*i*Journals: International Journal of Software & Hardware Research in Engineering (IJSHRE)
ISSN-2347-4890
Volume 9 Issue 4 April 2021

AN EMPIRICAL STUDY ON DATA MINING TECHNIQUES AND ITS APPLICATIONS

Author: Mrs. T. Pratheebha[1]; Mrs. V. Indhumathi[2], Dr. S. Santhana Megala[3]

Affiliation: Research Scholar, School of Computer Studies, RVS College of Arts and Science, Coimbatore.

Assistant Professor in Computer Science, P.A College of Education, Pollachi.[1]

Assistant Professor, School of Computer Studies, RVS College of Arts and Science, Coimbatore. [2, 3]

E-mail: prathe21@gmail.coml[1]; santhanamegala@rvsgroup.com[2]

10.26821/IJSHRE.9.4.2021.9410

ABSTRACT

Data mining is the process of bring out the information from huge amount of data. Data mining is also referred as Knowledge Discovery from the Data(KDD), it is the convenient extraction of patterns representing knowledge implicitly stored or captured in Large Databases, Data Warehouses, the Web, other massive Data Repositories or Information Streams. In this paper we have discourse various Data Mining Concepts, Process, Techniques and Data Mining Applications.

Keywords: KDD, Data Mining Concept, Knowledge discovery techniques, Application.

1. INTRODUCTION

The development of information technology has brought out a large quantity of databases and data in various aspects. It converts the raw data into useful information in various research fields. The analysis in information bases and information technology has given revolt to an approach in order to store and conserve the valuable data for further decision making. The hidden relationships and trends are not precisely distinct from reviewing the data. Data mining is a multi - dimensional process involves in bringing out the data by retrieving and assembling them, evaluate the results and express them. Data mining tools can analyze the huge amount of data in the databases by using high - efficiency client/server or parallel processing computers [2].

This paper consists of 6 sections. Section 1 is completely introduction about the multi – dimensional process of data mining where the data is retrieved and assembled. Section 2 consists of knowledge discovery process and its steps. Section 3 consists of data mining lifecycle. Section 4 focuses on several data mining techniques. Section 5 describes data mining applications are which are used in several areas. Section 6 is about the conclusion of further research.

2. DATA MINING:

Data mining may be a process of discovering of needed information from large sets of knowledge, it's also called as knowledge discovery process. Knowledge mining from data, knowledge extraction or data/pattern analysis, typically deals with data that have already been collected for a couple of purposes instead of the information mining analysis. This suggests that the objectives of knowledge mining exercise play no role within the data collection strategy. The data sets examined in processing are often large. Data processing uses mathematical analysis to derive patterns and trends that exist in data. Typically, these patterns cannot be discovered by traditional data exploration because the relationships are too complex or

because there's an excessive amount of knowledge. The data mining techniques are used to operate large volumes of knowledge to get hidden patterns and relationships helpful in deciding. So, many people use the term "Knowledge discovery in data" or KDD for data mining [1]. In Data mining, Knowledge extraction or discovery is done in several steps as in figure 1.

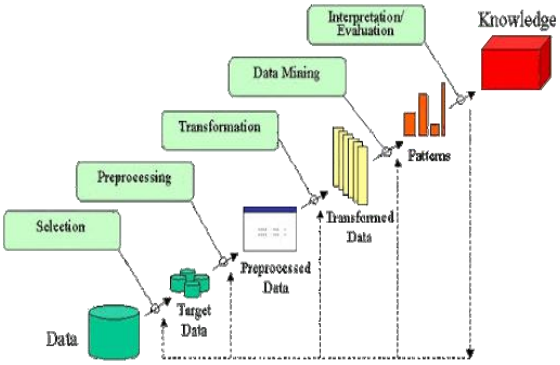

Figure 1: Knowledge Discovery in Data (KDD) Process

1. Data Selection: Data which is useful for the analysis is taken from the data source using Decision Trees, Naive Bayes, Clustering, Neural Networks and Regression techniques [3].
2. Data Cleaning/ Data Preprocessing: This is the basic step to remove noise and inconsistent data from the collected raw data.
3. Data Transformation: Data are transformed and consolidated into forms appropriate for mining by performing the aggregate operations.
4. Data Mining: In this step, an essential process of intelligent methods is applied to extract the data patterns.
5. Data Integration: At this step, where the multiple sources are combined and stockpile in a single source.
6. Interpretation and Evaluation: The patterns acquired in data mining stage are converted into knowledge by removing the irrelevant patterns and translate into useful patterns in terms to human understandable.
7. Knowledge Presentation: This is the last stage, where the visualization and knowledge representation techniques are used to understand and make clear the knowledge of data mining process to users.

