ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

A State of the Art Review on Various Data Mining Techniques

S.Yamini [1], Dr.V.Khanaa [2], Dr.Krishna Mohantha [3]

Research scholar, Department of Information Technology, Bharath University, Selaiyur, Chennai, India[1]

Dean, Centre for Information, Bharath University, Selaiyur, Chennai, India[2]

Professor, Department of CSE, Sri Ramanujar College of Engineering, Chennai, India3

ABSTRACT: Data Mining is a powerful technique with immense potential in any field of science and technology. It's a useful tool to explore information within data which would be challenging to retrieve through queries and reports. It reduces the traditional methods of data analogy which would consume more time. Data mining has a huge stake of applications in various fields. Yet it requires many research works and actions to begin with this field of study. It's vital that one begins with steep understanding of the key concepts in mining. This paper aims at fulfilling that task by providing a state of the art detailed review about various techniques and sub techniques involved in data mining and the allied techniques.

KEYWORDS: Data Mining, Text classification, Clustering, Association etc…

I. INTRODUCTION

Data mining has been a key prodigy in the field of information engineering. It helps in extracting specific information from a huge data set. Eventually the objective of data mining would be prediction and Predictive data mining is an extensively accepted norm which has direct impact in business applications. This process consists of three major stages or steps; the initial exploration, Model building or pattern identification and deployment [1].

As far as a business scenario is concerned, data mining is utilized to detect patterns and liaisons in data for the sake of improved business solutions. It can aid in discovering business trends in sales, establishing agile strategies and authentically anticipating the requirements of a customer. Precise usages of data mining techniques in a business perspective are as follows [2];

Market segmentation –Analyze the familiar attributes of customers who purchase similar products from a company.
Customer churn - Anticipating which customers might leave a company and might go to a competitor.
Fraud detection - Spotting which transaction is most likely to be a counterfeit.
Direct marketing - Determine which keywords should be used in a mailing list to fetch the best response rate.
Interactive marketing - Finding what each individual surfing a Web site is most likely interested in viewing.
Market basket analysis – Determining on what commodity or benefits customers would usually acquire together.
Trend analysis –Acknowledge the change between a common customer currently and in previous purchases.

Since the applications oriented with data mining are ample we consider a specific field of study in erecting this review paper. We therefore consider the healthcare domain for our review purposes. In current scenarios data mining is becoming attractive in health care industry since there was a need of interpretive technique for distinguishing anonymous and beneficial advice from health data. In health care industry, Data Mining could afford several aids such as disclosure of the graft in health insurance, Opportunities of medical solution at a reduced cost, disclosure of origin for diseases and recognition of medication methods. It also helps researchers for scripting effective healthcare policies, composing drug suggestion systems, establishing health profiles of individuals etc. [3]

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

The rest of the paper is structured as follows. Section II depicts the…..

II. TECHNIQUES IN DATA MINING

A. Classification:

Classification divides data samples into target classes. The classification technique predicts the target class for each data points. For example, patient can be classified as "high risk" or "low risk" patient on the basis of their disease pattern using data classification approach. It is a supervised learning approach having known class categories. Binary and multilevel are the two methods of classification. In binary classification, only two possible classes such as, "high" or "low" risk patient may be considered while the multiclass approach has more than two targets for example, "high", "medium" and "low" risk patient. Data set is partitioned as training and testing dataset. Using training dataset we trained the classifier. Correctness of the classifier could be tested using test dataset. Classification is one of the most widely used methods of Data Mining in Healthcare organization. Hu et al. used different classification method such as decision tree, SVM and ensemble approach for analyzing microarray data [4]. This research work performed comparative analysis of above mentioned classification method using 10-fold cross validation approach on the data set obtained from Kent Ridge Bio Medical Dataset repository. The experiment results indicate that among all classification method ensemble achieved good accuracy [4]. Further use of classifier in medical field is discussed by Hatice et al., to diagnosis the skin diseases using weighted KNN classifier [5]. Breast cancer is one of the fatal and dangerous diseases in women. Potter et al. has performed experiment on the breast cancer data set using Weka tool and then analyze the performance of different classifier using 10-fold cross validation method [6]. The research work revealed that there is no single best algorithm which yields better result for every dataset. Classification techniques are also used for predicting the treatment cost of healthcare services which is increases with rapid growth every year and is becoming a main concern for everyone [7]. Bestsimas et al. used classification tree approach to predict the cost of healthcare [8] by using the dataset of 3 years collected from the insurance companies to perform the experiment. The first two year data was used to train the classifier and last one year data was used for comparing the predicted results of classifier. Following are the various classification algorithms used in healthcare:

