Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey of Data Mining Techniques on Medical Data for Finding Locally Frequent Diseases

Mohammed Abdul Khaleel* *Research Scholar Sambalpur University, India*
Sateesh Kumar Pradham *P.G.Department of Computer Science Utkal University, India*
G.N. Dash *P.G.Department of Physics Sambalpur University, India*

Abstract - In the last decade there has been increasing usage of data mining techniques on medical data for discovering useful trends or patterns that are used in diagnosis and decision making. Data mining techniques such as clustering, classification, regression, association rule mining, CART (Classification and Regression Tree) are widely used in healthcare domain. Data mining algorithms, when appropriately used, are capable of improving the quality of prediction, diagnosis and disease classification. The main focus of this paper is to analyze data mining techniques required for medical data mining especially to discover locally frequent diseases such as heart ailments, lung cancer, breast cancer and so on. We evaluate the data mining techniques for finding locally frequent patterns in terms of cost, performance, speed and accuracy. We also compare data mining techniques with conventional methods.

Keywords - Data mining, frequent patterns, data mining techniques, medical data mining

I. Introduction

Data mining is the process of digging data for discovering latent patterns which can be translated into valuable information. Data mining usage witnessed unprecedented growth in the last few years. Of late the usefulness of data mining techniques has been realized in Healthcare domain. This realization is in the wake of explosion of complex medical data. Medical data mining can exploit the hidden patterns present in voluminous medical data which otherwise is left undiscovered. Data mining techniques which are applied to medical data include association rule mining for finding frequent patterns, prediction, classification and clustering. Traditionally data mining techniques were used in various domains. However, it is introduced relatively late into the Healthcare domain. Nevertheless, as on today lot of research is found in the literature. This has led to the development of intelligent systems and decision support systems in Healthcare domain for accurate diagnosis of diseases, predicting the severity of various diseases, and remote health monitoring. Especially the data mining techniques are more useful in predicting heart diseases, lung cancer, and breast cancer and so on.

The data mining techniques that have been applied to medical data include Apriori and FPGrowth [1], [2], [3], [4], [5], [6], [7], and [8], unsupervised neural networks [9][10], linear genetic programming [9], Association rule mining [11], [12], Bayesian Ying Yang [13], decision tree algorithms like ID3, C4.5, C5, and CART [14], [15], [16], [17], [18], [19], [20], outlier prediction technique [21], Fuzzy cluster analysis [22], classification algorithm [17], [23], [24], Bayesian Network algorithm [14], [25], Naive Bayesian [26], combination of K-means, Self Organizing Map (SOM) and Naïve Bayes [27], Time series technique [28], [29], combination of SVM, ANN and ID3 [16], clustering and classification [30],SVM [16], [31], FCM [29],k-NN [24], and Bayesian Network [14]. This paper provides the summary of all these techniques in terms of the problem they solve or the their utility in medical data mining or the tools which are implemented over them and so on.

The remainder of the paper is structured as follows. Section II reviews literature pertaining to data mining and applications of data mining techniques in Healthcare domain. Section III provides the summary of medical data mining techniques. Section IV presents the importance of discovering locally frequent patterns or trends or diseases in medical data. Section V concludes the paper.

II. Related Works

Data mining is a process of analyzing voluminous data in various perspectives in order to bring about trends or patterns that lead to business intelligence [32]. Data mining plays an important role in IT as it discovers knowledge from historical data of various domains. For instance data mining can be used to mine medical data as Healthcare domain produces huge amount of data about patients, diseases, diagnosis, medicine and so on. By applying data mining techniques in Healthcare domain, the administrators can improve the QoS (Quality of Service) by discovering latent potentially useful trends required by medical diagnosis [33]. Data mining is useful in medical applications such as medications, medical tests, prediction of surgical procedures, and discovery of relationships between pathological data and clinical data [34]. Apriori and FPGrowth are the most widely used frequent pattern mining algorithms [35]. These two algorithms and algorithms based on them are studied in [2], [3], [4], [5], [6], [7], and [8]. These two algorithms are also used in medical data mining. Goodwin et al. [36] applied data