The main motive of knowledge discovery and the process of data mining are to discover the patterns that are unknown among the large set of data and make useful knowledge and information.

3. DATA MINING LIFE CYCLE:

The life cycle of a data mining project consists of six phases. The sequence of the phases isn't rigid. Moving back and forth between different phases is usually required. It depends on the result of every phase. The following are the some of the phases,

3.1 Business Understanding:

This phase focuses on understanding the project objectives and requirements from a business perspective, then converting this data into a knowledge mining problem definition and a Preliminary plan designed to realize the objectives.

3.2 Data Understanding:

It starts with an initial data collection, to get well known with the data, to identify data quality problems, to get first insights into the data or to detect interesting subsets to make hypotheses for hidden information.

3.3 Data Preparation:

In this stage, it collects all the various data sets and constructs the activities based on the initial data.

4. MOST IMPORTANT DATA MINING TECHNIQUES:

There are several techniques of data mining such as Classification, Clustering, Regression, Outlier detection, Association rules and Sequential patterns are few most commonly used data mining techniques are used for knowledge discovery from database.

4.1 Classification:

Classification is the most usually applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the records of population at large. This method often employs decision tree or neural network-based classification

algorithms [4]. The common characteristics of classification tasks are as supervised learning, categories dependent variable and assigning new data to one of a set of well-defined classes. Classification technique is employed in client segmentation, modeling businesses, credit analysis, and lots of alternative applications. E.g., classifying the countries supported population, or bikes supported mileage.

4.2 Clustering:

Clustering could be a method of grouping a collection of physical or abstract object into categories of similar objects. A cluster is a group of objects which are "similar" between them and are "dissimilar" to the objects belonging to alternative clusters. Using the clustering techniques we can able to determine the dense and thin regions in object space and may additionally discover overall distribution pattern and correlations among data attributes. E.g., form a group of customers based on purchasing patterns, to categories genes with similar functionality.

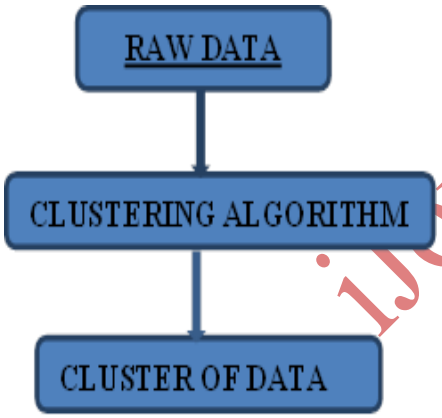

Figure 2: Clustering Techniques

4.3 Sequential patterns:

Sequential patterns technique is used to predict sequential dependencies and sub sequences. This Methods is used for finding sequential patterns are GSP (Generalized sequential pattern), free span, prefix span, SPADE (sequential pattern discovery using equivalent class). Some of the applications are DNA sequences, weblog click streams, telephone calling patterns, stocks and markets etc.

4.4 Outlier detection:

A data set may contain objects that do not comply with general behavior or model of the data. These data objects are called outliers. Many data mining methods discard outliers as noise or exceptions. In some applications the rare events can be more interesting than the regularly occurring events. The analysis of outlier data is referred to as outlier analysis or anomaly analysis. This analysis may help in detecting fraud and to predict abnormal values [5].

4.5 Association Rule:

Association rule is one of the best techniques in data mining. In association, a pattern is abstracted by a relationship of a particular item on other items in the same transaction. E.g. the association technique is used in market basket analysis to find out the products which the customer purchased regularly. On the base of this data, business can have corresponding marketing campaign to sell more products to make more profit [6].

4.6 Regression:

Regression analysis is a statistical methodology that is most often used for numeric prediction, although methods exist as well. It is used to model the relationship between one or more dependent variables and independent variables. In data mining independent variables are attributes already known and response variables are what we want to predict. This technique inspect the dependency of some attribute values, which is dependent upon the values of other attributes mainly, present in same item. Here the regression technique target values are known. For example, predicting the child's behavior with their family history [7].

