RESEARCH ARTICLE OPEN ACCESS

Data Mining In Healthcare: A Survey of Techniques and Algorithms with Its Limitations and Challenges

Prakash Mahindrakar[1], Dr. M. Hanumanthappa[2]
[1]Research Scholar, Department of Computer Science and Applications, Bangalore University, Bengaluru, India
[2]Associate Professor, Department of Computer Science and Applications, Bangalore University, Bengaluru, India

ABSTRACT
The large amount of data in healthcare industry is a key resource to be processed and analyzed for knowledge extraction. The knowledge discovery is the process of making low-level data into high-level knowledge. Data mining is a core component of the KDD process. Data mining techniques are used in healthcare management which improve the quality and decrease the cost of healthcare services. Data mining algorithms are needed in almost every step in KDD process ranging from domain understanding to knowledge evaluation. It is necessary to identify and evaluate the most common data mining algorithms implemented in modern healthcare services. The need is for algorithms with very high accuracy as medical diagnosis is considered as a significant yet obscure task that needs to be carried out precisely and efficiently.

***Keywords* -** data mining, data mining algorithms in healthcare, KDD, knowledge discovery, healthcare

I. INTRODUCTION

The KDD is the process of making low-level data into high-level knowledge. Knowledge discovery has the Preprocessing, Data mining and Post processing phases. KDD is an iterative or cyclic process that involves sequence steps of processes and data mining is a core component of the KDD process. Data mining is a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database [1].

Data mining involves choosing the data mining task, choosing the data mining algorithm(s) and use of data mining algorithms to generate patterns. A data mining system may generate thousands of patterns. A discovered pattern can correspond to prior knowledge or expectations. A pattern can refer to uninteresting attributes or attribute combinations. Patterns can be redundant.

Healthcare industry today generates large amounts of complex data about patients, hospitals resources, disease diagnosis, electronic patient records, medical devices etc. These large amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining brings a set of tools and techniques that can be applied to this data to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions.

Data mining techniques are used in healthcare management for, Diagnosis and Treatment, Healthcare Resource Management, Customer Relationship Management and Fraud and Anomaly Detection. Data mining can help Physicians identify effective treatments and best practices, and Patients receive better and more affordable healthcare services.

Data mining algorithms are needed in almost every step in KDD process ranging from domain understanding to knowledge evaluation. It is necessary to identify and evaluate the most common data mining algorithms implemented in modern healthcare services. Determining performance of data mining solutions require much time and effort. Data mining algorithms may give in better results for one type of problems while others may be suitable for different ones. The need is for algorithms with very high accuracy as medical diagnosis is considered as a significant yet obscure task that needs to be carried out precisely and efficiently.

This paper is organized as follows: Section 2 discusses Knowledge discovery process and data mining. Section 3 discusses Data mining techniques. Section 4 discusses Data mining algorithms in healthcare services. Section 5 discusses Limitations and challenges of data mining algorithms in healthcare services and conclusion in section 6.

II. KNOWLEDGE DISCOVERY PROCESS AND DATA MINING

Knowledge discovery process involves identifying a valid, potentially useful structure in data. The KDD is the process of making low-level data into high-level knowledge. It is an iterative or cyclic process that involves the process steps of Selection, Pre-processing, Transformation, Data mining and Interpretation.

Selection step retrieve data relevant to the analysis task. Selection and integration of the target data from possibly many different and heterogeneous sources. The target dataset which is created in this selection will undergo analysis as data in the real world is incomplete, noisy and inconsistent.

In *Pre-processing step* the dataset which is selected during the selection step is pre-processed to handle databases which are highly susceptible to noisy, missing, and inconsistent data due to huge size, complexity and their likely origin from multiple heterogeneous sources. Data preprocessing select relevant data with respect to the data mining task in hand. Preprocessing tasks of data cleaning fill in missing values, smooth noisy data, identify or remove outliers, and resolve inconsistencies. Data integration task integrate multiple databases, files, etc.

Transformation step transform data into forms appropriate for mining by performing smoothing, generalization, normalization, aggregation, discretization and feature construction operations. Data reduction task obtains reduced representation in volume, but produces the same or similar analytical results.