mining techniques for birth outcomes. Evans et al. [37] stated that hereditary syndromes can be detected automatically using data mining techniques. Doron Shalvi and Nicholas DeClaris, [10] discussed medical data mining through unsupervised neural networks besides a method for data visualization. They also emphasized the need for preprocessing prior to medical data mining. In the year 2000 Krzysztof J. Cior [38], bioengineering professor, identified the need for data mining methods to mine medical multimedia content. Tsumoto [39] identified problems in medical data mining. The problems include missing values, data storage with respect to temporal data and multi-valued data, different medical coding systems being used in Hospital Information Systems (HIS). Brameier and Banzhaf [9] explored and analyzed two programming models such as neural networks, and linier genetic programming for medical data mining. Abidi and Hoe [40] proposed and implemented a symbolic rule extraction workbench for generating emerging rule-sets. Abidi et al. [41] explored the usage of rule-sets as results of data mining for building rule-based expert systems. Olukunle and Ehikioya [11] proposed an algorithm for extracting association rules from medical image data. The association rule mining discovers frequently occurring items in the given dataset. Shim and Xu [13] proposed a classification method based on Bayesian Ying Yang (BYY) which is a three layered model. They applied this model to classify liver disease through automatic discovery of medical trends.

Brunie et al. [42] proposed architecture for mining geno-medical data in heterogeneous and grid-based distributed infrastructures. Mahmud Khan et al. [15] focused on decision tree data mining algorithm for medical image analysis. Especially they studied on lung cancer diagnosis through classification of x-ray images. Podgorelec et al. [21] presented an outlier prediction method for improving performance of classification as part of medical data mining. Wang et al. [22] applied fuzzy cluster analysis for medical images. They used decision tree algorithm to classify mammography into normal and abnormal cases. Cheng et al. [17] applied classification algorithm to diagnose cardio vascular diseases. For classification effectiveness they focused on two feature extraction techniques namely automatic feature selection and expert judgment.

Seng et al. [43] introduced web based data mining for the application of telemedicine. Ghannad-Rezaie et al. [44] presented an approach to integrate PSO rule mining methods and classifier on patient dataset. They used Particle Swarm Optimization technique as well. The results revealed that, their approach is capable of performing surgery candidate selection process effectively in epilepsy. Bethel et al. [12] developed an association rule learner which is based on the criteria collected from past breast cancer patients. The rule learner is used in a tool by name "Clinical Trial Assignment Expert System". Xue et al. [25] proposed and applied Bayesian Network algorithm for diagnosis of an ailment known as Coronary Heart Disease (CHD). Abraham et al. [26] proposed discrimination techniques to improve the accuracy of classification of medical data using Naive Bayesian classifier algorithm.

Hassan and Verma [27] proposed a hybrid approach for classification of medical data which combines K-means, Self Organizing Map (SOM) and Naïve Bayes with NN based classifier. Tsumoto [45] studied multi-stage medical diagnosis using experts' diagnostic rules and diagnostic taxonomy. They focused on automatic grouping of medical knowledge extracted from clinical database. Berlingerio et al. [28] studied Time Annotated Sequences (TAS) algorithm for mining medical data with temporal dimensions. The extracted patterns exhibited the attribute relationships in time domain which helps in accurate diagnosis. Xing et al. [16] developed data mining techniques for predicting the probability of survival of CHD patients. To achieve this they combined three prediction models such as SVM (Support Vector Machine), Artificial Neural Networks (ANN), and Decision trees using C4.5 or ID3, CART and C5. Abe et al. [46] proposed an integrated time-series data mining environment for mining huge amount of medical data for extracting more valuable rule-sets.

Jiquan et al. [47] proposed a framework known as term-mapping to combine multiple medical data sources for data mining. Barnathan et al. [30] presented a framework for clustering, classification and similarity search of biomedical images or 2D and 3D in nature. Shusaku et al. [48] proposed multi-scale matching and clustering technique on medical data. Their results revealed that their technique is capable of grouping hepatitis data based on temporal covariance of choline esterase, albumin and platelet. Hai Wang, and Shouhong Wang [49] studied on the role of medical experts in medical data mining. Medical experts can give expert advice that can be used as input in medical data mining. Abdullah et al. [1] applied apriori algorithm for medical data mining. They extracted frequent item sets by analyzing associations between treatments and diagnosis. Saraee et al. [18] applied data mining techniques to medical data pertaining to military with respect to mortality rate in children due to accidents. They used CART algorithm to generate a decision tree. Balakrishnan and Narayanaswamy [31] presented feature selection using SVM for classifying diabetes databases. Drugs and health effects are mined by Froelich and Wakulicz-Deja [29] using adaptive FCM (Fuzzy Cognitive Maps). Their work has led to improved decision support and planning in Healthcare domain.