5. CURRENT TRENDS IN DATA MINING APPLICATIONS:

The Data mining applications are widely used in several areas such as healthcare, Education, Market Basket Analysis, Intrusion Detection, E-commerce, Retail Industry, Telecommunication Industry, Financial Data Analysis, Bioinformatics, Web Mining, Data Mining in Insurance, Earthquake Prediction, Data Mining in Agriculture, Cloud Computing, Research Analysis, Customer Relationship Management.

5.1 Medicare and health care:

Data mining in Medicare enables to characterize the activities of the patient to see incoming office visits. Data mining supports to predict the patterns of medical therapies in different sickness, it is also used to predict the volume of patients in future and for preventing from disease and also to sort out the complex problem of,

- Prediction on medical diagnosis.
- predicting of various diseases,

- Assisting with all diagnosis and advising the doctors in making further clinical decisions.

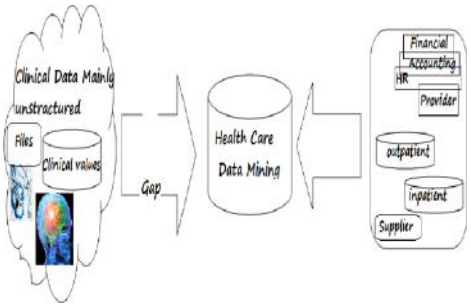

Figure 3: Data mining process in Health Care sector

5.2 Education:

Data Mining in Education is a blooming field which provides knowledge from educational Environment data. This study improves the educating methods by understanding the ward and to take accurate decisions respectively. The goals of EDM are identified as,

- Predicting students' learning behavior, emotions and skills [8].
- Predicting Problems on educational process management
- Predicting Problems of automatic extraction of interpreted knowledge from educational data.

Figure 4: Data Mining Techniques in Education domain

5.3 Market Basket Analysis:

Market Basket Analysis is a technique that uses association rule mining to understand the purchasing conduct of the customer. It also approves the vendor to understand his business, customer's needs and to make profitable alternate accordingly. The ultimate goal of market basket analysis is to find the products where the customers routinely purchase together [9].Here is some of the examples of data mining,

- Prediction of association relationships between large quantities of business transaction data which can help in catalog design,
- Cross-marketing and
- Decision making process in various business activities.

Figure 5: Data Mining Techniques in Market Basket Analysis

5.4 The Intrusion Detection in the Network:

The intrusion detection in the Network is very difficult and needs a very close watch on the data traffic. It plays a necessary role which is taken in the process of computer security. The classification method of data mining classifies the network traffic, normal traffic or abnormal traffic [10]. If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly. Here are some of the fields in where the data mining technology has been adapted −

- Development of data mining process for intrusion detection.
- Association and alternation analysis, collection to help select and build discriminating aspect.
- Analysis of Stream data.
- Distributed data mining.

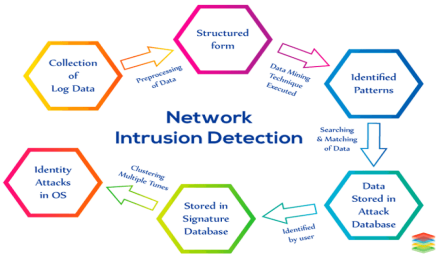

Figure 6: Data Mining Techniques in Intrusion Detection

5.5 A Malicious Executable is Threat:

A malicious executable is a threat to system's security, it damage a system or obtaining sensitive

information without the user's permission. The data mining methods can be used to,

- To detect malicious executable before they have been given chance to run [11].
- To detect patterns in large amounts of data, such as byte code, and use these patterns to detect future instances in similar data.

Classification algorithms RIPPER, Naïve Bayes, and a Multi-Classifier system are used to detect new malicious executable. This classifier had shown detection rate 97.76%

Figure 7: Data Mining Techniques in Threat Detection

5.6 Retail Industry:

Data Mining has its great appositeness in retail industry because it collects huge amount of data from sales, purchase history of the customer, transportation of goods, consumption and services. By nature the number of data collected will continue to expand briskly because of the increasing ease, availability and popularity of the web. Data mining in retail industry serve in identifying customer buying patterns and bias that lead to improve the quality of customer retention, good customer service and their satisfaction [12]. Here are some of the examples of data mining in the retail industry.

- Data construction of stockpiles based on the advantages of data mining.
- Multidimensional investigation of customers sales, product details, region and time.
- Determining the efficiency of sales campaigns.
- Customer retention and service.

