A Survey on Medical Data Classification

Sharmilla.J[1] Sumathi V.P[2]
[1]Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]Kumaraguru College of Technology, Coimbatore-641049, India

Abstract— Data mining techniques are used on medical data for discovering patterns that are used in diagnosis and decision making. Data mining techniques such as clustering, classification are widely used in healthcare domain. But in this paper, classification of medical data is analyzed, as clustering is highly sensitive to noise and outliers. Data mining algorithms are capable of improving the quality of prediction, diagnosis and disease classification. This paper is about the survey of classifying the medical data using various feature selection and classification method. The missing values in data are handled. The main focus of this paper is to analyze data mining techniques required for medical data mining to discover the diseases such as heart ailments, lung cancer, breast cancer, etc, and to compare the evaluation metrics attained through different techniques for finding locally frequent patterns.

Key words: Data Mining, Medical Data Classification

I. INTRODUCTION

Data mining is the process of analyzing the data from different perspectives and summarizing it into useful information. It is the mining of knowledge from large amounts of data. Data mining techniques have been applied to medical services in many areas including prediction of effectiveness of surgical procedures, performance evaluation and for the diagnosis of disease.

Medical data classification is the process of converting the descriptions of medical diagnoses into universal medical code numbers. The diagnoses are usually taken from a variety of sources within the health care record, such as the transcription of physician's notes, laboratory results, radiologic results, and other sources. Those codes are used to track the diseases and other health issues. The data collected in medicine is generally collected as a result of patient-care activity. Usually the medical data will be noisy, inconsistent with missing values. So these data must be free of noise filled with appropriate values. Then these data must be classified using different classification methods.

A. Flow Graph:

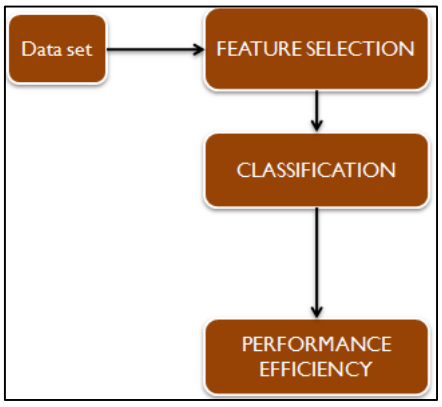

Fig. 1: Flow Graph

II. LITERATURE REVIEW

Murat Karabatak [2015] proposed weighted Naïve Bayesian classifier to detect the breast cancer. Several experiments were conducted to evaluate the performance of the weighted Naïve Bayesian on the breast cancer database. Experiments were realized with 5-cross fold validation test. The metrics namely accuracy, specificity, sensitivity were evaluated. The pros of this paper are that the performance of the weighted Naïve Bayesian classifier is better than regular Naïve Bayesian classifier. This algorithm uses grid search mechanism to find the optimum weight values. But the disadvantage of this search is computationally expensive and the initialization of the weights vector is crucial and application dependent. The accuracy obtained in this paper is 98.5%.

Wei-Chang Yeh et al. [2009] developed an efficient hybrid data mining approach to separate from a population of patients who have and do not have breast cancer. The proposed data mining approach consists of two phases. In first phase statistical method is used to preprocess and in second phase Discrete Particle Swarm Optimization is used. In this phase each particle is coded in positive integer numbers and has a feasible system structure. The main advantage of this Discrete Particle Swarm Optimization can improve the accuracy, sensitivity and scpecificity at a higher rate. It can also be used as the reference for making decision in hospital. The accuracy obtained here is 98.71%.

Thanh Nguyen et al [2015] introduced an automated medical data classification method using wavelet transform (WT) and interval type-2 fuzzy logic system (IT2FLS). Wavelet coefficient serves as an input to the IT2FLS. The integration between WT and IT2FLS aims to cope with both high-dimensional data challenge and uncertainty. The IT2FLS utilizes a hybrid learning process comprising unsupervised structure learning by the FCM clustering and supervised parameter tuning by genetic algorithm. The application of WT reduces the computational burden and enhances the performance of IT2FLS making this paper beneficial. Experiments are taken from the dataset from UCI machine library. Although the Genetic algorithm diminishes computational burden of the original SAM in the supervised learning process, the entire learning process of GSAM still requires a relatively large amount of time compared to PNN, SVM which is consider to be the weakness of this paper. The accuracy obtained here is 96%.