Pradhan and Prabhakaran [50] proposed an approach through association rule mining to mine high-dimensional, time series medical data for discovering high confidence patterns. Karegowda and Jayaram [23] proposed a model to classify diabetic database using two techniques in cascading fashion for classification accuracy. The techniques are known as Correlation based Feature Selection (CFS) and Genetic Algorithm (GA). CHAO and WONG [19] proposed a decision tree learning methodology which could interpret attributes in medical data classification for higher accuracy when compared with Incremental Tree Induction (ITI) algorithm. TANG and TSENG [24] studies three classifiers for medical data mining. They are weighting fuzzy k-NN, fuzzy k-NN, and crisp k-NN to classify diabetic and cancer datasets. Tu et al. [20] proposed an intelligent medical decision support system which provides diagnosis of heart diseases through decision tree algorithm C4.5 and bagging algorithm Naïve Bayes. Su et al. 2011 [14] explored three techniques namely Back Propagation Network (BPN), C4.5 (decision tree algorithm), and Bayasian Network (BN) for mining medical databases. Hogl [51] introduced a language

known as Knowledge Discover Question Language for preparing questions that are used to discover knowledge from medical data. They explored ways and means for intelligent medical data mining.

III. Summary of Techniques for Medical data Mining

Data mining techniques have shown significant improvement in medical industry in terms of prediction and decision making with respect to various diseases like cancer, cardio vascular abnormalities, diabetes, and others. Table 1 summary the medical data mining, its areas of application and the utility of the techniques.

TABLE 1 – Summary of medical data mining techniques

REFERENCES	TECHNIQUES	UTILITY	DISEASE
[1], [2], [3], [4], [5], [6], [7], and [8]	Appriori and FPGrowth	Association rule mining for finding frequent item sets (diseases) in medical databases.	
[9], [10]	Neural Networks	Extracting patterns, detecting trends	
[9],	Genetic Algorithm	Classification of medical data.	Diabetic Diseases
[11], [12]	Association Rule Mining	Finding frequent patterns	
[13]	Bayesian Ying Yang (BYY)	Classification	Liver diseases
[14], [15], [16], [17], [18], [19], [10]	Decision Tree Algorithms such as ID3, C4.5, C5, and CART.	Decision Support	
[21]	Outlier Prediction Technique	For improving classification accuracy	
[22]	Fuzzy cluster analysis	Analyzing medical images	
[17], [23], [24]	Classification Algorithm	Disease classification	Cardio Vascular Diseases
[14], [25]	Bayesian Network algorithm	Modeling and analysis of medical data	Coronary Heart Disease
[26]	Naive Bayesian	Improving classification accuracy.	Coronary Heart Disease
[27]	Combined use of K-means, SOM and Naïve Bayes	Accurate Classification of medical data.	
[28], [29]	Time Series Technique	Medical diagnosis	
[16]	combination of SVM, ANN and ID3	Medical data classification	
[30]	Clustering and classification	Clustering and classification of biomedical databases	
[16], [31]	SVM	Disease Classification	Diabetes
[29]	Fuzzy Cognitive Maps	Drugs and Health effects classification	
[24]	k-NN	Classification of diseases	Diabetes, Cancer

IV. Conclusion

In this paper we survey various data mining techniques that have been employed for medical data mining. Data mining techniques have higher utility in medical data mining as there is voluminous data in this industry. Due to the rapid growth of medical data, it has become indispensable to use data mining techniques to help decision support and predication systems in the field of Healthcare. The medical mining yields required business intelligence to support well informed diagnosis and decisions. This paper has provided the summary of data mining techniques used for medical data mining besides the diseases they classified. It also throws light into the importance of locally frequent patterns and the mining techniques used for the purpose.

References

[1] Umair Abdullah (2008). Analysis of Effectiveness of Apriori Algorithm in Medical Billing Data Mining1. *IEEE*.p1-5.

[2] Cong-Rui Ji and Zhi-Hong Deng. (n.d). Mining Frequent Ordered Patterns without Candidate Generation. *IEEE*. 0 (0), P1-5.

[3] Hai-Tao He and Shi-Ling Zhang. (2007). A New method for Incremental Updating Frequent patterns mining. *IEEE*. 0 (0), p1-4.

[4] Carson Kai-Sang Leung∗ Christopher L. Carmichael and Boyu Hao. (2007). Efficient Mining of Frequent Patterns from Uncertain Data. *IEEE*. 0 (0), p489-494.

[5] Shariq Bashir, Zahid Halim, A. Rauf Baig. (2008). Mining Fault Tolerant Frequent Patterns using Pattern Growth Approach. *IEEE*. 0 (0), p172-179.

[6] Sunil Joshi and Dr. R. C. Jain. (2010). A Dynamic Approach for Frequent Pattern Mining Using Transposition of Database. *IEEE*. 0 (0), p498-501.

