Volume 6, Issue 3, March 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Data Mining: Techniques, Key Challenges and Approaches for Improvement

N. Mlambo
Lecturer, College of Science and Technology, University of Rwanda, Rwanda
Jomo Kenyatta University of Agriculture and Technology, Kigali, Rwanda

Abstract: Data mining is an emerging multidisciplinary field which facilitates discovering of previously unknown correlations, patterns and trends from large amounts of data stored in multiple data sources. It is a powerful new technology with great potential to help businesses make full use of the available data for competitive advantages. Data mining application success stories have been told in different areas among them; healthcare, Banking and finance and telecommunication. This survey paper reviews some major mining techniques and key challenges. It also draws attention to useful applications, giving a small collection of real-life examples of data mining implementations from the business to the scientific world.

Key words: Data Mining, Data Mining Techniques, Challenges, Applications

I. INTRODUCTION

Data Mining is the process of sifting through stores of data to extract previously unknown, valid patterns and relationships that provide useful information [1]. Once these patterns are found they can further be used to make certain decisions for development of businesses. For decades major components of data mining technology have been under development in research areas such as statistics, artificial intelligence and machine learning. But in recent times, the maturity of these techniques coupled with high performance relational database engines and broad data integration efforts make these technologies more effective for current data warehouse environments [2]. Data Mining uses sophisticated data analysis tools and visualization techniques to segment the data and evaluate the probability of future events.

This technology has become popular with many businesses because it allows them to learn more about their customers and make smart marketing decisions [16]. In health care [11], [16], [18], data mining techniques have been applied in the diagnosis of diseases such as tuberculosis, diabetes, cancer and heart diseases.

II. DATA MINING TECHNIQUES

Several major data mining techniques have been developed and used in data mining projects. These include association, classification, clustering, prediction and sequential patterns etc.

2.1 Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large. Fraud detection and credit risk applications are particularly well suited to this type of analysis. This approach frequently employs decision tree or neural network-based classification algorithms. The data classification process involves learning and classification. In Learning the training data are analyzed by classification algorithm. In classification test data are used to estimate the accuracy of the classification rules. If the accuracy is acceptable the rules can be applied to the new data tuples. For a fraud detection application, this would include complete records of both fraudulent and valid activities determined on a record-by-record basis. The classifier-training algorithm uses these pre-classified examples to determine the set of parameters required for proper discrimination. The algorithm then encodes these parameters into a model called a classifier [3].

Among several types of knowledge representation present in the literature, classification normally uses prediction rules to express knowledge. Prediction rules are expressed in the form of IF-THEN rules, where the antecedent (IF part) consists of a conjunction of conditions and the rule consequent (THEN part) predicts a certain predictions attribute value for an item that satisfies the antecedent.

Types of Classification Models

The following are some of the common classification models;

(a) Neural Networks
(b) Support Vector Machines
(c) Bayesian Classifiers
(d) Classification based on Associations
(e) Decision Trees

In a survey paper on classification techniques, [4] concludes that Decision trees and Bayesian Network (BN) generally have different operational profiles, when one is very accurate the other is not and vice versa. On the contrary, decision trees and rule classifiers have a similar operational profile. The goal of classification result integration algorithms is to generate more certain, precise and accurate system results. Numerous methods have been suggested for the creation of ensemble of classifiers. Although or perhaps because many methods of ensemble creation have been proposed, there is as yet no clear picture of which method is best. Classification methods are typically strong in modeling interactions. Several of the classification methods produce a set of interacting loci that best predict the phenotype. However, a straightforward application of classification methods to large numbers of markers has a potential risk picking up randomly associated markers.

2.2 Neural Networks

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example [5]. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process.

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. A trained neural network can be thought of as an "expert" in the category of information it has been given to analyze. This expert can then be used to provide projections given new situations of interest and answer "what if" questions.