- K-Nearest Neighbour (K-NN)
- Decision Tree (DT)
- Support Vector Machine (SVM)
- Neural Network (NN)
- Bayesian Methods

B. Regression

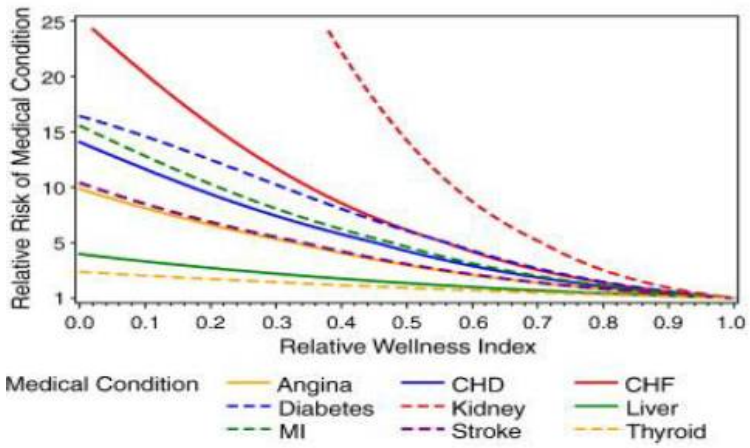
Regression is used to find out functions that explain the correlation among different variables. A mathematical model is constructed using training dataset. In statistical modeling two kinds of variables are used where one is called dependent variable and another one is called independent variable and usually represented using 'Y' and 'X'. There is always one dependent variable while independent variable may be one or more than one. Regression is a statistical method which investigates relationships between variables. By using Regression dependences of one variable upon others may be established [9]. Based on number of independent variables regression is of two types, one is Linear and another one is Non-linear. Linear regression identifies relation of a dependent variable and one or more independent variables. It is based on a model which utilizes linear function for its construction. Linear regression finds out a line and calculates vertical distances of points from the line and minimize sum of square of vertical distance. In this approach dependent and independent variables are already known and purpose is to spot a line that correlates between these variables [9]. But, linear regression is limited to numerical data only and cannot be use for categorical data. Logistic regression, a type of non-linear regression can accept categorical data and predicts the probability of occurrence using logit function. Logistic regression is of two types, one is Binomial and other is multinomial. Binomial regression predicts the result for a dependent variable when there occurs only two possible outcomes such as either a person is dead or alive while the multinomial handles the situation when dependent variable has three or more outcome. For example either a patient is at 'low risk', 'medium risk' and 'high risk'. Logistic regression does not consider linear relationship between variables [10]. Regression is widely used in medical field for predicting the diseases or survivability of a patient. Figure 1 represents an application of logistic regression for the estimation of relative risk for various medical conditions such as Diabetes, Angina, stroke etc [11]. In another research work, Weighted Support Vector Regression (WSVR) is used for monitoring the daily activities of patient [12]. This paper presents a model based on WSVR to overcome the over-fitting problem occurred due to noise and outliers.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016



C. Clustering

Clustering is an unsupervised learning method that is different from classification. Clustering is unlike to classification since it has no predefined classes. In clustering large database are separated into the form of small different subgroups or clusters. Clustering partitioned the data points based on the similarity measure [13]. Clustering approach is used to identify similarities between data points. Each data points within the same cluster are having greater similarity as compare to the data points belongs to other cluster. Various clustering techniques are established and used over the last few decades. As pointed out earlier clustering need less or no information for analyzing the data. So it is mainly used for analysing microarray data because very little details are available for genes. Tapia et al. analysed the gene expression data with the help of a new hierarchical clustering approach using genetic algorithm [14].

Partitioned Clustering: In this clustering method the datasets having 'n' data points partitioned into 'k' groups or clusters. Each cluster has at least one data point and each data point must belong to only one cluster. In this clustering approach there is a need to define the number of cluster before partitioning the datasets into groups. Based on the choice of cluster centroid and similarity measure, partition clustering method is divided into two categories-K-means and K-Mediods. K-Means clustering approach is one of the most widely used approach that partition the given 'n' data points into 'k' cluster based on similarity measure in such a way that data points belong to the same cluster have high similarity as compare to the data points of other cluster [15].

