forms hence this method is also called class/concept description.
  - Association Rules- it refers to mining of association rules that describe relation between two entities in the database.
  - Sequence Discovery- Discovering patterns or sequences that often occur in transactional databases.
- **Classification and Predictive Tasks-**
  These further include the following sub-tasks.
  - Classification-There are certain objects whose class labels are unknown.Classification is a process to find a model that will describe theseentities' classes and concepts.
  - Prediction- It is used to estimate missing numerical value or unavailable values instead of class labels. It can also be used to determine distribution trends using available data.
  - Regression- It is one such general method of prediction.
  - Time Series Analysis- this analysis showcases model trends, regularities and description for those data objects whose performance changes with time.

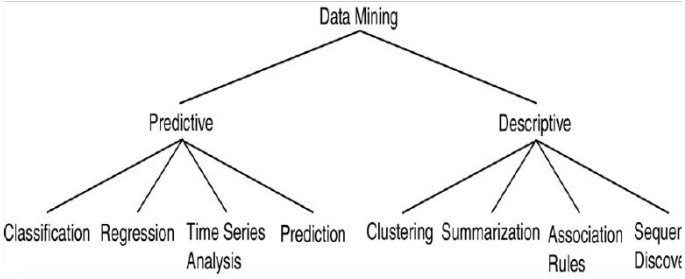


**Figure 2-Data Mining Tasks**

Out of these tasks we will explain Association Rules andTime series analysis in detail.

## II. ASSOCIATION RULE MINING

### *A. Concept*

It is one of the most important technique in Data Mining, mainly used to find frequent patterns, associations or correlations among itemsets of a transactional, relational database or any other information repository.

Mining association rules from large amount of data is helpful in taking business decision like market basket analysis, retail store designing, storage planning and user classification as per the buying patterns.

Association rules are mainly expressed as X->Y where X, Y are itemsets. Its main concepts are expressed as: -

- Itemset- Two or more items together form an itemset. K-itemset will contain k items.
- Count- It denotes the number of itemset is encountered in database records.
- Support- Fraction of transactions/records that contain both X and Y to the total number of records in the database. It is used to find the strongest association rules in the itemset and is expressed as – Support (XY): support count of XY/Total number of transactions in D.
- Minimum Support- It is a user-defined value which says that itemsets below this support threshold is not interesting.
- Confidence- It is also used to find the association rules. Confidence of an association rule can be explained as a fraction or percentage of number of transactions that contain X and Y to the total number of records that contain only X. It is thus expressed as: -
  Confidence(X/Y):Support (XY)/Support(X)
- Minimum Confidence- It is a user-defined value which says that association rules below this threshold value are not interesting.
- Lift- The lift of a rule is defined as: -
  Lift(X->Y): Support (X n Y)/Support(X) * Support(Y)
- Association Rules- Given two itemsets X and Y such that (X n Y) = null (empty form). The association rule of the form X-> Y can be generated. It depicts that the presence of X increases the probable presence of Y.

### *B. Algorithm*

Apriori Algorithm is used to discover Association Rules. It employs an iterative approach known as a breadth-first search (level-wise search) through the search space, where k-itemsets are used to explore (k+1)-itemsets. It includes the following steps: -

- Generate and Test- Here, first the database is scanned and the 1-itemset frequent elements L which satisfy the minimum support criteria are found.
- Join Step- Now, the Cartesian product (Ck) is obtained by self- joining the previous frequent elements i.e. Lk-1* Lk-1. Let Ck denote candidate k-itemset and Lk be the frequent k-itemset.
- Prune step-Ck is the superset of Lk so members of Ck may or may not be frequent but all K ' 1 frequent itemsets are included in Ck thus prunes the Ck to find K frequent itemsets with the help of Apriori property. I.e. This step eliminates some of the candidate k-itemsets using the Apriori property A scan of the database to determine the count of each candidate in Ck would result in the determination of

Lk (i.e., all candidates having a count no less than the minimum support count are frequent, and therefore belong to Lk). Ck, however, can be huge, and so this could involve grave computation.