Data mining is a core component of the KDD process. Data mining involves choosing the data mining task, data mining algorithm(s) and use of data mining algorithms to generate patterns. Data mining is defined as a process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database [1].

In *interpretation/evaluation step* the mined patterns and models are interpreted. The results are presented in understandable form. Post-processing methodology find all potentially interesting patterns. Evaluation of the results of data mining is done with statistical validation and significance testing. Different kinds of knowledge require different kinds of representation association, classification, clustering, etc. Clustering results can be shown in a graph or in a table etc. Visualization techniques are important for making the results useful.

III. DATA MINING TECHNIQUES

Data mining techniques are used in healthcare management for, diagnosis and treatment, healthcare resource management, customer relationship management and fraud and anomaly detection. Data mining uses two strategies; supervised and unsupervised learning. In supervised learning, a training set is used to learn model parameters whereas in unsupervised learning no training set is used. Unsupervised learning refers to modeling with an unknown target variable. The models are solely descriptive. The goal of the process is to build a model that describes interesting regularities in the data. Tasks of Data mining can be separated into descriptive and predictive. Descriptive tasks have a goal on finding human interpreted forms and associations, after reviewing the data and the entire construction of the model, prediction tasks tend to predict an outcome of interest [2].The prediction is one of the data mining techniques that discover relationship between independent variables and relationship between dependent and independent variables. The main predictive and descriptive data mining tasks can be classified as following:

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. It is the process of finding a function that allows the classification of data in one of several classes. Classification approach often makes use of decision tree or neural network-based classification algorithms. The accuracy of the classification rules are estimated using test data. Applicable techniques if the target attribute is categorical are decision tree induction, Bayesian classification, back propagation (neural network), based on concepts from association rule mining, k-nearest neighbor, case based reasoning, genetic algorithms, rough set theory, support vector machine and fuzzy set. If the target attribute is continuous then we use linear, multiple and non-linear regression [3].

Association rule analysis is descriptive data mining task which includes determining patterns, or associations, between elements in data sets. Associations are represented in the form of rules, or implications [2]. There are several classifications of these algorithms such as according to the types of values handled (e.g. Boolean and quantitative rules), according to the number of dimensions (e.g. single and multi dimensional) and according to the level of abstraction (e.g. single and multi level rules) [3].

Clustering can be considered as identification of similar classes of objects. Clustering techniques can find out overall distribution pattern and correlations among data attributes. Clustering various methods are applicable as partitioning methods (e.g. k-means, k-medoids), hierarchical methods (e.g. chameleon, CURE), density-based methods (e.g. DBSCAN, OPTIC), grid-based methods (e.g. STING, CLIQUE) and model-based methods (e.g. statistical and neural network approaches) [3].

IV. DATA MINING ALGORITHMS IN HEALTHCARE SERVICES

Data mining also recognized as Knowledge Discovery in databases is very frequently utilized in the field of medicine. The process of supporting medical diagnoses by automatically searching for valuable patterns undergoes evident improvements in terms of precision and response time [4]. Data mining algorithms are needed in almost every step in KDD process ranging from domain understanding to knowledge evaluation. Every data mining technique serves a diverse purpose depending on the modeling objective. The two most common modeling objectives are classification and prediction. Classification models predict categorical labels (discrete, unordered) while prediction models predict continuous-valued functions [5] Decision Trees and Neural Networks use classification algorithms while Regression, Association Rules and Clustering use prediction algorithms [6, 7]. Here are some of the data mining algorithms which are successfully used in healthcare.

4.1. Naïve Bayes - The Naïve Bayes is a simple probabilistic classifier. Naïve Bayes is based on the assumption of mutual independency of attributes. The algorithm works on the assumption, that variables provided to the classifier are independent. The probabilities applied in the Naïve Bayes algorithm are calculated using Bayes Rule [8, 9]. In 1980s Bayesian networks were introduced. Their first applications in medicine were revealed in 1990s. The Bayesian formalism is a way of representation of uncertainties what is essential during diagnosis, prediction of patients' prognosis and treatment selection [10]. It is possible to present the interactions among variables using Bayesian networks. These networks are often understood as cause-and-effect relationships.

The application of a Bayesian network in medicine was presented for instance in diagnosis and antibiotic treatment of pneumonia by P. Lucas [10]. In addition, the Naïve Bayes method's performance was tested against a colorectal cancer in [11] in which the authors enhanced the effectiveness of this method [4].