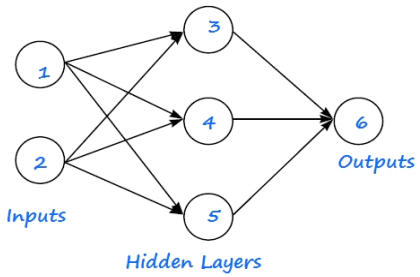

Fig 1: Example of Neural Network

Applications of Neural Networks

Since neural networks are best at identifying patterns or trends in data, they are well suited for prediction or forecasting needs including: sales forecasting, industrial process control, customer research, data validation, risk management, target marketing, etc. To give some more specific examples; ANN are used in the following specific paradigms: recognition of speakers in communications; diagnosis of hepatitis; recovery of telecommunications from faulty software; interpretation of multi meaning Chinese words; undersea mine detection; texture analysis; three dimensional object recognition; hand-written word recognition; and facial recognition [5].

2.3 Clustering

Clustering is a data mining technique of grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. Dissimilarities and similarities are assessed based on the attribute values describing the objects. Clustering algorithms are used to organize data, categorize data, for data compression and model construction, for detection of outliers etc. Common approach for all clustering techniques is to find clusters centre that will represent each cluster. Cluster centre will represent with input vector can tell which cluster this vector belong to by measuring a similarity metric between input vector and all cluster centre and determining which cluster is nearest or most similar one [6]. Cluster analysis can be used as a standalone data mining tool to gain insight into the data distribution, or as a preprocessing step for other data mining algorithms operating on the detected clusters. Many clustering algorithms have been developed and are categorized from several aspects such as partitioning methods, hierarchical methods, density- based methods, and grid-based methods. Further data set can be numeric or categorical. Inherent geometric properties of numeric data can be exploited to naturally define distance function between data points. Whereas categorical data can be derived from either quantitative or qualitative data where observations are directly observed from counts [7].

Data Clustering Techniques

(a) K-Means Clustering
(b) Hierarchical Clustering
(c) DBSCAN Clustering (Density Based Spatial Clustering of Application with Noise). (d) OPTICS
(e) STING

2.4 Support Vector Machines (SVM)

Support vector machines (SVMs) belong to a new class of machine learning algorithms with their origins firmly rooted in statistical learning theory. Due to the strong theoretical foundation these algorithms possess desirable

properties such as the ability to learn from very small sample sets and a firm estimation of the generalization capacity of the learned model.

Unlike traditional methods (e.g. Neural Networks), which minimize the empirical training error, SVMs aim at minimizing an upper bound of the generalization error through maximizing the margin between the separating hyper plane and the data. Since SVMs are known to generalize well even in high dimensional spaces under small training sample conditions and have shown to be superior to traditional empirical risk minimization principle employed by most of neural networks, SVMs have been successfully applied to a number of applications ranging from face detection, verification, and recognition, object detection and recognition handwritten character and digit recognition, text detection and categorization, speech and speaker verification, recognition information and image retrieval [8].

2.5 Association Rule Mining

In data mining [9], association rule learning is a popular and well researched method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using different measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal et al [10], a typical and widely-used example of association rule mining is Market Basket Analysis. The problem is to generate all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence. Association rule mining algorithms include; Apriori, AprioriTid, Apriori hybrid and tertius algorithms.

SNO	PROPERTIES	APRIORI	APRIORITID	APRIORIHYBRID	TERTIUS
1	Candidate generation	Candidate item-sets are generated using only the large item-sets of the previous pass without considering the transactions in the database.	The database is not used at all for counting the support of candidate item-sets after the first pass.	Hybrid algorithm can be designed that uses Apriori in the initial passes and switches to AprioriTid in the later passes.	The Tertius algorithm builds rules out of the attribute pair values in the training data.
2	Methodology used	Uses Join & prune step	Uses Join & prune as well as Tids	Uses Apriori + Aprioritid	Uses first order logic representation.
3	Database scan	Multiple scan over the database.	Uses the database only once.	Uses Apriori + Aprioritid	It is depend on the number of literals in the rules.
4	Memory usage	It takes more space and memory for candidate generation process.	In the kth pass, AprioriTid needs memory for L_{k-1} and C_{k-1} during candidate generation. An additional cost is incurred if it cannot completely fit into the memory.	An extra cost is incurred when shifting from Apriori to AprioriTid.	When the program runs out of memory, the best rules found so far are printed, and a message indicates that the search was interrupted.
5	Execution Time	It takes more execution time for candidate generation process.	For small problem, it's better than Apriori but it takes more time for large problem.	It's better than Apriori and Aprioritid	Its relatively long runtime. Tertius can take up to several hours for some of our larger tests.

