Journal of Computer & Robotics Education Research (JOCRE)
Vol. 1, No. 1, 2024

APPLICABILITY OF DATA MINING SOFTWARE IN PROBLEM-SOLVING AND DATA ANALYSIS

[1]Eze John E, [2]Eze L. Ifeoma, & [3]Eze G. Nnedinso

[1]Department of Mathematics Education

[1&2] Department of Computer Science Education

School of Secondary Education Sciences

Federal College of Education (Technical), Bichi, Kano state

E-mail: [1]justusezeone@gmail.com , [2] ifeomaezelov@gmail.com,

[3] gloriahchinedu@gmail.com

Abstract

Information plays vital role in every sphere of the human life. It is very important to gather data from different data sources, store and maintain the data, generate information, generate knowledge and disseminate data, information and knowledge to every stakeholder. Due to vast use of computers and electronics devices and tremendous growth in computing, power and storage warehouse enables entire enterprise to access a reliable current database. To analyze this vast amount of data and drawing fruitful conclusions and inferences, it needs the special tools called data mining tools. This paper gives an overview of the data mining software tools, data mining techniques, data mining task, its applications in problem solving and data analysis and challenges of data mining usage.

Keywords

Data mining, Data mining task, Data mining software ,Data mining techniques, Knowledge discovery process, Data mining applications

Introduction

To generate information it requires massive collection of data. The data can be simple numerical figures and text documents, to more complex information such as spatial data, multimedia data, and hypertext documents. To take complete advantage of data; the data retrieval is simply not enough, it requires a tool for automatic summarization of data, extraction of the essence of information stored, and the discovery of patterns in raw data. With the enormous amount of data stored in files, databases, and other repositories, it is increasingly important, to develop powerful tool for analysis and interpretation of such data and for the extraction of interesting knowledge that could help in decision-making. The only answer to all above is 'Data Mining'.

Data mining is the extraction of hidden predictive information from large databases; it is a powerful technology with great potential to help organizations focus on the most important information in their data warehouses (Larose, 2005; Dunham & Srilhar; 2006)..Data mining tools predict future trends and behaviors, helps organizations to make proactive knowledge-driven decisions . The automated, prospective analyses offered by data mining, move beyond the analyses of past events provided by prospective tools typical of decision support systems. Data mining tools can answer the questions that traditionally were too time consuming to resolve. They prepare databases for finding hidden patterns, finding predictive information that experts may miss because it lies outside their expectations.). According to Osman and Abass,(2016), data mining software forecast behaviors and potential trends, permitting business personnel to make proactive, knowledge-driven decisions. Data mining software can provide answer to business questions that will be difficult and time consuming to do manually.

Data mining, popularly known as Knowledge Discovery in Databases (KDD), is the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases (Neelamadhab, Pragnyaban & Rasmita; 2012). It is actually the process of finding the hidden information/pattern of the repositories**.**

The paper therefore describes data mining tasks and how the data will be stored, how to retrieve and how to analyze the data. Data mining software tools, data mining techniques or methods used to analyze data, applicability of data mining software in problem-solving, its challenges and suggestions for improvement were also discussed.

The Data Mining Task

According to Jiawarie and Micheline (2006), Data mining is an interdisciplinary field of astronomy, business, computer science, economics and others to discover new patterns from large data sets. The actual data mining task is to analyze large quantities of data in order to extract previously unknown patterns such as groups of data records (cluster analysis), unusual records (anomaly detection) and dependencies (association rule mining .

The data mining tasks according to Neelamadhab, etal; (2012)are classified as follows:

- **Anomaly detection** (Outlier/change/deviation detection)**:** The identification of unusual data records, that might be interesting or data errors which require further investigation.
- **Association rule learning** (Dependency modeling)**:** 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 purposes.
- **Clustering:** It is a task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- **Classification**: It is the task of generalizing known structure to apply for new data. For example, an email program might attempt to classify an email as legitimate or spam.
- **Regression:** It attempts to find a function which models the data with the least error.
- **Summarization:** It providing a more compact representation of the data set, including visualization and report generation

Data Mining Techniques Used in Data Analysis

Monika and Rajan(2012), identified the following data mining techniques used to Analyze data.