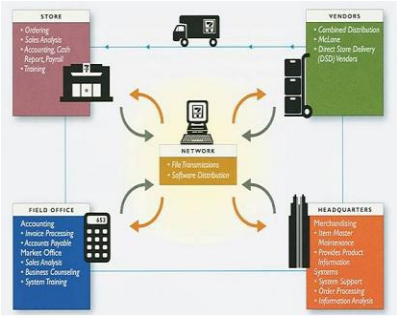

Figure 8: Data Mining Techniques in Retail Industry

5.7 Telecommunication Industry:

At present the telecommunication corporation is one among the foremost emerging industries providing different services like fax, pager, telephone, internet messenger, images, e-mail, web data transmission, etc. Due to the event of recent computer and communication technologies development, the telecommunication industry is briskly expanded. This is often the rationale behind data processing becoming very essential to assist and understand the business, data processing in telecommunication corporation helps in identifying the telecommunication patterns, catch fraudulent activities, make better use of resource, and amend quality of service. Here is that the arrangement of examples that data processing improves telecommunication aids

- Multidimensional Analysis of Telecommunication data.
- Fraudulent Pattern Analysis.
- Identification of Uncommon Patterns.
- Multidimensional Association and Sequential Patterns Analysis.
- Mobile Telecommunication Aids.

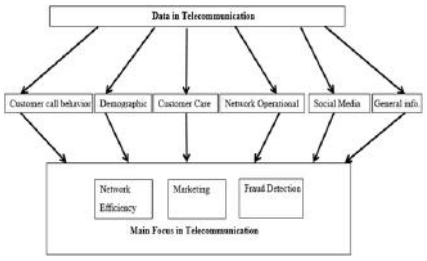

Figure 9: Data Mining Techniques in Telecommunication

5.8 Financial Data Analysis:

The financial data in banking and financial industry is broadly reliable and of top quality which facilitates systematic data analysis and data extracting. Some of the classical cases are as follows,

- Design and construction of data stockpiles for multidimensional data analysis and data mining.
- Customer credit policy analysis and Loan payment.
- Clustering and classification of customers for aspire marketing.
- Detecting money laundering and other financial delinquency.

Figure 10: Data Mining Techniques in Financial Analysis

5.9 Bioinformatics:

Bioinformatics [13] is the collection of various methods to manage, store and study biological data using computers. The data in this field were increasing every day and used extensively for research purposes. The following are some of the areas where data mining techniques applied,

- Gene sequence finding,
- Protein sequence analysis,
- Gene and protein communication network construction,
- Disease detection, DNA sequencing and aligning etc.

Sequence data set is used in bioinformatics. Based on the application type, this sequence dataset is given to tools of data mining to retrieve required results. Some of the mining tools used in bioinformatics are BLAST (Basic Local Alignment Search Tool), FASTA, CS-BLAST for Finding Sequence Alignment, GenScan, Gene Mark for Gene finding, Pfam, Blocks, ProDom for Protein Analysis etc.

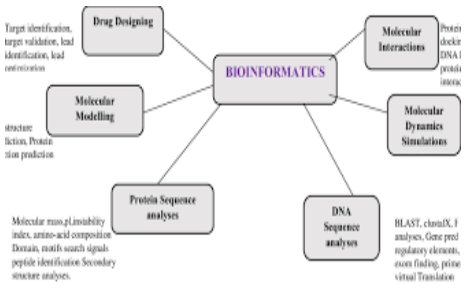

Figure 11: Data Mining Techniques in Bio Informatics

5.10 Web Mining:

Web mining uses methods of data mining to find relevant web documents and patterns from websites. Several data mining techniques are used in some applications such as,

- Web Content Mining (to extract useful information from web documents)
- Web Structure Mining (to discover structure information from the website)
- Web Pattern Mining (log mining). Data mining tools used here are SAS (Statistical Analysis System), Scrappy, Page rank etc.

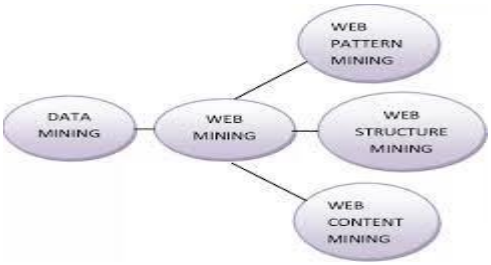

Figure 12: Data Mining Techniques in WEB

5.11 Data Mining in Insurance:

Data mining enables to forecasts which customers will potentially purchase new policies. Data mining permits insurance companies to bring out risky customer's behavior patterns. Data mining helps detect fraudulent behavior [14] such as,

- Eligibility fraud,
- Auto insurance fraud,
- Credit fraud etc.