Fig. 2: Comparison of Apriori, AprioriTid, AprioriHybrid and tertius [9]

2.6 Sequential Pattern Mining

Sequential pattern mining [11] deals with finding statistically relevant patterns between data examples where the values are delivered in sequence. It is closely related to time series mining and special case of structural data mining. Some of the applications are analysis of customer purchase patterns or Web access patterns, analysis of time related processes involved in scientific experiments, disease treatments, DNA sequencing etc.

Classification Of Sequential Pattern Mining Algorithms

The three main categories of Sequential pattern mining algorithms that have been in use are as follows:

1. Apriori Based algorithms such as GSP, SPADE, SPAM algorithms
2. Pattern Growth algorithms such as FreeSpan and PrefixSpan
3. Early Pruning algorithms such as LAPIN-SPAM. Some other algorithms are hybrids of these techniques.

III. KEY CHALLENGES IN DATA MINING

As an emerging and powerful technology, it is evident that data mining is bound to face a wide range of challenges. Application of the current data mining techniques and algorithms has faced challenges because of inadequacies. Notable

challenges include the need to scale up for high dimensional data and high speed streams, contamination in sequential and time series data, distributed data mining and mining multi-agent data, security and privacy of data, mining complex knowledge from complex and heterogeneous data, interpretation of results (including visualization or any other methods so that the knowledge can be easily understood and directly usable by humans), applying algorithms designed for small datasets when dealing with big data sets, etc. On the interpretation of results, Dunren et al [15] argues that delayed discovery of even highly valuable knowledge invalidates the usefulness of the discovered knowledge. Given below are discussions for some of the most prominent challenges:

3.1 Private and Sensitive Data

Data mining has proved to be a significant technology for gaining knowledge from big and diverse quantities of data. However, there has been growing concern that use of this technology is violating individual privacy. There are many popular data mining applications that deal with sensitive data, such as people's medical and financial records. Ganga [12] argues that the central collection of such data is not desirable as it puts their privacy into risk. In certain cases (e.g. banking, telecommunication) the data might belong to different, perhaps competing, organizations that want to exchange knowledge without the exchange of raw private data.

A lot of research is being carried out on how to deal with this problem. Al-Hamami and Suhad Abu Shehab [13] designed an application that provides protection for privacy and knowledge in data mining. In this application, privacy protection of individuals is achieved by adding a white Gaussian noise to selected columns in a database to be mined for an unauthorized user and the knowledge protection is done by encrypting the result of data mining before it appears to the unauthorized users by using Rijndael algorithm. Fig. 3 shows the steps for privacy protection in data mining.