Cluster Analysis

Clustering is the task of assigning a set of objects into groups called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. This approach is used in many fields like machine learning, pattern recognition,image analysis, information retrieval, and bioinformatics. Application of clustering in education can helpi nstitutes group individual student into class of similar

Behaviour, partition the students into clusters, so that students within a cluster (e.g. Average) are similar to each other while dissimilar to students in other clusters (e.g.Intelligent, Weak). The partition of students in cluster is shown in figure 1

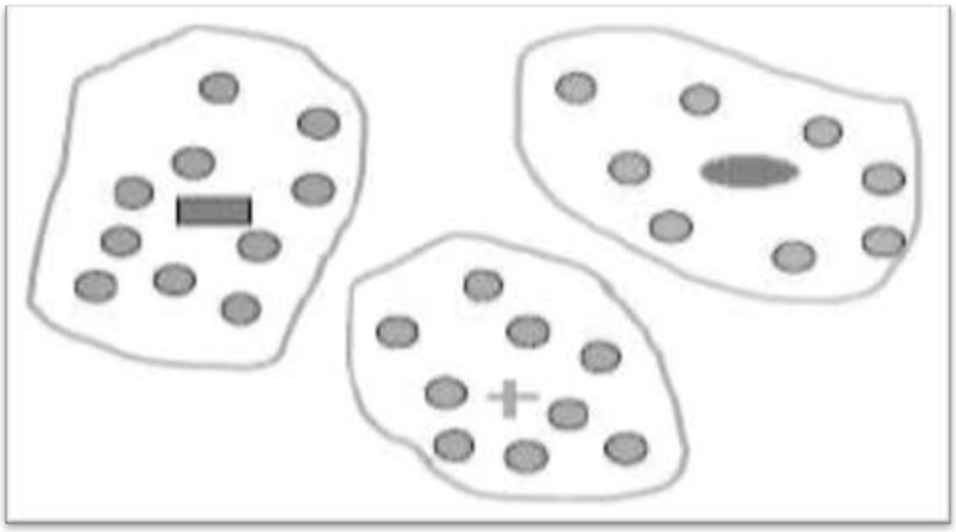


Figure1. Picture showing the partition of students in cluster

Decision Tree

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. Decision trees are commonly used in operations research, specifically in decision analysis, to help identify a strategy most likely to reach a goal. Another use of decision trees is as a descriptive means for calculating conditional probabilities. Decision trees can be used to analyze the admission criteria of an institute. Decision trees are simple to understand and interpret and moreover they give good results even with small data. This

approach may not be suitable for data including categorical variables with different number of levels

Factor Analysis

Factor analysis is a statistical method used to describe variability among observed, correlated variables in terms of a potentially lower number of unobserved, uncorrelated variables called factors. Factor analysis searches for such joint variations in response to unobserved latent variables. The observed variables are modeled as linear combinations of the potential factors, plus "error" terms. Factor analysis originated in psychometrics and used in behavioural sciences, social sciences, marketing, product management, operations research, and other applied sciences that deal with large quantities of data. It can be of two types, Exploratory factor analysis (EFA) and Confirmatory factor analysis(CFA)

Exploratory factor analysis (EFA): is used to uncover the underlying structure of a relatively large set of variables. The researcher's a priori assumption is that any indicator may be associated with any factor. This is the most common form of factor analysis.

Confirmatory factor analysis (CFA): seeks to determine if the number of factors and the loadings of measured(indicator) variables on them conform to what is expected on the basis of pre-established theory. Indicator variables are selected on the basis of prior theory and factor analysis is used to see if they load as predicted on the expected number of factors

Regression Analysis:

In statistics, regression analysis includes techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps us to understand how the typical value of the dependent variable changes when any one of the independent variables is varied while the other independent variables are held fixed. Regression analysis is widely used for prediction and forecasting, It is used to explore relationship between independent variables and dependent variable, Regression methods mainly used are linear regression and ordinary least squares regression.

Data Mining Software Tools: An Overview

There are several open-source and proprietary data mining tools in the market. It helps implement the organization's data mining strategy. The tools vary in the amount of computational analysis they perform. Therefore, choosing a data mining tool depends on the objective the company wishes to achieve. At its core, these tools help classify, identify trends and run queries on large data sets to provide insights. Some of the most popular free software tools for general data mining include:: RapidMiner, R, Weka, KNIME,

Orange, and Scikit-learn. The description of these tools are as follows.