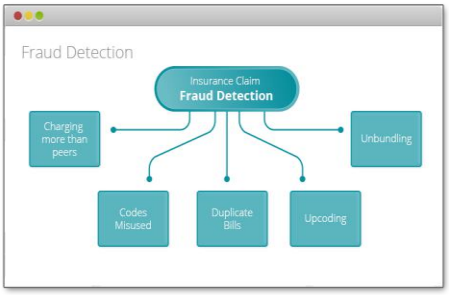

Figure 13: Data Mining Techniques in Insurance domain

5.12 Data Mining in Earthquake Prediction:

Predict the earthquake from the satellite maps. Earthquake is that the sudden movement of the Earth's crust caused by the abrupt release of stress accumulated along a geologic fault within the interior. There are two basic categories of earthquake predictions,

- Forecasts (months to years in advance) and
- Short – terms predictions (hours or days in advance).

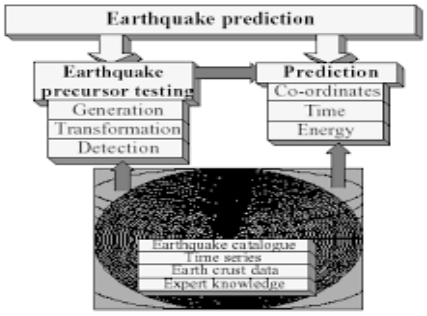

Figure 14: Data Mining Techniques in Earthquake prediction

5.13 Data Mining in Agriculture:

Data mining than emerging in agriculture field for crop yield analysis a with reference to four parameters namely year, rainfall, production and area of sowing.Yield prediction may be a vital agricultural problem that is still to be solved supported the available data. The yield prediction problem are often solved by employing data processing techniques like K Means, K nearest neighbor (KNN), and Artificial Neural Network [15].The following are a number of the issues in agriculture,

- Prediction of crop yield,
- Planning of oil seed or cash crops for cultivation and grain,
- Enhance the use of pesticide by data mining,
- Detecting the diseases from sounds issued by animals by neural networks.

Figure 15: Data Mining Techniques in Agriculture

5.14 Cloud Computing:

The implementation of techniques in Data Mining through Cloud computing will permits the users to bring out the useful information from virtually integrated data storehouse that reduces the costs of framework and storage [16]. Cloud computing uses the Internet services that rely on clouds of servers to handle tasks. The techniques of data mining in Cloud Computing support to perform reliable, efficient, and secure services for their users. Here are some of the examples of data mining in the Cloud computing.

- Security of data,
- Insufficient of resources and expertise,
- Dealing with multi cloud environment,
- Cloud cost management.

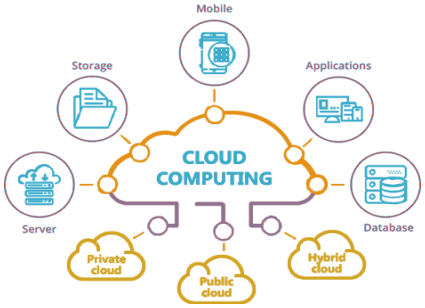

Figure 16: Data Mining Techniques in Cloud Computing

5.15 Criminal Investigation:

Data mining applied within the context of enforcement and intelligence analysis holds the promise of alleviating crime related problems. Criminal analysis includes investigations of crimes and criminal relationships with those crimes. Different crimes like cyber-crimes, violent crimes, fraud detection, drug offences, we get high volumes of criminal datasets. Data mining is utilized in this field like,

- Counter-Terrorism Activities,
- Legal Judgement Summarization[17],
- Crime Matching, Crime Trends and
- Sex Crime, Theft, Fraud, Arson, Gang Drug Offences, Violent Crime and Cyber-Crime.

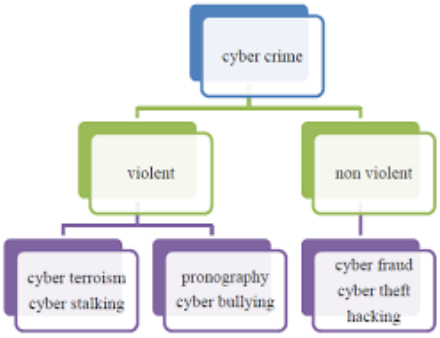

Figure 4: Data Mining Techniques in Criminal Investigation

6. CONCLUSION:

According to the techniques and applications of data mining mentioned above, it is learned that it is a powerful and necessary technique for manipulate the data where data mining gives proper and targeted outcome from huge and vast growing data worldwide. This paper discusses the concept of data mining, the process of KDD, different techniques such as clustering, association, classification, prediction, outlier detection and so on and also discussed some insights of the data mining applications. Data Mining has its great application in Retail Industry because it collects large amount of data from sales, customer purchasing history, goods transportation, consumption and services. Even though it focuses on different aspects, there is a need in identifying customer buying patterns and bias that lead to improve the quality of good customer service, customer retention and their satisfaction. So Here I am focusing my research towards retail industry.