[7] Thanh-Trung Nguyen. (2010). An Improved Algorithm for Frequent Patterns Mining Problem. *IEEE*. 0 (0), p503-507.

[8] Xiaoyong Lin and Qunxiong Zhu. (2010). Share-Inherit: A novel approach for mining frequent patterns. *IEEE*. 0 (0), p2712-2717.

[9] Markus Brameier and Wolfgang Banzhaf. (2001). A Comparison of Linear Genetic Programming and Neural Networks in Medical Data Mining. *IEEE*.p1-10.

[10] Doron Shalvi and Nicholas DeClaris., (n.d). An Unsupervised Neural Network Approach to Medical Data Mining Techniques. *IEEE*. 0 (0), p1-6.

[11] Adepele Olukunle and Sylvanus Ehikioya, (n.d). A Fast Algorithm for Mining Association Rules in Medical Image Data. *IEEE*. p1-7.

[12] Cindy L. Bethel and Lawrence O. Hall and Dmitry Goldgof (n.d). Mining for Implications in Medical Data. *IEEE*. p1-4.

[13] Jeong-Yon Shim, Lei Xu (n.d). MEDICAL DATA MINING MODEL FOR ORIENTAL MEDICINE VIA BYY BINARY INDEPENDENT FACTOR ANALYSIS. *IEEE*. p1-4.

[14] Jenn-Lung Su, Guo-Zhen Wu, I-Pin Chao (2001). THE APPROACH OF DATA MINING METHODS FOR MEDICAL DATABASE. *IEEE*. p1-3.

[15] Safwan Mahmud Khan Md. Rafiqul Islam Morshed U. (n.d). Medical Image Classification Using an Efficient Data Mining Technique. *IEEE*, p1-6.

[16] Yanwei Xing, Jie Wang and Zhihong Zhao (2007). Combination data mining methods with new medical data to predicting outcome of Coronary Heart Disease. *IEEE*. p1-5.

[17] Tsang-Hsiang Cheng, Chih-Ping Wei, Vincent S. Tseng (n.d). Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches . *IEEE*. p1-6.

[18] Mohammad Saraee, George Koundourakis, Babis Theodoulidis. (n.d). EASYMINER: DATA MINING IN MEDICAL DATABASES. *IEEE*. p1-3.

[19] SAM CHAO, FAI WONG, "AN INCREMENTAL DECISION TREE LEARNING METHODOLOGYREGARDING ATTRIBUTES IN MEDICAL DATA MINING". Proceedings of the Eighth International Conference on Machine Learning and Cybernetics, Baoding, 12-15 July 2009.

[20] My Chau Tu AND Dongil Shin (2009). A Comparative Study of Medical Data Classification Methods Based on Decision Tree and Bagging Algorithms. *IEEE*. p1-5.

[21] Vili Podgorelec, Marjan Heri□koMaribor, (n.d). Improving Mining of Medical Data by Outliers Prediction. *IEEE*. p1-6.

[22] Shuyan Wang Mingquan Zhou Guohua Geng (n.d). Application of Fuzzy Cluster Analysis for Medical Image Data Mining. *IEEE*. p1-6.

[23] Asha Gowda Karegowda M.A.Jayaram (2009). Cascading GA & CFS for Feature Subset selection in Medical Data Mining. *IEEE*. p1-4.

[24] Graduate Institute of Applied Information Sciences (2009). MEDICAL DATA MINING USING BGA AND RGA FOR WEIGHTING OF FEATURES IN FUZZY K-NN CLASSIFICATION. *IEEE*. p1-6.

[25] Weimin Xue, Yanan Sun, Yuchang Lu (n.d). Research and Application of Data Mining in Traditional Chinese Medical Clinic Diagnosis. *IEEE*.p1-4.

[26] Ranjit Abraham, Jay B.Simha, Iyengar (n.d). A comparative analysis of discretization methods for Medical Datamining with Naïve Bayesian classifier. *IEEE*. p1-2.

[27] Syed Zahid Hassan and Brijesh Verma,(n.d). A Hybrid Data Mining Approach for Knowledge Extraction and Classification in Medical Databases. *IEEE*. p1-6.

[28] Michele Berlingerio (n.d). Mining Clinical Data with a Temporal Dimension: a Case Study. *IEEE*. p1-8.

[29] Wojciech Froelich, Alicja Wakulicz-Deja (2009). Mining Temporal Medical Data Using Adaptive Fuzzy Cognitive Maps. *IEEE*. P1-8.