(i). RAPIDMINER

RapidMiner (previously: Rapid-I, YALE) is a mature,Java-based, general DM tool currently in development by the company RapidMiner, Germany. Previous versions (v.5 or lower) were open source. The latest one (v. 6) is proprietary for now, with several license options (Starter,Personal, Professional, Enterprise). RapidMiner according to Hofmann and Klinkenberg (2013), offers an integrating environment with visually appealing and user-friendly GUI. Everything in RapidMiner is focused on processes that may contain subprocesses. RapidMiner also offers the option of application wizards that construct the process automatically based on the required project goals (e.g.direct marketing, churn analysis, sentiment analysis). There are tutorials available for many specific tasks so the tool has a stable learning curve.

(ii). WEKA

Weka is a Java-based, open-source DM platform developed at the University of Waikato, New Zealand.The software is free under GNU GPL 3 for noncommercial purposes. Weka has had mostly stable popularity over the years, which is mainly due to its user friendliness and the availability of a large number of implemented DM algorithms . Weka offers four options for DM: command-line interface (CLI), Explorer, Experimenter, and Knowledge flow. The preferred option is the Explorer which allows the definition of data source, data preparation,

machine learning algorithms, and visualization . Weka supports many model evaluation procedures and metrics, but lacks many data survey and visualization methods. It is also more oriented towards classification (Hall, Frank, Holmes and Pfahringer (2009).
and regression problems and less towards descriptive statistics and clustering methods.

(iii). R

The open-source tool and programming language of choice for statisticians, R , is also a strong option for data mining(DM) tasks. R has been in development for the last 15 years and is the successor of S, a statistical language originally developed by Bell Labs in 1970s. The source code of R is written in C++, Fortran, and in R itself. According to Zhao (2012), it is an interpreted language and is mostly optimized for matrix based calculations, comparable in performance to commercially available Matlab and freely available GNU Octave. The main language is extended by a myriad collection
of packages for all sorts of computational tasks.

(iv). ORANGE

Orange is a Python-based tool for DM being developed at the Bioinformatics Laboratory of the Faculty of Computer and Information Science at the University of Ljubljana. It can be used either through Python scripting as a Python plug-in, or through visual programming. Its
visual programming interface, Orange Canvas, offers a structured view of supported
functionalities grouped into nine categories: data operations, visualization, classification, regression, evaluation, unsupervised learning, association, visualization using Qt, and prototype implementations (Demsar & Erjavec (2013).

(v).SCIKIT-LEARN

Scikit-learn is a free package in Python that extends the functionality of NumPy and SciPy packages with numerous DM algorithms. It also uses the matplotlib package for plotting charts. The package keeps improving by accepting valuable contributions from many contributors and is supported by both INRIA and Google Summer of Code. One of its main strong points according to Pedregosa, Varoquaux, Gramfort & Michel (2011), is a well written online documentation for all of the implemented algorithms. The package is strong in function-based methods including

many general linear models and various types of SVM implementations. It is also quite fast despite being written in an interpreted language. This is mainly because the contributors are asked to optimize the code in various aspects, such as calling array based NumPy number crunching algorithm or writing wrappers for existing
C/C++ implementations in Cython

(vi).KNIME

KNIME (Konstanz Information Miner) is a general purpose DM tool based on the Eclipse platform, developed and maintained by the Swiss company KNIME.com AG. The tool adheres to the visual programming paradigm present in most DM tools, where building blocks are placed on a canvas and connected to obtain a visual program. In KNIME, these building blocks are called nodes, and according to the official website, more than 1000 nodes are available through the core installation and various extensions. Nodes are organized in a hierarchy and can be searched by name within an intuitive interface. Each node is documented in detail, and the documentation is automatically shown within the interface once the node is selected.