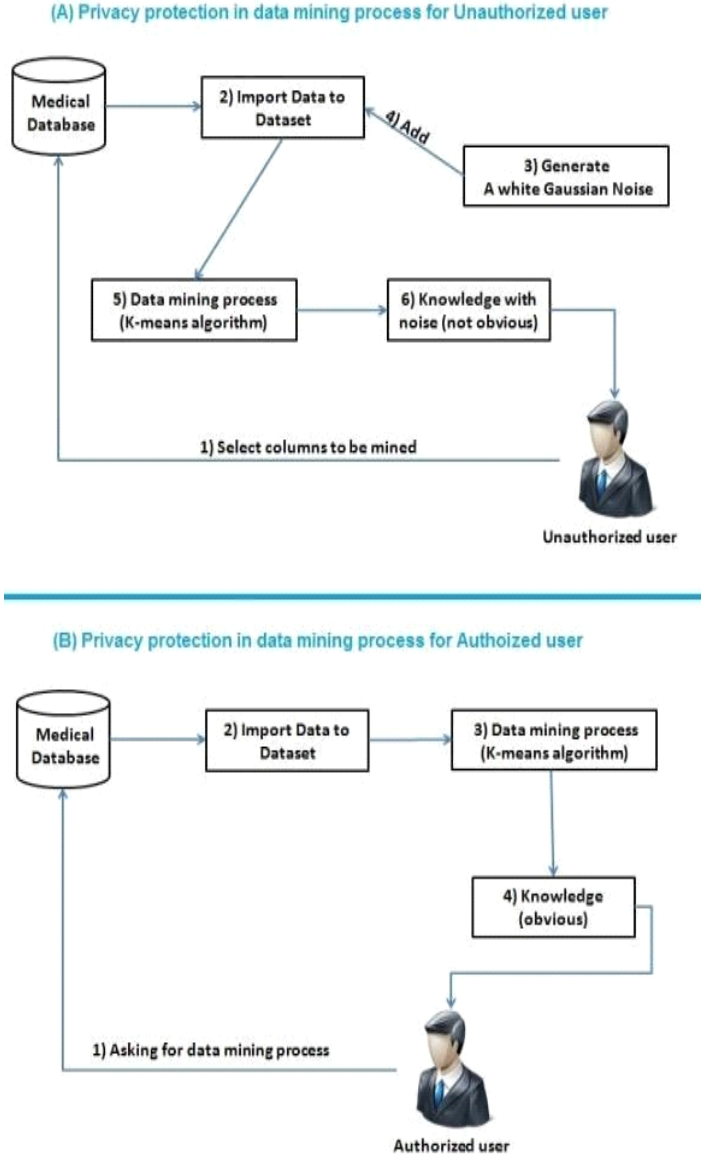

Fig. 3: Steps for Privacy Protection in Data mining process for both authorized and unauthorized user [13].

3.2 Distributed data and operations

The shift towards intrinsically distributed complex problem solving environments is prompting a range of new data mining research and development problems [14]. As reported in [12], [14] and [15], the data which is stored in

distributed computing environments on heterogeneous platforms become impossible to bring to a centralized place because of both technical and organizational reasons. Consequently, development of algorithms, tools, and services is required that facilitate the mining of distributed data [14].

As for distributed operations, more data mining operations and algorithms will be required on the grid and to facilitate seamless integration of these resources into distributed data mining systems for complex problem solving, novel algorithms, tools, grid services and other IT infrastructure need to be developed.

3.3 Data Quality

The quality of data is a very important factor when considering solutions for managing data in support of performance of organisations. It is common that when acquiring and entering data, simple and complex errors can be committed. The errors in a large database may be due to a number of factors [16], [17], among them missing attribute values and corrupted vales.

To eliminate the errors, data cleaning methods and data analysis methods that can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases. Common data cleaning tasks include data acquisition and metadata, filling in missing values, unifying date formats, converting nominal to numeric values, identifying outliers and smooth out noisy data and correcting inconsistent data.

IV. APPLICATIONS OF DATA MINING

Data mining is now deployed in a wide range of fields including Health/healthcare/Insurance, Telecommunication, Corporate surveillance, Bioinformatics, Text Mining and Web Mining, Banking and Finance, Bibliomining, Customer Segmentation and Targeted Marketing, etc.

Health/Healthcare/Insurance: Medical data mining in healthcare is regarded as an important yet complicated task that needs to be executed accurately and efficiently. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases [16] [18], such as heart and liver diseases. Vikas [18] apply different methods of classifier techniques in the diagnosis of heart disease patients. The results show that bagging algorithm gives an accuracy of 85.03%, which makes it one of the most successful data mining techniques used in the diagnosis of heart diseases. In health insurance, data mining is applied in claims analysis such as identifying which medical procedures are claimed together. It also helps to forecast on the customers with potential to purchase new policies and those with fraudulent behavior.

Examples:

(i) NeuroMedical Systems used neural networks to perform a pap smear Diagnostic aid.