4.2. Decision Trees - Decision trees are one of the most regularly used techniques of data analysis [8]. Decision trees are easy to visualize and understand and resistant to noise in data [12]. Generally, decision trees are used to classify records to a proper class. Besides, they are applicable in both regression and associations tasks. In the medical field decision trees specify the sequence of attributes values and a decision that is based on these attributes [13, 4].

Decision Tree algorithms comprise CART (Classification and Regression Tree), ID3 (Iterative Dichotomized 3) and C4.5. These algorithms are at variance in selection of splits, when to stop a node from splitting, and assignment of class to a non-split node [14]. CART uses Gini index to measure the impurity of a partition or set of training tuples [5]. It can handle high dimensional categorical data. Decision Trees can also handle continuous data (as in regression) but they must be converted to categorical data [7]. The decision trees are effectively applied in medicine for instance in prostate cancer classification [15]. Here C4.5 algorithm was used. The article [16] presents a study carried out to create a decision tree model to describe how women in Taiwan make a decision whether or not to have a hysterectomy. The qualitative study was conducted and a tree model was built. This method, based on the Galwin's methodology, had accuracy of 90%. The problem of cervical cancer was described in [17]. It covered women all over the world. The study evaluated performance of four different decision tree techniques and Bayesian network. 10-fold cross validation was used for testing. The research with the use of CART for breast cancer analyses was presented in [18]. It confirms that decision trees are an important technique whose outcome should be compared with other methods [4].

4.2.1. ID3 - Quinlan introduced ID3 algorithm. This algorithm belongs to the family of decision tree. The algorithm is based on Occam's razor, which means that the smaller trees are preferred. The Occam's razor is formalized using information entropy concept. The construction of a tree is top-down and start with the appropriate attribute for the root node. The choice is tested and the procedure is repeated until all the attributes are used. The choice is based on calculating the entropy for each attribute. The entropy is a measure of information impurity. The ID3 uses an information gain as a measure of information carried by each of the attributes. The information gain measure is the reduction in entropy caused by the partition of the dataset [19, 9].

The ID3 searches an entire space of finite discrete-valued functions [19]. This helps to avoid the risk that the hypothesis space does not contain a target function. It does not backtrack to reconsider the former choices. It is a greedy algorithm. There is a threat of converging to locally optimal solutions that are not globally optimal. On the other hand the ID3 algorithm has a very significant advantage: it is less sensitive to errors as the decisions are based on all the instances not just the current one. It prefers short and small trees which have the attributes with the greatest information value closer to the root. The ID3 algorithm was effectively applied in supporting medical diagnosis. The authors of [20] made use of it in case of staging of cervical cancer [4].

4.2.2. C4.5 - The C4.5 algorithm is an expansion of ID3 algorithm [19]. The C4.5 algorithm is competent of handling continuous attributes, which are required in case of medical data. Other very common aspect missing values was also taken into consideration in C4.5. The C4.5 algorithm generates rules from a single tree. It can transfer multiple decision trees and create a set of classification rules [9].

The C4.5 algorithm is competent of handling continuous attributes, which are necessary in case of medical data (e.g. blood pressure, temperature, etc.). The worth of C4.5 algorithm was widely proven in medicine [15]. This algorithm suits medical data because it copes with missing values. The C4.5 algorithm handles continuous data which are common among medical symptoms. The efficiency of C4.5 was revealed in breast cancer and prostate cancer classification [15] to generate a decision tree and rules which may be helpful in medical diagnosing process [4].

4.3. Neural networks - Artificial neural networks are analytical techniques that are formed on the basis of superior learning processes in the human brain. As the human brain is capable to, after the learning process, draw assumptions based on earlier observations, neural networks are also capable to predict changes and events in the system after the process of learning. Neural networks are groups of connected input/output units where each connection has its own weight. The learning process is performed by balancing the net on the basis of relations that exist between elements in the examples. Based on the significance of cause and effect between certain data, stronger or weaker connections between "neurons" are being formed. Network formed in this manner is ready for the unknown data and it will react based on previously acquired knowledge [2]. One of the key advantages of Artificial Neural Networks is their high performance [19]. The core function of Artificial Neural Networks is prediction. The disadvantage of this method is its complexity and difficulty in understanding the predictions. Their effectiveness and usefulness was proven in medicine [21]. The successful implementation of the neural networks was in the development of novel antidepressants [22]. The notable success is the application of a neural network in coronary artery disease [23] and processing of EEG signals [24, 4].