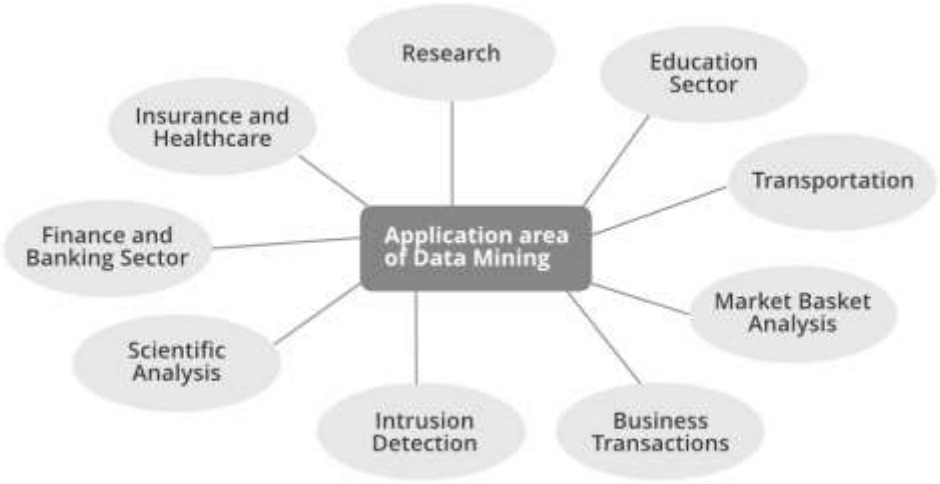
One of the greatest strengths of KNIME is the integration with Weka and R . Weka integration enables using almost all the functionality available in Weka as KNIME nodes, while R integration enables running R code as a step in the workflow, opening R views and learning models within R. Several other interesting free extensions are also available, e.g. JFreeChart extension that enables advanced charting, Open Street Map extension that enables working with geographical data, etc. (Berthold ; Cebron,; Dill; Gabriel; Kötter, T; & Meinl, T. et al ;2008).

Applications of Data Mining Software in Problem Solving and Data Analysis

The data mining applications can be generic or domain specific. The generic application is required to be an intelligent system that by its own, can take certain decisions like: selection of data, selection of data mining method, presentation and interpretation of the result. The domain specific applications are focused to use the domain specific data and data mining algorithm that are targeted for specific objective. The applications studied in this context are aimed to generate

the specific knowledge. In the different domains, the data generating sources generate different type of data. Data can be from a simple text, numbers to more complex audio-video data.

The following are the measurable benefits that have been achieved in different application areas from data mining.



Application in Higher Education:

Data mining can be effectively used to address students and alumni challenges. Data mining facilitate organizations to use their current reporting capabilities to uncover and understand hidden patterns in huge databases. These patterns are then built into data mining models and used to predict individual behavior accurately. As a result of their insight, institutions are able to allocate resources and staff efficiently. This data mining can provide an entity the information necessary to take action before a student drops out, or to efficiently allocate resource with an accurate estimate of how many students will take a particular course(Deshpande & Thakare;2010).

By using education data mining(EDM) we can perform some educational tasks such as:

- Predicting students admission in higher education
- Predicting students profiling
- Predicting student performance
- Teachers teaching performance

- Curriculum development
- Predicting student placement opportunities

Health care and Insurance

Data mining applications in health can have tremendous potential and usefulness. However, the success of healthcare data mining hinges on the availability of clean Health care data. In this respect, it is critical that the healthcare industry look into how data can be better captured, stored, prepared and mined. Possible directions include the standardization of clinical vocabulary and the sharing of data across organizations to enhance the benefits of healthcare data mining applications (Thakare, 2010).

Insurance industry growth is completely depends on the ability of transforming data into information regarding customers, competitors and its market. The insurance industries have implemented the Data Mining successfully and have achieved tremendous competitive advantages. The data mining applications in insurance industry can be used in the form that, data mining is applied in claims analysis such as identifying the medical procedures which are claimed together. Data mining enables to forecasts the potential customers who will buy new schemes. Data mining also helps in detecting fraudulent behavior

Retail Industry:

Retailers have been collecting large amount of data like customer information and related transactions, product information etc. This significantly improves the applications like product demand forecasting, assortment optimization, product recommendation and assortment comparison across retailers and manufacturers (Ghani, Probst, Liu, Krema,& Fano; 2015.). To update the product details database is thus the main issue. The text mining application for extraction of implicit attributes and explicit attributes from product descriptions documents is the main task in such applications. Retail data mining helps in analyzing client behavior, client patterns of shopping and trends which increases the quality of client service, enhance things

consumption ratios, design more effective goods transportations and distribution policies, achieve better customer retention and satisfaction and minimize the cost of business.