(ii) The University of Rochester Cancer Center and the Oxford Transplant Center use Knowledge SEEKER, a decision tree technology, to help with their Research

Telecommunication: Telecommunication companies routinely generate and store large amounts of data, have a large customer base and operate in a rapidly and highly competitive environment. One important feature of mobile telecommunication data is its association with spatiotemporal information. The most common areas of data mining application in telecommunication are broadly classified into 4 types, i) Telecommunication Fraud Detection ii) Telecommunication Churn Prediction iii) Network Fault Identification and Isolation and Marketing. In a study on churn prediction in mobile telecommunication systems using data mining techniques [19], Balasubramanian and Slvarani observed that decision tree model surpasses the neural network model in the prediction of churn and it is also easy to construct.

Examples:

(i) Coral Systems of Longmont, Colorado is a company that incorporates data mining techniques in their FraudBuster product which tracks down known types of fraud by modeling subscriber usage patterns and predicting when a carrier is suspected of fraud.

(ii) RightPoint Corporation focuses on data mining issues in the telecommunication industry, and in particular, customer retention or churn.

Text Mining and Web Mining: Text mining is the process 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 established. Using text mining however, we can easily derive certain patterns in the comments that may help identify a common set of customer perceptions not captured by the other survey questions. An extension of text mining is web mining. Web mining is an exciting new field that integrates data and text mining within a website. It enhances the web site with intelligent behavior, such as suggesting related links or recommending new products to the consumer [18].

Examples;

(i) Practical Text Mining using Perl.

(ii) PackMOLE (Mining Online Expert on Packaging Patents), a Text Mining tool, designed for mining patent information in the packaging field.

(iii) AUTINDEX- a commercial text mining software package based on sophisticated linguistics by IAI (Institute for Applied Information Sciences), Saarbrücken.

(iv) NetOwl - suite of multilingual text and entity analytics products, including entity extraction, link and event extraction, sentiment analysis, geotagging, name translation, name matching, and identity resolution, among others.

Banking and Finance: Data Mining is now widely used in banking and finance. It is mainly used for credit fraud prediction, risk evaluation and for analyzing trends and profitability. Neural networks have been used extensively in financial markets to forecast stock prices, bond rating, commodity price prediction as well as forecasting financial disasters.

Examples:

(i) Mellon Bank (USA) has used the data on existing credit - card customers to characterize their behavior and they try to predict what they will do next. Using IBM Intelligent Miner, Mellon developed a credit card - attrition model to predict which customers will stop using Mellon' s credit card in the next few months. Based on the prediction results, the bank can take marketing actions to retain these customers' loyalty.

Bibliomining (Data Mining in Libraries): Data mining techniques can help libraries in knowing the trends of popular subjects to enable better focus of acquisitions and budgets, analysis of usage, borrowing and interlibrary loan patterns to plan collection, time-of-day traffic to plan opening hours and staffing, etc. Bibliomining is the application of data mining and bibliometric tools to data produced from library services to aid decision-making and justify services. The bibliomining process [2] consists of: determining areas of focus; identifying internal and external data sources; collecting, cleaning, and anonym zing the data into a data warehouse; selecting appropriate analysis tools; discovery of patterns through data mining and creation of reports with traditional analytical tools; and analyzing and implementing the results.

Examples:

(i) Bibliomining on North South University (USA) library data **[20]**

V. CONCLUSION AND FUTURE IMPROVEMENTS

Over the years data mining has enjoyed tremendous success, the application areas expanded continuously but the mining techniques also kept up improving. Diverse problems have emerged and have been solved by data mining researchers. However, there are areas and issues that still require attention for future improvements in this technology. More research on how to deal with the social issue of sometimes, unknowing and unsuspecting individuals' privacy need to be conducted. In the paper, we identified the distributed data and operations as a big challenge. Data mining techniques should therefore evolve to match up with this challenge.

More work should be done on standardization of interaction languages to make it convenient to users, given that currently we have different mining tools with different syntaxes dealing with non-standard data types. From this survey I have established that different researchers concur that it is not easy to interpret the results of a data mining exercise. The existing software packages [21] lack sufficient support for both directing the analysis process and presenting the analysis results in a user understandable manner.

REFERENCES

[1] R. Agrawal, T. Imielinski and A. Swami, "Data Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering 5(6) (1993), 914–925.