Hierarchical Clustering: Unlike partitioned clustering there is no need to define the number of cluster in advance. Hierarchical Clustering algorithm decomposes the data points in hierarchical way. It decompose the data points either using bottom up approach or top down approach. Hierarchical clustering is classified into two categories – Agglomerative and Divisive that depends on the decomposition process. Agglomerative approach initially consider each data point as a separate group and further it merges the data points that have some similarity with each other and repeat this process until all the data points are merged into one group or class or until it gets some termination condition [15].

Density Based Clustering: The problem with partition and hierarchical clustering method is that they can handle only spherical shaped cluster and are not suitable for discovering cluster of arbitrary shapes. Density clustering methods remove this drawback and efficiently handle outliers and arbitrary shaped cluster. DBSCAN and OPTICS are two approach of Density based clustering which discover cluster on the basis of density connectivity analysis. DENCLUE is another approach of density based clustering methods that form the grouping of data points on the basis of distribution value analysis of density function [15]. The research work [16] extracts the useful and interesting patterns from biomedical images using density based clustering. This research discovers the area of homogeneous color in biomedical images. This method separates the unhealthy skin or wound from healthy skin and discovers the sub regions of varied color or spotted part inside the unhealthy skin which is again useful for classification and association task [16].

Association: Association is one of the most vital approaches of data mining that is used to find out the frequent patterns, interesting relationships among a set of data items in the data repository. It is also known as market basket analysis due to its capability of discovering the association among purchased item or unknown patterns of sales of customers in a transaction database. For example if a customer is buying a computer then the chance of buying antivirus software is high. This information helps the storekeeper to further enhance their sales [17]. Association also

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

has great impact in the healthcare field to detect the relationships among diseases, health state and symptoms. Ji et al., used association in order to discover infrequent casual relationships in Electronic health databases.

The following table describes the summary of data mining approaches that are used in health domain:

Author	Publication Year	Approaches	Accuracy
Yan et al.	2003	Multilayer Perceptron	63.60%
Andreeva, P et al.	2006	Naïve Bayes	78.56%
		Decision Tree	75.73%
		Neural Network	82.77%
		Kernel Density	84.44%
	2008	Automatically Defined Groups	67.80%
		Immune Multi-agent Neural Network	82.30%
Sitar-Taut et al.	2009	Naïve Bayes	62.03%
		Decision Tree	60.40%
Chang et al.	2009	Decision Tree	90.89%
		Artificial Neural Network	92.62%
		Decision Tree combined with ANN	86.89%
		Decision Tree with sensitivity Analysis	80.33%
		ANN with sensitivity Analysis	83.61%
Rajkumar et al.	2010	Naïve Bayes	52.33%
		Decision tree	52%
		KNN	45.67%
Srinivas et al.	2010	Naïve Bayes	84.14%
		One Dependency Augmented Naïve Bayes classifier	80.46%
Kangwanariyakul, et al.	2010	Back-Propagation Neural Network	78.43%
		Bayesian Neural Network	78.43%
		Probabilistic Neural Network	70.59%
		Linear Support Vector Machine	74.51%

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

		Polynomial Support Vector Machine	70.59%
		RBF- kernel Support Vector Machine	60.78%
Anbarasi, et al.	2010	Genetic with Decision tree	99.20%
		Genetic with Naïve Bayes	96.50%
		Genetic with Classification via Clustering	88.30%
Fan et al.	2010	CHAID	69.75%
		C & RT	69.73%
		QUEST	67.25%
		C 5.0	71.17%
Sonali et al.	2010	One-against-many with POLY kernel	85.14%
		One-against-many with Gaussian kernel	95.98%
		M-SVM with polynomial kernel	83.25%
		M-SVM with Gaussian kernel	97.19%
Osareh et al.	2010	PNN	92.86%
		KNN	94.06%
		SVM-RBF	95.45%
		SVM-POLY	95.19%

II. APPLICATIONS IN HEALTH CARE

Data mining provides several benefits to healthcare industry. Data Mining helps the healthcare researchers to make valuable decision. Following are the several applications of Data Mining in healthcare:

A. Effective management of Hospital resource

Data mining provides support for constructing a model for managing the hospital resources which is an important task in healthcare. Using data mining, it is possible to detect the chronic disease and based on the complication of the patient disease prioritize the patients so that they will get effective treatment in timely and accurate manner. Fitness report and demographic details of patients is also useful for utilizing the available hospital resources effectively. An automated tool using data mining is proposed by J.Alapont et al., for managing hospital resources such as physical and human resources [90]. Group Health Cooperative provides various healthcare services at lower cost using data mining techniques [1]. It is a non-profit organization of healthcare that offers patients to online access their medical information, online fill the prescription form and allow safe exchanging of e-mail with the healthcare provider. Seton Medical centre also used data mining to enhance the healthcare quality, provide various details regarding patient's health and reduce admitted duration of the patients in the hospitals.