Telecommunications industry*:*

Telecommunication industries generally generate and store large amount of high quality data, having a very huge customer base, and operate in rapidly changing and highly competitive environment. Telecommunication companies use data mining to enhance their marketing efforts to detect fraud and to betterment of their telecommunication

Pharmaceutical industry

The pharmaceutical industry is well known for performing quantitative analysis for clinical research and market research (Bhramaramba, Allam, Kumar, & Sridhar; 2011)). In the marketing departments data mining applications are used for sales force planning and direct marketing to doctors and consumers. Data mining techniques used quite well to a variety of critical business decisions in the pharmaceutical industry. It also used for forecasting production schedules for the manufacturing plants, determining market potential in critical go/no decisions on continuing work on development compounds, or making financial projections for stock holders and investors (Cohen; Olivia; Rud; 2005).

Market Basket Analysis

Data mining technique is used in MBA(Market Basket Analysis).When the customer want to buy some products then this technique helps in finding the associations between different items that the customer put in their shopping buckets. Here the discovery of such associations promotes the business technique .In this way the retailers uses the data mining technique so that they can identify the customers intension (buying the different pattern).In this way this technique is used for profits of the business and also helps to purchase the related items. Examples of market Basket Analysis includes:

- Data mining concepts are in use for Sales and marketing to provide better customer service, to improve cross-selling opportunities, to increase direct mail response rates.
- Customer Retention in the form of pattern identification and prediction of likely defections is possible by Data mining.

Risk Assessment and Fraud area also use the data-mining concept for identifying inappropriate or unusual behavior etc

Research: A data mining technique can perform predictions, classification, clustering, associations, and grouping of data with perfection in the research area. Rules generated by data mining are unique to find results. In most of the technical research in data mining, we create a training model and testing model. The training/testing model is a strategy to measure the precision of the proposed model. It is called Train/Test because we split the data set into two sets: a training data set and a testing data set. A training data set used to design the training model whereas testing data set is used in the testing model. Example:

- Classification of uncertain data.
- Information-based clustering.
- Decision support system
- Web Mining
- Domain-driven data mining
- IoT (Internet of Things)and Cybersecurity

Smart farming IoT(Internet of Things

Medical Science

In medical science there is large scope for application of data mining. Diagnosis of diesis, health care, patient profiling and history generation etc. are the few examples. Mammography is the method used in breast cancer detection. Radiologists face lot of difficulties in detection of tumors. Computer-aided methods could assist medical staff and improve the accuracy of detection. The neural networks with back-propagation and association rule mining is used for tumor classification in mammograms. The data mining effectively used in the diagnosis of lung abnormality that may be cancerous or benign (Shu-Hsien; Pei-Hui; & Pei-Yuan; 2012). The data mining algorithms significantly reduce patient's risks and diagnosis costs.

Sports

Sports are ideal for application of data mining tools and techniques. In the sports world the vast amounts of statistics are collected for each player, team, game, and season. Data mining can be

used by sports organizations in the form of statistical analysis, pattern discovery, as well as outcome prediction. Patterns in the data are often helpful in the forecast of future events. Data mining can be used for scouting, prediction of performance, selection of players, coaching and training and for the strategy planning (Solieman; 2006). The data mining techniques are used to determine the best or the most optimal squad to represent a team in a team sport in a season, tour or game(Chodavarapu;2007).

Transportation: A diversified transportation company with a large direct sales force can apply data mining to identify the best prospects for its services. A large consumer merchandise organization can apply information mining to improve its business cycle to retailers.

- Determine the distribution schedules among outlets.
- Analyze loading patterns

Financing and Banking Sector

There are numerous fields in which data mining can be used like in financial and banking sector for credit analysis, fraudulent transactions, customer segmentation and profitability, optimizing stocks portfolios, predicting payment default, ranking investments, marketing, high risk loan applicants, cash management and forecasting operations and most profitable credit card customers and cross selling. Bankruptcy is the major threat to the banking sector(Foster & Stine,2004)]. It increases the cost of lending. The data mining algorithms effectively used for prediction of the personal bankruptcy. Predicting bankruptcy has become the province of computer science rather than statistics. The data mining method least squares regression; neural nets and decision trees are proved to be the suitable for prediction of bankruptcy.