[30] Michael Barnathan, Jingjing Zhang, Vasileios (n.d). A WEB-ACCESSIBLE FRAMEWORK FOR THE AUTOMATED STORAGE AND TEXTURE ANALYSIS OF BIOMEDICAL IMAGES. *IEEE*. p1-3.

[31] Sarojini Balakrishnan (n.d). SVM Ranking with Backward Search for Feature Selection in Type II Diabetes Databases. *IEEE*. p1-6.
[32] Arun K Pujari "Data Mining Techniques", Edition 2001.
[33] M. Ilayaraja Department of Computer Science & Engineering Alagappa University Karaikudi, India ilayarajaalu@gmail.com. (2013). Mining Medical Data to Identify Frequent Diseases using Apriori Algorithm. *IEEE*. 0 (0), p1-6.
[34] J. C. Prather, D. F. Lobach, L. K. Goodwin, J. W.Hales , M. L. Hage, W. Edward Hammond, "MedicalData Mining: Knowledge Discovery in a Clinical DataWarehouse", 1997.
[35] HAI-BING MA, JIN ZHANG, YlNG-JIE FAN, YUN-FA W. (2004). MINING FREQUENT PATTERNS BASED ON IS+-TREE. *IEEE*. 0 (0), P1208-1213.
[36]. Goodwin L, Prather J, Schlitz K, Iannacchione My Hammond W, Grzymala J, DataMining Issues for Impproved Birth Outcomes, Biomed. Science Instrum, 34, 1997, pp. 291-296.
[37]. Evans S, Lemon S, Deters C, Fusaro R and Lynch H, Automated Detection of hereditary Syndromes Using Data Mining, Computers and Biomedical Research 30, 1997, pp. 337-348.
[38] **Krzysztof** J. **Cior** , Medical Data Mining and Knowledge Discovery. (n.d). From the guest Editor. *IEEE*. p1-2
[39] Shusaku Tsumoto (n.d). Problems with Mining Medical Data. *IEEE*. p1-2.
[40] Syed Sibte Raza Abidi Kok Meng (n.d). Symbolic Exposition of Medical Data-Sets: A Data Mining Workbench to Inductively Derive Data-Defining Symbolic Rules. p1-6.
[41] S. S. R. Abidi, K. M. Hoe, A. Goh, "Analyzing data clusters: A rough set approach to extract cluster definingsymbolic rules, Fisher, Hand, Hoffman, Adams (Eds.) Lecture Notes in Computer Science: Advances inIntelligent Data Analysis, 4th Intl. Symposium, IDA-01. Springer Verlag: Berlin, 2001.
[42] Lionel Brunie, Maryvonne Miquel, Jean-Marc Pierson, and Anne Tchounikine, "Information grids: managing and mining semantic data in a grid infrastructure; open issues and application to geno-medical data. 2003, 14th International workshop on Database and Expert Systems Applications.
[43] Wong Kok Seng, Rosli Bin Besar, Fazly Salleh Abas trosli, (n.d). Collaborative Support for Medical Data Mining in Telemedicine. *IEEE*. p1-6.
[44] M. Ghannad-Rezaie, H. Soltanain-Zadeh, M.-R. Siadat, K.V. Elisevich. (2006). Medical Data Mining using Particle Swarm Optimization for Temporal Lobe Epilepsy. *IEEE*. p1-8.
[45] Shusaku Tsumoto , (n.d). Problems with Mining Medical Data. *IEEE*. p1-2.
[46] Hidenao Abe AND Hideto Yokoi (n.d). Developing an Integrated Time-Series Data Mining Environment for Medical Data Mining. *IEEE*. p1-6.
[47] Liu Jiquan Deng Wenliang Xudong Lu (n.d). Liu Jiquan Deng Wenliang Xudong Lu Huilong Duan College of Biomedical Engineering & Instrument Science Zhejiang University Hangzhou 310027, China liujiquan@gmail.com. *IEEE*. p1-4.
[48] Shusaku Tsumoto (n.d). Problems with Mining Medical Data. *IEEE*. p1-2.
[49] Hai Wang, Shouhong Wang 1. (n.d). Medical Knowledge Acquisition through Data Mining. *IEEE*. 0 (0), p1-4.
[50] Gaurav N. Pradhan AND B. Prabhakaran (n.d). ASSOCIATION RULE MINING IN MULTIPLE, MULTIDIMENSIONAL TIME SERIES MEDICAL DATA. *IEEE*. p1-4.
[51] Oliver Hogl, Michael Müller (2001). On Supporting Medical Quality with Intelligent Data Mining. *IEEE*. p1-10.