Thanh Nguyen et al. [2014] proposed an integration of fuzzy standard additive model (SAM) with genetic algorithm (GA), called GSAM, to deal with uncertainty and computational challenges. Wavelet transformation is employed to extract discriminative features for high-dimensional datasets. The strength of this paper is that GSAM becomes highly capable when deployed with small number of wavelet features as its computational

burden is reduced. The proposed method is evaluated using medical datasets: the Wisconsin breast cancer and Cleveland heart disease from the UCI Repository for machine learning. Experiments are organized with a five-fold cross validation and performance of classification techniques are measured by a number of important metrics: accuracy, F-measure, mutual information and area under the receiver operating characteristic curve. The negative side is that GSAM requires a relatively larger amount of time compared to SVM. The accuracy obtained here is 78.78%.

MRI has been a widely-used method of high quality medical imaging, especially in brain imaging. Classification is an important part in retrieval system. The classifications of brain MRI data as normal and abnormal are important to prune the normal patient and to consider only those who have the possibility of having abnormalities or tumor. This step was done by Noramalina Abdullah et al. [2015] using support vector machine (SVM). The aim of this paper is to compare percentage of accuracy in classification data with and without the implementation of principal component analysis (PCA). As a result, we found that by using PCA method, the number of feature vector has been reduced and increased the percentage of accuracy. The cons are SVM doesn't work precisely for a large data due to the training complexity of SVM is highly dependent on the size of data. The accuracy obtained here is 85%.

The existing system identifies the disease using particular type of classification techniques. A comparison of existing methodologies is given below:

III. COMPARISON

S.No	Author	Year	Title	Methods	Accuracy Obtained	Pros	Cons
1	Murat Karabatak	2015	A new classifier for breast cancer detection based on Naïve Bayesian and wavelets	WNB classifier	96%	Optimum weight values found easily by GA	This search is computationally expensive and the initialization of the weights vector is crucial
2	Wei-Chang Yeh	2009	A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method	DPSO	98.5%	Improved in accuracy and robustness, reduced computational cost.	Can be used just as a reference for decision making.
3	Thanh Nguyen,	2015	Medical data classification using interval type 2 fuzzy logic system and wavelets	WT, IT2FLS	78.78%	The fuzzy system helps to handle the noisiness, uncertainity and complexity of the medical data.	Difficulty in selecting the features
4	Thanh Nguyen	2014	Classification Of Healthcare Data Using Genetic Fuzzy Logic System And Wavelets	GA, WT	85%	It has lower computational costs and higher efficiency.	Time consumption
5	Noramalina Abdullah	2015	Improvement of MRI Brain Classification using Principal Component Analysis	SVM, PCA	98.71%	PCA can be used to reduce the number of feature vectors and lead to improve the percentage of accuracy	SVM doesn't work precisely for a large data. SVM is highly dependent on the size of data.

Table 1: Comparison

IV. CONCLUSION

Medical data classification is used to build more powerful information systems in healthcare. It can also support the healthcare process for transmission, re-using and sharing patient's data. The need for medical data classification is that it is highly helpful in decision making. Using this medical classification, patterns can be recognized and classified in multivariate patient attributes. From the above survey it is clearly known that disease diagnosis cannot be done perfectly without the help of medical data classification.

REFERENCES

[1] Murat Karabatak, "A new classifier for breast cancer detection based on Naïve Bayesian", 2015, pp 32-36.
[2] Wei-Chang Yeh, Wei -Wen Chang, Yuk Ying Chung " A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method", 2009 pp 8204-8211
[3] Thanh Nguyen, Dougla Creighton, "Medical data classification using interval type 2 fuzzy logic system and wavelets", 2015, pp 812-822.

[4] Thanh Nguyen, Abbas Khosravi, Douglas Creighton, Saied Nahavandi, "Classification Of Healthcare Data Using Genetic Fuzzy Logic System And Wavelets", 2014, pp 2184- 2197.
[5] Noramalina Abdullah, Lee Wee Chuen, Umi Kalthum Ngah, Khairul Azman Ahma , "Improvement of MRI Brain Classification using Principal Component Analysis", 2015 pp 557-561
[6] Ashutosh Malhotra, Erfan Youesi, Michaela G€undel, Bernd M€uller, Michael T.Heneka, Martin Hofmann-Apitius "ADO: A disease ontology representing the domain knowledge specific to Alzheimer's disease", 2014, pp 238-246.
[7] Audrey Baneyx, Jean Charlet, Marie-Christine Jaulent "Building an ontology of pulmonary diseases with natural language processing tools using textual corpora", 2007, pp 208-215.
[8] Yu-Liang Chi, Tsang-Yao Chen, Wan-Ting Tsai "A chronic disease dietary consultation system using OWL-based ontologies and semantic rules", 2015, pp 208-219.
[9] Prabath Chaminda Abeysiriwardana , Saluka R. Kodituwakku "Ontology based information Extraction for disease intelligence" 2012, pp 7-19.
[10] Vijay N. Garla, Cynthia Brandt "Ontology-guided feature engineering for clinical text classification" 2012, pp 992-998.