B. Hospital Ranking

Different data mining approaches are used to analyze the various hospital details in order to determine their ranks [93]. Ranking of the hospitals are done on the basis of their capability to handle the high risk patients. The hospital with higher rank handles the high risk patient on its top priority while the hospital with lower rank does not consider the risk factor.

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

C. Better Customer Relation

Data Mining helps the healthcare institute to understand the needs, preferences, behavior, patterns and quality of their customer in order to make better relation with them. Using Data Mining, Customer Potential Management Corp. develops an index represent the utilization of Consumer healthcare. This index helps to detect the influence of customer towards particular healthcare service.

D. Hospital Infection Control

A system for inspection is constructed using data mining techniques to discover unknown or irregular patterns in the infection control data [93]. Association rules are used to produce unexpected and interesting information from the public surveillance and hospital control data. To control the infection in the hospitals, this information is reviewed further by an Expert.

E. Smarter Treatment Techniques

Using Data Mining, physicians and patients can easily compare among different treatments technique. They can analyze the effectiveness of available treatments and find out which technique is better and cost effective. Data Mining also helps them to identify the side effects of particular treatment, to make appropriate decision to reduce the hazard and to develop smart methodologies for treatment.

F. Improved Patient care

Large amount of data is collected with the advancement in electronic health record. Patient data which is available in digitized form improve the healthcare system quality. In order to analyze this massive data, a predictive model is constructed using data mining that discover interesting information from this huge data and make decision regarding the improvement of healthcare quality. Data mining helps the healthcare providers to identify the present and future requirements of patients and their preferences to enhance their satisfaction levels.

G. Decrease Insurance Fraud

Healthcare insurer develops a model to detect the fraud and abuse in the medical claims using data mining techniques. This model is helpful for identifying the improper prescriptions, irregular or fake patterns in medical claims made by physicians, patients, hospitals etc. US taxpayers also reported to lost hundred dollars in 1997 due to fraudulent in the hospitals bill. ReliaStar financial corp. has improved the annual savings by 20% by detected the fraud and abuse. Doctor's prescriptions and treatment materials are produced large amount of data.

H. Recognize High-Risk Patients

American Healthways system construct a predictive model using data mining to recognize the patients having high risk. The main concern of this system is to handle the diabetic patients, improve their health quality and also offers cost savings services to the patient. Using Predictive model, healthcare provider recognize the patient which require more concern as compare to other patients [99].

I. Health Policy Planning

Data mining play an important role for making effective policy of healthcare in order to improve the health quality as well as reducing the cost for health services. COREPLUS and SAFS models were developed using data mining techniques to analyse the results of medical care services provided by hospitals and treatment cost.

IV. CHALLENGES

One of the most significant challenges of the data mining in healthcare is to obtain the quality and relevant medical data. It is difficult to acquire the precise and complete healthcare data. Health data is complex and heterogeneous in nature because it is collected from various sources such as from the medical reports of laboratory, from the discussion with patient or from the review of physician. For healthcare provider, it is essential to maintain the quality of data because this data is useful to provide cost effective healthcare treatments to the patients. Health Care Financing Administration maintains the minimum data set (MDS) which is recorded by all hospitals. In MDS there are 300 questions which are answered by the patients at check-in time. But this process is complex and patients face problem to respond the entire questions. Due to this MDS face some difficulties such as missing information and incorrect entries. Without quality data there is no useful results. For successful data mining, complication in medical data is one the significant hurdle for analyzing medical data. So, it is essential to maintain the quality and accuracy data for data mining to making effective decision. Another difficulty with healthcare data is data sharing. Healthcare organizations are unwilling to share their data due to privacy concern. Most of the patients do not want to disclose their health data. So, the Health Maintenance Organization and Health insurance Organization are not distributing their data for preserving the privacy of patient. This poses hurdle in the fraud detection studies in health insurance. The startup

ISSN(Online): 2319-8753
ISSN (Print): 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 3, March 2016

cost of data warehouse is very high. Before applying data mining techniques in healthcare data it is essential to collect and record the data from different sources into a central data warehouse which is a costly and time consuming process. Faulty data warehouse design does not contribute to effective data mining.