REFERENCES:

[1] Sathiyapriya, S & Kanagaraj, A, 'Basics of Data Mining Techniques and its Application', IJCAT - International Journal of Computing and Technology, Vol: 5, Is: 4, 2018.

[2] Siddhesh Chavan, Aditya Jadhav, Prashik Suryagandh, & Prof. Nidhi Sharma, 'Data Mining Techniques to Improve Customer Relationships Management', International Research Journal of Engineering and Technology (IRJET), Vol: 05 Is: 02, 2016.

[3] Koti Neha & Yogi Reddy, M, 'A Study on Applications of Data Mining', International Journal of Scientific & Technology Research Vol: 9, Is: 02, February 2020, ISSN 2277-8616.

[4] Prema, K & Kumar Kombaiya, A, 'A survey on use of Data Mining Methods Techniques and Application', IJARSE : International Journal of Advanced Research in Science and Engineering, Vol.6, Is.12, PP.532-539, 2017.

[5] Dr. Varun Kumar & Anupama Chadha, 'An Empirical Study of the Applications of Data Mining Techniques in Higher Education', (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 2, No.3, 2011.

[6] Kalyani M Raval, 'Data Mining Techniques and Applications', Indian Journal of Computer Science and Engineering, Vol.1, No. 4, PP: 301-305, 2016.

[7] Smita & Priti Sharm, 'Use of Data Mining in Various Field: A survey Paper', IOSR-JCE: IOSR Journal of Computer Engineering, Vol.16, Issue.3, PP.18-21, 2014.

[8] Brijesh Kumar Baradwaj & Saurakh Pal, 'Mining Educational Data to Analyze Students Performance', IJACSA: International Journal of Advanced Computer Science and Applications, Vol.2, Issue.6, PP.63-69, 2011.

[9] Arun K Pujari, 'Data Mining Techniques', Universities Press, India, PP.98-120, 2017.

[10] Cai, W & Li L., 'Anomaly Detection using TCP Header Information, STAT753 Class Project Paper, May 2004.". Web Site:http://www.scs.gmu.edu/~wcai/stat753/stat753report.pdf.

[11] Schultz, M.G, Eskin, Eleazar, Zadok, Erez&Stolfo, Salvatore , J, 'Data Mining Methods for Detection of New Malicious Executables', Proceedings of the IEEE Symposium on Security And Privacy, IEEE Computer Society Washington, DC, USA , ISSN:1081-6011, 2011.

[12] Gandhimathi, D & Al-Rehnaz Anupama Sudeep, 'Study On Application Of Data Mining And Security Measures', International Journal of Advanced Technology in Engineering and Science, Vol.No.4, Issue No.09, 2016.

[13] Stefano Lonardi & Jake Chen, 'Data Mining in Bioinformatics: Selected Papers from BIOKDD in IEEE/ACM Transactions on Computational Biology and Bioinformatics', Vol. 7, No. 2, 2010.

[14] ParamjitKaur&Kanwalpreet Singh Attwal, 'Data Mining: Review', (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (5), PP: 6225-6228, 2014.

[15] ParamjeetKaur, 'Data Mining Techniques and its Applications: An Approach to Discover Knowledge in Data',

International Journal of Scientific Research in Computer Science, Engineering and Information Technology © 2017 IJSRCSEIT | Volume 2 | Issue 6 | ISSN : 2456-3307.

[16] Porkodi K, SanthanaMegala S, "An Empirical Study on Cryptographic Algorithms implemented in Cloud Computing Environment", IOSR Journal of Engineering, Volume: 4, PP :23-33, 2018, ISSN : 2250 - 3021.

[17] SanthanaMegala S, "Classification of Legal Judgement Summary using Conditional Random Field Algorithm" International Journal of Computer Sciences and Engineering, Vol:5, Iss : 8, June 2018, ISSN : 2347 2693.