4.4. Genetic algorithms - Genetic algorithms are based on the principle of genetic modification, mutation and natural selection. These are algorithmic optimization strategies inspired by the principles observed in natural evolution. Genetic algorithms are used in data mining to formulate hypotheses about the dependencies between variables in the form of association rules or other internal formalism [2]. Genetic algorithms are one of those key techniques for feature selection problems in medical informatics research.

V. LIMITATIONS AND CHALLENGES OF DATA MINING ALGORITHMS IN HEALTHCARE SERVICES

Medical diagnosis is considered as a significant yet obscure task that needs to be carried out precisely and efficiently. The need is for algorithms with very high accuracy because it is an issue of life or death. However powerful these data mining techniques are it has to be used with great care in the biomedical applications. Within the issue of knowledge integrity assessment, two biggest challenges are: (1) How to develop efficient algorithms for comparing content of two knowledge versions (before and after). This challenge require development of efficient algorithms and data structures for evaluation of knowledge integrity in the data set; and (2) How to develop algorithms for evaluating the influence of particular data modifications on statistical importance of individual patterns that are collected with the help of common classes of data mining algorithm. Algorithms that measure the influence that modifications of data values have on discovered statistical importance of patterns are being developed, although it would be impracticable to develop a universal measure for all data mining algorithms [25, 2].

To find out performance of data mining solutions a great deal of effort is dedicated to empirical studies. Some methods may yield improved results for one type of problems while others may be appropriate for different ones. It is necessary to uncover advantages and limitations of these methods as it helps in deciding appropriate algorithm. In determining performance of a method different problems are to be dealt like limited sample of data, difficulty in evaluating hypothesis's performance for unseen instances and how to use an available dataset both for training and testing. When estimating performance of an algorithm issue of bias and variance of an estimate need to be considered. To compare various data mining solutions different notions from statistics and sampling theory are utilized [19].The algorithms needs to be evaluated with complex experiments before application them in real medical system. Precision of the medical decisions would increase along with a decrease in the time spent for the diagnosing with the use of ideal algorithms. Each of the algorithms has some specific features applicable to different problems so it may be impracticable to find the best data mining algorithm suitable for all of the medical domains [4].

VI. CONCLUSION

This paper provided an overview of knowledge discovery process, data mining techniques and algorithms in healthcare services along with limitations and challenges of data mining algorithms in healthcare services. Data mining brings a set of tools and techniques that can be applied to the large amount of data in healthcare industry to discover hidden patterns that provide healthcare professionals an additional source of knowledge for making decisions. Data mining algorithms are needed in almost every step in KDD process ranging from domain understanding to knowledge evaluation. The need is for algorithms with very high accuracy as medical diagnosis is a significant task that needs to be carried out precisely and efficiently. It is necessary to identify and evaluate the most common data mining algorithms implemented in modern healthcare services as data mining algorithms may give in better results for one type of problems while others may be suitable for different ones. It may be impracticable to find the best data mining algorithm suitable for all of the medical domains.

REFERENCES

[1] Fayyad, U, Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases, *Proc. of the 9th Int. Conf. on Scientific and Statistical Database Management, Olympia, Washington*, USA, 2-11, 1997.

[2] Boris Milovic, Milan Milovic, Prediction and

Decision Making in Health Care using Data Mining , *International Journal of Public Health Science (IJPHS) Vol. 1, No. 2,* December 2012, pp. 69~78 ISSN: 2252-8806

[3] Shaker H. El-Sappagh, Samir El-Masri, A. M. Riad, Mohammed Elmogy, Data Mining and Knowledge Discovery: Applications, Techniques, Challenges and Process Models in Healthcare, *International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 Vol. 3, Issue 3,* May-Jun 2013, pp.900-906

[4] Kamila Aftarczuk, *Evaluation of selected data mining algorithms implemented in Medical Decision Support Systems,* Thesis no: MSE-2007-21, September 2007, School of Engineering, Blekinge Institute of Technology, Sweden.