V. CONCLUSION

A key finding is that Data Mining is more oriented towards applications than the basic nature of the underlying phenomena. In other words, Data Mining is relatively less concerned with identifying the specific relations between the involved variables. For effective utilization of data mining in health organizations there is a need of enhance and secure health data sharing among different parties. Some propriety limitations such as contractual relationships among researcher and health care organization are mandatory to overcome the security issues. There is also a need of standardized approach for constructing the data warehouse. In recent years due to enhancement of internet facility a huge datasets (text and non-text form) are also available on website. So, there is also an essential need of effective data mining techniques for analyzing this data to uncover hidden information

REFERENCES

[1] http://www.statsoft.com/Textbook/Data-Mining-Techniques#mining
[2] http://www.laits.utexas.edu/~anorman/BUS.FOR/course.mat/Alex/
[3] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
[4] H. Hu, J. Li, A. Plank, H. Wang and G. Daggard, "A Comparative Study of Classification Methods For Microarray Data Analysis", Proc. Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS, vol. 61, pp. 33-37, (2006)
[5] C. Hattice and K. Metin, "A Diagnostic Software tool for Skin Diseases with Basic and Weighted K-NN", Innovations in Intelligent Systems and Applications (INISTA), (2012).
[6] R. Potter, "Comparison of classification algorithms applied to breast cancer diagnosis and prognosis", advances in data mining, 7th Industrial Conference, ICDM 2007, Leipzig, Germany, pp. 40-49, (2007) July.
[7] G. Beller, "The rising cost of health care in the United States: is it making the United States globally noncompetitive?", J. Nucl. Cardiol., vol. 15, no. 4pp. 481-482, (2008),.
[8] D. Bertsimas, M. V. Bjarnadóttir, M. A. Kane, J. C. Kryder, R. Pandey, S. Vempala and G. Wang, "Algorithmic prediction of health-care costs", Oper. Res., vol. 56, no. 6, pp. 1382-1392, (2008).
[9] J. Fox, "Applied Regression Analysis, Linear Models, and Related Methods", (1997).
[10] P. A. Gutiérrez, C. Hervás-Martínez and F. J. Martínez-Estudillo, "Logistic Regression by Means of Evolutionary Radial Basis Function Neural Networks", IEEE Transactions on Neural Networks, vol. 22, no. 2, pp. 246-263, (2011).
[11] C. Gennings, R. Ellis and J. K. Ritter, "Linking empirical estimates of body burden of environmental chemicals and wellness using NHANES data", http://dx.doi.org/10.1016/j.envint.2011.09.002,2011.
[12] Divya and S. Agarwal, "Weighted Support Vector Regression approach for Remote Healthcare monitoring", IEEE-International Conference on Recent Trends in Information Technology, ICRTIT 2011, 978-1-4577-0590-8/11/$26.00 ©2011 IEEE MIT, Anna University, Chennai, June 3-5(2011).
[13] M. Silver, T. Sakara, H. C. Su, C. Herman, S. B. Dolins and M. J. O'shea, "Case study: how to apply data mining techniques in a healthcare data warehouse", Healthc. Inf. Manage, vol. 15, no. 2pp. 155-164, (2001).
[14] J. J. Tapia, E. Morett and E. E. Vallejo, "A Clustering Genetic Algorithm for Genomic Data Mining", Foundations of Computational Intelligence, vol. 4 Studies in Computational Intelligence, vol. 204, pp. 249-275(2009).
[15] J. Han and M. Kamber, "Data mining: concepts and techniques", 2nd ed. The Morgan Kaufmann Series, (2006).
[16] M. E. Celebi, Y. A. Aslandogan and R. P. Bergstresser, "Mining Biomedical Images with Density-based Clustering", Proceedings of the International Conference on Information Technology: Coding and Computing (ITCC'05), (2005).
[17] R. Agrawal, T. Imielinski and A. N. Swami, "Mining Association Rules between Sets of Items in Large Databases. SIGMOD", vol. 22, no. 2, pp. 207-16, (1993) June.
[18] J. Alapont, A. Bella-Sanjuán, C. Ferri, J. Hernández-Orallo, J. D. Llopis-Llopis and M. J. Ramírez-Quintana, "Specialised Tools for Automating Data Mining for Hospital Management", http://www.dsic.upv.es/~abella/papers/HIS_DM.pdf, (2005).
[19] H. C. Koh and G. Tan, "Data Mining Application in Healthcare", Journal of Healthcare Information Management, vol. 19, no. 2, (2005).
[20] O. Mary K., Mat, "Application of Data Mining Techniques to Healthcare Data", Infection Control and Hospital Epidemiology, August (2004).
[21] M. Ridinger, "American Healthways uses SAS to improve patient care", DM Review, vol. 12, no.139, (2002).