To shrink the size of Ck, the Apriori property is used as follows. Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset. Hence, if any (k-1)-subset of candidate k-itemset is not in Lk-1 then the candidate cannot be frequent either and so can be removed from Ck. Step 2 and 3 is repeated until no new candidate set is generated.

## III FORECASTING USING TIME SERIES ANALYSIS

### *A. Concept*

Data Mining has developed some powerful methods which enable us to see things ahead of time. One of such a method is Forecasting. Time series modelling helps to derive hidden insights and make informed decisions. For e.g. Business houses may use this modelling to analyse sales for next year, competition position, website traffic and much more.

Time series modelling requires the series to be stationary. By stationary we mean that firstly mean of the time series should not be a function of time. Similarly, next check is that variance should exhibit homoscedasticity i.e. again, it should not vary with time. Finally, the same property must be possessed by the covariance of (i)th term and (i+m)th term. In case the series is non-stationary we need to stationeries the series by differencing, transforming or detrending and then choose the appropriate time series model. The popular models are discussed below.

### *B. Models*

- Auto-Regressive Time Series Model

This model These are used to model the time series in which the next instance is purely dependent on the previous ones. Any shock in the series gradually fades down. Thus, its equation can be given by the following graph and equation.This figure shows an example of sudden increase (shock) and gradual decrease of cold drink demands in winter season.

**x(t) = alpha * x (t – 1) + error (t)**

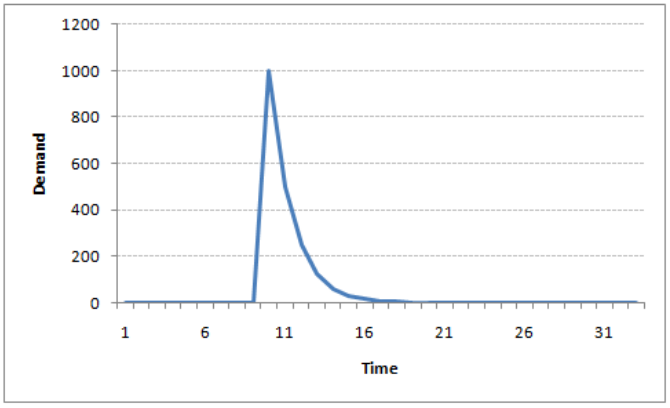


**Figure 3-Graph of Auto regressive Model**

1. **x(t) = beta * error(t-1) + error (t)**

Also, in MA, unlike AR the shock/noise will quickly vanish. An example below says the sudden increase in a shopkeeper's bag sale. All the bags get sold on same day leaving no bags or sale for the next day. Thus, the graph drops down quickly.

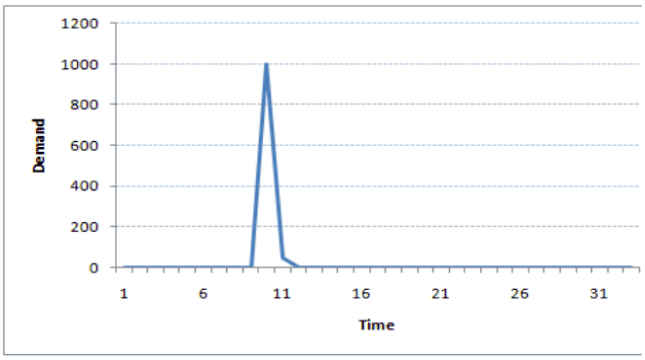


**Figure 4 -Graph of Moving Average Model**

### *C. Deciding models to be used*

Whether the series must be modelled using AR, MA or both i.e. ARIMA depends on two important factors ACF (Autocorrelation factor) and PACF (Partial Autocorrelation Factor). ACF is a plot of total correlations between different lag functions. Similarly, PACF is a plot obtained by removing any partial correlations once formed between different lag functions.

If ACF decreases gradually and PACF drops sharply then it is surely AR series.