[5] Han, J., Kamber, M., *Data Mining Concepts and Techniques* (Morgan Kaufmann Publishers, 2006).

[6] Charly, K., Data Mining for the Enterprise, *31st Annual Hawaii Int. Conf. on System Sciences, IEEE Computer,* 7, 295-304, 1998.

[7] V. Krishnaiah, G. Narsimha & N. Subhash Chandra, A Study on Clinical Prediction using Data Mining Techniques, *International Journal of Computer Science Engineering and Information Technology Research (IJCSEITR) ISSN 2249-6831 Vol. 3, Issue 1,* Mar 2013, 239-248 TJPRC Pvt. Ltd.

[8] Nong Y., *The Handbook of Data Mining* (Lawrence Earlbaum Associates, 2003)

[9] N. Abirami, T. Kamalakannan, Dr. A. Muthukumaravel, A Study on Analysis of Various Data mining Classification Techniques on Healthcare Data, *International Journal of Emerging Technology and Advanced Engineering ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 3, Issue 7,* July 2013.

[10] Lucas P.J.F., Boot H., Taal B.G., *Decision-theoretic network approach to treatment management and prognosis* (Knowledge-based Systems, 1998, vol. 11) 321–330

[11] Sarkar M., XinZhi Q., Peng-Kheong L., Tze-Yun L., Performance of existing prognostic factors for colorectal cancer prognosis - a critical view, *Engineering in Medicine and Biology Society, 2001 vol. 4, 3909-3912*

[12] Witten I. H., Frank E., *Data Mining Practical Machine Learning Tools and Techniques* (2nd Elsevier, 2005)

[13] Lavrac N., *Selected techniques for data mining in medicine* (Artificial Intelligence in Medicine 1999, vol. 16) 3-23

[14] Ho, T. J., *Data Mining and Data Warehousing* (Prentice Hall, 2005).

[15] Tahir M.A., Bouridane A., Novel Round-Robin Tabu Search Algorithm for Prostate Cancer Classification and Diagnosis Using Multispectral Imagery, *IEEE Transactions on Information Technology in Biomedicine*, 2006, 782-793.

[16] Shu-Mei W., Yu C., Yu-Mei, Cheng-Fang Y., Hui-Lian C., Decision-making tree for women considering hysterectomy, *Journal of advanced nursing, Blackwell Publishing*, 2005, 361-368.

[17] Jorng-Tzong H., Kai-Chih H., Li-Cheng W., Hsien-Da H., Horn-Cheng L., Ton-Yuen C., Identifying the combination of genetic factors that determine susceptibility to cervical cancer, *Bioinformatics and Bioengineering*, 2004, 325-330

[18] Land W.H. Jr., Verheggen E.A., Experiments using an evolutionary programmed neural network with adaptive boosting for computer aided diagnosis of breast cancer, *Soft Computing in Industrial Applications*, 167-172, 2003

[19] Mitchell T. M., *Machine Learning* (Redmond, McGraw-Hill, 1997)

[20] Pabitra M., Sushmita M., Sankar P, *Evolutionary Modular MLP with Rough Sets and ID3 Algorithm for Staging of Cervical Cancer,* Springer, 2001, vol. 10, 67-76

[21] Tsymbal A., Bolshakova N., Guest Editorial Introduction to the Special Section on Mining Biomedical Data, *IEEE Transactions On Information Technology In Biomedicine*, 2006, vol. 10, no. 3, 425-428.

[22] Lesch K.P., Serotonergic gene expression and depression: implications for developing novel antidepressants, *Journal of Affective Disorders, 2001, vol. 62, 57-76.*

[23] Liping A. and Lingyun T., A rough neural expert system for medical diagnosis, *Services Systems and Services Management*, 2005, vol. 2, 1130-1135

[24] Kutlu, Y., Isler, Y., Kuntalp, D., Kuntalp, M., Detection of Spikes with Multiple Layer Perceptron Network Structures, *Signal Processing and Communications Applications,* 2006, 1-4

[25] Yang, Q., & Wu, X. (2006), 10 Challenging problems in data mining research, *International Journal of Information Technology & Decision Making Vol. 5, No. 4* , 597–604.