[2] Shieh, Jiann-Cherng, " The Integration System for Librarians' Bibliomining", Asia- Pacific Conference on Library & Information Education & Practice, 2009

[3] Bharati M. Ramageri, "Data Mining Techniques and Applications": Indian Journal of Computer Science and Engineering Vol. 1 No. 4 301-305, 2010

[4] Kalyani et al., International Journal of Advanced Research in Computer Science and Software Engineering 2 (10), October- 2012, pp. 439-442

[5] Md. Adam Baba, Mohd Gouse Pasha, Shaik Althaf Ahammed, S. Nasira Tabassum, "Introduction to Neural Networks Design Architecture", International Journal of Scientific & Engineering Research Volume 4, Issue 2, February-2013 1 ISSN 2229-5518

[6] Manish Verma, Mauly Srivastava, Neha Chack, Atul Kumar Diswar, Nidhi Gupta," A Comparative Study of Various Clustering Algorithms in Data Mining," International Journal of Engineering Reserch and Applications (IJERA), Vol. 2, Issue 3, pp.1379-1384, 2012

[7] Aastha Joshian and Rajneet Kaur, "A Review: Comparative Study of Various Clustering Techniques in Data Mining", International Journal of Advanced Research in Computer Science and Software Engineering 3(3), March - 2013, pp. 55-57

[8] Dr. Maya Nayak and Er. Jnana Ranjan Tripathy: "Pattern Classification Using Neuro Fuzzy and Support Vector Machine (SVM) – A Comparative Study", *International* Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 5, May 2013.

[9] Jyoti Arora, Nidhi Bhalla, Sanjeev Rao, "A REVIEW ON ASSOCIATION RULE MINING LGORITHMS", International Journal of Innovative Research in Computer and Communication Engineering Vol. 1, Issue 5, July 2013.

[10] R. Agrawal, T. Imielinski, and A. Swami, "*Mining association rules between sets of items in large databases*". In Proc. of the ACM SIGMOD International Conference on Management of Data - *SIGMOD '93*. p. 207 Washington, D.C., May 1993.

[11] V. Uma, M. Kalaivany and G. Aghila, "Survey of Sequential Pattern Mining Algorithms and an Extension to Time Interval Based Mining Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering 3(12), December -2013, pp. 1178-1183

[12] S. V. S. GANGA DEVI, "A SURVEY ON DISTRIBUTED DATA MINING AND ITS TRENDS", International Journal of Research in Engineering & Technology (IMPACT: IJRET) ISSN(E): 2321-8843; ISSN(P): 2347-4599 Vol. 2, Issue 3, Mar 2014, 107-120

[13] Alaa H Al-Hamami and Suhad Abu Shehab, "An Approach for Preserving Privacy and Knowledge In Data Mining Applications", Journal of Emerging Trends in Computing and Information Sciences, Vol. 4, No. 1 Jan 2013, ISSN 2079-8407.

[14] Bhoj Raj Sharma, Daljeet Kaur and Manju, " A Review on Data Mining: Its Challenges, Issues and Applications", International Journal of Current Engineering and Technology, Vol.3, No.2 (June 2013).

[15] Dunren Che, Mejdl Safran, and Zhiyong Peng, "From Big Data to Big Data Mining: Challenges, Issues, and Opportunities", DASFAA Workshops 2013, LNCS 7827, pp. 1–15, 2013. © Springer-Verlag Berlin Heidelberg 2013.

[16] Vikas Gupta, "A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues", International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013 ISSN 2229-5518.

[17] Dileep Kumar Singh and Vishnu Swaroop, "Data Security and Privacy in Data Mining: Research Issues & Preparation", International Journal of Computer Trends and Technology- volume 4 Issue 2- 2013 ISSN: 2231-2803.

[18] Vikas Chaurasia and Saurabh Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739.

[19] M.Balasubramanian and M.Selvarani, "Churn Prediction In Mobile Telecommunication System Using Data Mining Techniques", International Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014 1 ISSN 2250-3153.

[20] North South University (USA) library

[21] Michael Goebel and Le Gruenwald, "A Survey of Data Mining and Knowledge Discovery Software Tools", SIGKDD Explorations, Volume 1, Issue 1, June 1999.