E- commerce

E-commerce is also the most prospective domain for data mining (Neelamadhab, etal; 2012)

It is ideal because many of the ingredients required for successful data mining are easily available: data records are plentiful, electronic collection provides reliable data, insight can easily be turned into action, and return on investment can be measured. The integration of e-commerce and data mining significantly improve the results and guide the users in generating knowledge and making correct business decisions. This integration effectively solves several

major problems associated with horizontal data mining tools including the enormous effort required in pre-processing of the data before it can be used for mining, and making the results of mining actionable.

Application in Digital Library

The Digital Library retrieves, collects, stores and preserves the digital data. The advent of electronic resources and their increased use in libraries has brought about significant changes in Library.(Jadhar & Kumbargoudar, 2007). The data and information are available in the different formats. These formats include Text, Images, Video, Audio, Picture, Maps, etc. therefore digital library is a suitable .domain for application of data mining.

Web Education

Data mining methods are used in the web Education which is used to improve courseware.

The relationships are discovered among the usage data picked up during students' sessions.

This knowledge is very useful for the teacher or the author of the course, who could decide what modifications will be the most appropriate to improve the effectiveness of the course. .In the 21st century the beginners are using the data mining techniques which is one of the best learning method in this era (Anjewierden, Koll¨offel & Hulshof ; 2007). This makes it possible to increase the awareness of learners.

Intrusion Detection in the Network

Intrusions are the set of actions that threatens the availability and integrity of a network resource. Network intrustion detection has been considered to be one of the most promising method for defending complex and dynamic intrusion behaviors. The intrusion detection in the Network is very difficult and needs a very close watch on the data traffic. The intrusion detection plays an essential role in computer security. The classification method of data mining is used to classify the network traffic, normal traffic or abnormal traffic.(Cai & Li, 2004). If any TCP header does not belong to any of the existing TCP header clusters, then it can be considered as anomaly.

A malicious Executable is Threat

A malicious executable is threat to system's security, it damage a system or obtaining sensitive information without the user's permission. The data mining methods are used to accurately detect malicious executable before they run(Schultz, Eskin, Eleazar, Zadok, Erez, Stolfo & Salvatore; 2005). Classification algorithms RIPPER, Naïve Bayes, and a Multi-Classifier system are used to detect new malicious executable. This classifier had shown detection rate of 97.76%.

Language Research and language Engineering

In language research and language engineering many time extra linguistic information is needed about a text. A linguistic profile that contains large number of linguistic features can be generated from text file automatically using data mining (Halteren & Osttdijk; 2007). This technique found quite effective for authorship verification and recognition. A profiling system using combination of lexical and syntactic features shows 97% accuracy in selecting correct author for the text. The linguistic profiling of text are effectively used to control the quality of language and for the automatic language verification. This method verifies automatically if the text is of native quality. The results show that language verification is indeed possible.

Data mining in Crime Detection

The Intelligence Agencies collect and analyze information to investigate terrorist activities.

One challenge to law enforcement and intelligent agencies is the difficulty of analyzing large volume of data involve in criminal and terrorist activities. Data mining makes it easy, convenient and practical to explore very large databases for organizations. The different data mining techniques are used in crime data mining (Poonam; 2015). Entity extraction is used to automatically identify person, address, vehicle, narcotic drug, and personal properties from police narrative reports. Clustering techniques are used to automatically associate different objects such as persons, organizations, vehicles etc. in crime records. Deviation detection is applied in fraud detection, network intrusion detection, and other crime analyses that involve tracing abnormal activities

Challenges of Data Mining Usage and Suggestions for Improvement

In current situation of affairs, data mining research is "too"ad-hoc" and there are so many challenges to unify different data mining tasks. Some of the challenges according to Yang & WU (2012) are as follows.