This can be shown as: -

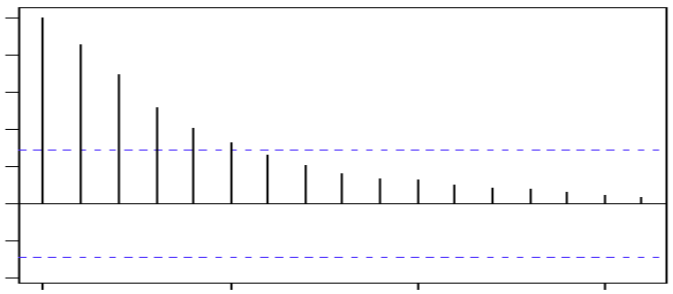

**Figure 5 Graph of Autocorrelation Function in case of AR Model**

Just the opposite case is seen in case of MA series. There are some series which AR or MA are not purely then in that case ARIMA modelling comes into play.

- Moving Average Time series Model

This model is used to depict time series which forecast values on basis of previous terms errors. Thus, its equation can be given by –

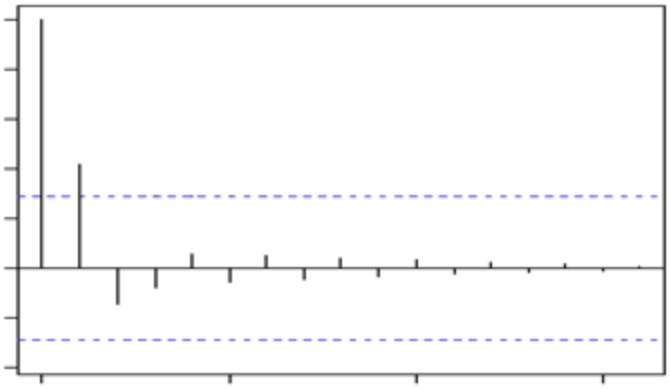

**Figure 6-Graph of Partial-autocorrelation function in case of AR Model**

## CONCLUSION

This paper thus discusses the need and concept of Data Mining. It also describes the two most trending techniques namely Association Rule Mining and Time Series Analysis. We have presented a basic of the research work done in these areas. A lot has been done and much more awaits in discovering efficient methods of rule mining and forecasting. We hope that this paper gives a deep insight of the concept and inspires researchers to make necessary advancements in this area.

## REFERENCES

[1] R. Agarwal and G. Psalia, "Active Data Mining", In Proceeding of the first Intel Conference on Knowledge Discovery and Data Mining,1995.

[2] M.S. Chen,J. Han and P.S. Yu, "Data Mining: An Overview from a database perspective.", IEEE Transactions on Knowledge and Data Mining,8(6): pp. 866-883,1996.

[3] R. Agarwal and R. Shrikant, "Fast algorithm for mining association rules", In Proceeding 20th Intel conference very large databases, VLDB, 1215: pp. 487-499,1994.

[4] Hemlata Sahu, Shailin Sharma and Seema Gondhalakar," A brief overview on Data Mining Survey", IJCTEE, Vol-1, Issue-3,2012.

[5] Dr. Poonam Chaudhary, "Data Mining System, Functionalities and Applications: A Radical Review", IJIET, Vol-5, Issue-2,2015.

[6] G. Vasmi Krishna, "An Integrated approach for weather Forecasting based on Data Mining and Forecasting Analysis", International Journal of Computer Applications (IJCA), Vol-120, no-11, June 2015.

[7] Qin Yu, Lyu Jibin and Lirui Jiang, "An Improved ARIMA-Based Traffic Anomaly Detection Algorithm for Wireless Sensor Networks", International journal of Distributed sensor Networks, vol-2016, no-9653230,9pages, December 2015

[8] Arivarasi, R. And Madhavi Ganesan, "Time series analysis of vegetable production and forecasting using ARIMA model", Asian Journal of Science and Technology, Vol-06, no-10, pp. 1844-1848, October 2015.

[9] Neerja Dhingra, "Yield of Principal Crops in India: growth and Trends", International Journal of Advances in Management and Economics, Vol-4,no-6,pp. 24-28,December 2015.

[10] Mrinmoy Ray, Anil Rai, Ramasubramanian V. And K.N. Singh, "ARIMA-WNN hybrid model for forecasting wheat yield time series data.", Journal of the Indian society of Agricultural Statistics, Vol-70, no-1, pp. 63-70, April 2016.