International Journal of Research in Engineering and Science (IJRES)
ISSN (Online): 2320-9364, ISSN (Print): 2320-9356
www.ijres.org Volume 1 Issue 2 ǁ June 2013 ǁ PP.45-49

Deriving High Quality Information for Knowledge Discovery in Medical Systems by Structuring Text

Sofia Sunny [1], Paul P Mathai [2]
[1](Computer Science and Engineering, Federal Institute of Science and Technology, Angamaly)
[2](Computer Science and Engineering, Federal Institute of Science and Technology, Angamaly)

ABSTRACT : *The paper proposes a new approach for knowledge discovery in medical systems by deriving relevant information from text .This approach is a natural representation and makes diagnosis easier. The approach is to extract high quality data from an unstructured text. This paper would be very helpful for researchers involved in medical diagnosis.*

Keywords - *Information extraction, knowledge discovery, medical diagnosis, neural networks, text mining.*

I. Introduction

Medical diagnosis is an important area of research which helps to identify the occurrence of a disease. Medical data is an ever-growing source of information. The diagnostic procedure contains classification. Diagnosis is said to be a cognitive process. Naming the disease, symptoms, dysfunction or disability are different forms of diagnosis.

Data mining means mining knowledge from large data. It is used to search the patterns that exist in large databases [1].Prediction and analysis plays an important role in data mining [2]. Data mining methods that can be applied to extract knowledge from existing data set and using these extracted pattern the diseases are diagnosed very well. It can be used to predict class labels and the data is classified based on these class labels and training set. Thus it is an unavoidable part of data mining [3].

Text mining is an important area in medical diagnosis. Medical data is an unstructured text of information .So text mining is used to retrieve the useful and relevant information from the unstructured text. This makes diagnosis of diseases easier. The purpose is to retrieve high quality information from the unstructured text that is the medical data and then using this information for medical diagnosis. The LAMSTAR network, Naïve classifier, differential diagnosis and text mining are combined together to make diagnosis easier and faster.

II. Related Works

Bayesian networks (BN) plays a vital role in medical diagnostics. A set of random variables and the conditional dependencies among them are represented using this network .The network can compute the probabilities of the occurrence of various diseases when the symptoms are given[4].Bayesian networks are directed acyclic graphs which contains nodes and edges representing random variables and conditional dependencies. A BN represents joint probability distribution. The JPD can be computed by taking the probability of each possible combination of values for all variables used. The probabilities computed are stored in joint probability table (JPT).To reduce the number of probabilities stored in the table the conditional independence is incorporated. The BN can be used to identify the variables that are conditionally independent [5].The BN has two tasks which is represented by a pair B = {G, P} where G is the directed acyclic graph which contain nodes and edges. The nodes are connected using the arrows. The P represents conditional probability distribution [2]. If arcs are not present between nodes means they are conditionally independent. Domain expert knowledge of BN can be used in the knowledge discovery process. Due to the use of nodes and arrows BN are easily understandable when compared to other techniques. Graphical diagrams are used to represent the domain expert knowledge with which the output of network can be obtained. It is used to represent knowledge with uncertainty and efficient reasoning. The main advantage is that BN uses prior knowledge. These are the main strengths of Bayesian Network.BN contain both quantitative and qualitative part. For proper operation BN needs quantitative information. To fill the numeric information the statistics are not available so indirect statistics are required [4].

The k-nearest neighbor (k-NN) is used for classifying objects in the feature space. The k-nearest neighbour is the simplest of all classifiers. Majority vote of neighbours are considered for classification of the objects. Class that matches higher among the k nearest neighbours is selected and the object gets assigned to it. The k-NN performs well with large number of classes [3].The data is contained in a feature space. To find the distance among the points Euclidean distance is used [2]. The k is used to determine the number of neighbours

with which classification can be obtained. An integer k, a training set and metric to measure closeness are certain requirements for a k-NN. This classification can be used in medical diagnosis to classify diseases into subgroups. Preference is given to that subgroup that matches the given symptoms. Then searching is done in that subgroup thereby reducing the database access [6].Implementation of k-NN mechanism is easy. The debugging process is very faster. Learning is faster and neighbour points can be determined fastly. Increased value for k reduces the noisy points in the training data set. Classifier mechanism can be improved by incorporating noise reduction techniques. A small value for k increases the noise level. Large value for k makes it computationally expensive and can also increase the computational complexity and time [14].As all process is done at run time it is comparatively a slow technique.