Interpretability and Explainability. Complex data mining models can be difficult to interpret and explain. These challenges can be addressed as follows:

- Use transparent and interpretable data mining models
- Implement model explainability techniques
- Provide clear insights and visualization

Scalability:

One important challenge is mining data from huge data bases. Computer data network and satellite data can easily be of this scale but to-days technology in data mining are too slow to handle data of this scale. If data mining algorithms are efficient enough to control these massive data sets then they must be scalable. Future data mining should be a continuous, online process instead of one time tiny process. The said scalability also warrants the execution of novel data structure to access individual records in a smooth manner.

High Dimensional Data and High Data Streams:

One challenge is to design classifiers to control ultra-high dimensional classification problems for mining vast, enormous and high dimensional data set out-of-memory, parallel and distributed algorithms. Algorithm is needed to be developed. The traditional data analysis techniques developed for low-dimensional data do not work for high dimensional data.(Poonam, 2015)

Complex and Heterogeneous Data*:*

Another challenge erupted in these years is emergence of more data complex. A good system must scale the complexity from users. Previous analysis of 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 arisen to develop techniques which can handle heterogeneous attributes. Such developed techniques for mining such complex objects ought to have taken care the relationships in data, like temporal and spatial auto-correction, graph

connectivity and parent-child relationships between the components in semi-structures text and XML documents.

Data Ownership, Security and Privacy:

It is a big challenge to find out data for an analysis at one location or to be owned by one location or to be owned by one entity. An automatic data mining in distributed environment can develop serious issues in terms of data privacy or its security. These issues can be addressed by developing of an efficient algorithms and data structures to evaluate the knowledge integrity of a collection of data and further to measure the impact on the modification of data values on discovered pattern's statistical significance.

Data Distribution:

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

a) To minimize the amount of communication needed to perform the distributed computation.

b) To consolidate the data mining results obtains from multiple sources in a efficient manner.

c) To tackle data security issues

Overfitting and Bias. Data mining models can be prone to overfitting and bias, leading to inaccurate decision. This issue can be addressed by use of regularization techniques and cross-validation, implementation of bias detection and mitigation strategies and use of diverse and representative data sets.

Ethical Concerns. Data mining raises ethical concerns such as using data for discriminatory purposes. This can be addressed by the establishment of clear ethical guidelines and principles and implementation of fairness and bias detection mechanisms.

Conclusion

In this paper various data mining applications were reviewed. This review would be helpful to researchers to focus on the various issues of data mining. Data mining technology is an application oriented technology and is having extensive applications in various fields. It also

evaluates, integrates and reasons to guide the solution of practical problems and find the relationships between events. Furthermore predictions of further activities can be easily made by using the prevailing data. Several efforts have been made to design the generic data mining system but no system found completely generic. However, the decisions at different stages in data mining techniques are influenced by factors like context parameters, aim of data mining, domain and details of data. These specific applications are aimed to extract explicit ideas. The results gathered from the domain specific applications are more accurate and useful.

References

Anjewierden, A., Koll¨offel, B., & Hulshof C.(2007). Towards educational data mining: Using data mining methods for automated chat analysis to understand and support inquiry learning processes. *International Workshop on Applying Data Mining in e-Learning*, ADML'07,(305), 23- 32 Sissi,LassithiCrete Greece, 18 September, .

Berthold, M. R; Cebron, N; Dill, F;. Gabriel, T. R; Kötter, T Meinl, T. et al.(2008*). KNIME: The Konstanz Information Miner*, in Data Analysis, Machine Learning and Applications (Studies in Classification, Data Analysis, and Knowledge Organization), Springer Berlin Heidelberg, pp. 319–326,

Bhramaramba, R., Allam, A. R., Kumar, V. V., & Sridhar, G. R. (2011). Application of data mining techniques on diabetes related proteins. *International Journal of Diabetes in Developing Countries*, 31(1), 22– 38

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

Chodavarapu Y.(2007). Using data-mining for effective (optimal) sports squad selections.

WebSite:http://insightory.com/view/74//using_data mining_for_effective_(optimal)_sports_squad_selections

Cohen, J. J., Olivia, C., Rud, P.(2005). Data Mining of Market Knowledge in The Pharmaceutical Industry. *Proceeding of 13th Annual Conference of North-East SAS Users Group Inc*., NESUG2000, Philadelphia Pennsylvania, September 24-26

Demšar, J; Curk, T & Erjavec, A. (2013) "Orange: Data MiningToolbox in Python," Journal of Machine Learning Research, 14(1), 2349−2353.

Deshpande, S. P. Thakare, & V. M. (2010). Data Mining System and applications: A Review. *International Journal of Distributed and Parallel systems (IJDPS*) 1(1). 32-44

Dunham, M. H., & Sridhar S. (2006). *Data Mining: Introductory and Advanced Topics*, Pearson Education, New Delhi, ISBN: 81-7758-785-4, 1st Edition.

Foster, D. P. & Stine, R. A (2004). Variable Selection in Data Mining: Building a Predictive Model for Bankruptcy. *Journal of the American Statistical Association, Alexandria, VA, ETATS-UNIS, 99(1), ISSN 0162-1459, 303-313*

Ghani, R., Probst, K., Liu, Y., Krema, M.& Fano, A.(2015). Text Mining for Product Attribute Extraction ,*SIGKDD Explorations* 8(1), 54 -64.

Han, J., Kamber, M., & Pei. J. (2006). Data mining: concepts and techniques. Morgan Kaufmann Publishers 225 Wyman Street, Waltham, MA 02451, USA.https ://www.google.com.ng.

Halteren, H. V. & Oostdijk, N.,(2007). Linguistic profiling of texts for the purpose of language verification, The ILK research group, Tilburg centre for Creative Computing and the Department of Communication and Information Sciences of the Faculty of

Humanities,TilburgUniversity,The Netherlands."Website: www.ilk.uvt.nl/~antalb/textmining/LingProfColingDef.pdf

Jadhav, S. R.& Kumbargoudar, P.(2007,) Multimedia Data Mining in Digital Libraries: Standards and Features *READIT* , 54-59,

Jiawei, H. & Micheline, K,(2006). Data Mining- Concepts and Techniques: Elsevier Publishers,

Larose, D. T.,(2005). *Discovering Knowledge in Data: An Introduction to Data Mining*, John Wiley & Sons, Inc, , 67-71

Lindell, Y.& Pinkas, B. (2000). Privacy preserving data mining, *in Advances in Cryptology CRYPTO*. *Springer, 2000*, 36–54.

Monika, G. & Rajan, V. (2012). Applications of Data Mining in Higher Education. *International Journal of computer Science.* 9(2), 113-120

Neelamadhab, P, Pragnyaban ,M, & Rasmita, P. (2012). The Survey of Data Mining Applications And Feature Scope. *International Journal of Computer Science, Engineering* and Information Technology (IJCSEIT), 2(3). 43-58

Osman, A. & Abbas, S. (2016). The role of data mining in information security. *International Journal of Computer (IJC)* Global Society of Scientific Research and Researchers. ISSN 2307-4523, 2-4.

Poonam, C. (2015). Data mining system, functionalities and applications: A radical review. *International Journal of innovations in Engineering and Technology(IJIET*), 5(2), 449-455.

Schultz, M. G., Eskin, N; Eleazar, P; Zadok, R; Erez, B; Stolfo. K; & Salvatore, J.(2005), "Data Mining Methods for Detection of New Malicious Executables". *Proceedings of the 2001 IEEE Symposium on Security And Privacy*, IEEE Computer Society Washington, DC, USA.

Shu-Hsien, L., Pei-Hui, C. & Pei-Yuan, H. (2012). Data mining techniques and applications- A decade review from2000 to 2011. *Expert system with application*, 39, 11303-11311

Solieman, O. K.(2008), Data Mining in Sports: A Research Overview. A Technical Report, MIS Masters Project,". Web Site: http://ai.arizona.edu/hchen/chencourse/Osama-DM_in_Sports.pdf

Thakar, V.M. (2010).,Data Mining System and Applications: A Review. ,*International Journal of Distributed and Parallel Systems (IJDPS)*;1(1) DOl : 10-5121/ijdps.2010. 1103

Yang, Q., WU, X. (2006). 10 Challenging Problems in Data Mining Research., *International Journal of Information Technology & Decision Making 5(4):597-604.*

Zhao, Y (2012). R and Data Mining: Examples and Case Studies, San Diego: Academic Press,