LAMSTAR network can be used for large memory storage and retrieval in several medical diagnosis problems. Self-organizing maps are the basic processing module of this network. The neurons are stored in separate SOM modules. Correlation links between individual neurons are used to store information. Neurons are interconnected horizontally and vertically by arrays of link weights. Input word consists of subset of subwords. Each subword is given to corresponding SOM module.SOM stores data concerning that subword. The network is to find a neuron from a class that is best matching with the given input pattern[7].Information in the SOM module is stored in the form of weight .When an input word is presented to the network it checks weight stored in the SOM module. The weight is updated with a match. If no match found it creates a new pattern in the corresponding SOM module. LAMSTAR network plays an important role in medical diagnosis. The entered and original symptoms are compared. This information is given to the LAMSTAR network and then the weight assigning takes place. Learning in LAMSTAR can be improved by increasing case experience. To avoid rigidity in its decisions a degree of stochasticity is involved. It is faster in assigning of weights. Due to the link forgetting capability it gives more priority to the recent weight. Network efficiency can be increased by this link forgetting capability that is it avoids the need of large number of links. These are the main advantages of LAMSTAR network [6].The network does not allow the rare events a role in decision making. It does not incorporate confidence measure with which comparison across different network inputs can be made. This comparison can help user to choose the best solution [8].

ANN consists of a group of artificial neurons that are interconnected. They are used for representing complicated relationships between inputs and outputs. It is also called Neural Network [3].The network is used to assign patients to one of the classes of diseases. It can be used in decision making process and with this diagnosis of diseases can be done. So ANN is a powerful tool to disease diagnosis [9].ANN has different layers. The first layer contains input neurons that send data through a connection called synapses. Each of these connections has a weight associated with it. This weight is the coefficient of connectivity. The data is send to the second layer of neurons via synapses and then to the output neurons. The output layer gives the prediction or the result. More complicated network has more number of layers. Every neuron has a threshold value. When the weight sum is larger than threshold the neuron is activated and this activation signal generates the output.ANN has three parameters that is the interconnection pattern, learning process to update weight and the activation function [11].ANN is good in identifying diseases. The network does not need any details of how to recognize as it learns by example [10].Has good capacity that is the ability to model any given function. Hardware implementation is possible and is easy to maintain [3].The network continue to work even though an element failure occurs.ANN has good computational power thus it has good accuracy. There are some disadvantages for ANN that is difficulty in incorporating knowledge of a given problem. Neural network behaves like a black box that is there is no knowledge of internal workings. So it is difficult to interpret the network solution [12].

Back propagation is a common method to train artificial neural networks. It is a supervised learning method. The network without a feedback is called feedforward neural networks. These networks are widely used for classification [10]. The network contains three layers that is the input, hidden and output layer. The information flows from input layer through hidden layer to the output layer. As the information flows forward there will be no loops or cycles. The connection between the layers has weights and is trained using back propagation learning algorithm. The weight is adjusted by propagating error between network outputs and correct diagnosis backward through the network [24]. Feedforward neural networks are widely used in diagnosis of diseases. The hidden layer of the network requires transfer function to process information from input layer and this processed information is send to the output layer. It works better in real world applications and requires only small amount of training data [18].The result of the network converges to local minimum and this convergence is very slow in this network and it is not guaranteed too. Weight assigning is comparatively slow here [6].Parameters that allows prior learning is required and is also less sensitive to training sets [19].

Support Vector Machine (SVM) are supervised learning methods that can be used for classification.SVM takes input data and predicts to which of the class the corresponding input belongs. It contains a feature space and the input is considered as points in this feature space. These points are separated using the hyperplane in such a way that points of one class is on one side of hyperplane and other class on other

side of the hyperplane [20].The hyperplane with maximum minimum distance to a point in any of the class can achieve better separation. Margin of separation for a hyperplane is the closest data point [21].Support vectors are points that are closer to this separation hyperplane.SVM uses kernel function in case of non linearly separable samples. The kernel function transforms data to high dimensional space where it is linearly separable [22]. It is simple and is easy to be analyzed mathematically [23].Generalization performance is good and it delivers a unique solution. It is widely used in many applications such as classification and prediction. The choice of kernel is a big limitation. Speed and size is another limitation of SVM. There is a shortage of transparency of results in SVM.

III. Importance of Naïve Bayes and Structuring of Text in Medical Diagnosis

Medical data is an ever-growing source of information. It is generated in the form of patient records from hospitals. Medical diagnosis is done to predict whether a disease is present or not. The probability of occurrence of a particular disease is evaluated. The LAMSTAR network, k-NN and differential diagnosis are combined together to increase the accuracy. The most probable disease is predicted using the pattern matching along with k-NN classification and the next probable disease is obtained by performing differential diagnosis using LAMSTAR network. The advantage of using LAMSTAR is its link forgetting capability [6].The implementation of k-NN is easy but there is some difficulty in using k-NN. Increased value for k reduces the noisy points in the training data set. Large value for k makes it computationally expensive and can also increase the computational complexity and time [14].As all process is done at run time it is comparatively a slow technique.

The better enhancement to this is the use of naive bayes classifier. Naive bayes algorithm outperforms most of the sophisticated algorithms. It is a good tool in medical diagnosis. For example given a list of symptoms, it predicts occurrence of a disease [13], [14].NB assumes its attributes to be conditionally independent [16].The classifier computes the probability of each attribute in a class. The result of the classification is the class with the highest posterior probability. Posterior probability is proportional to the product of prior probability and likelihood. NB's main strength is its simplicity, efficiency and good classification performance. It combines efficiency with good accuracy. Due to its good accuracy it is used in medical diagnosis. A small amount of training data is required for the estimation of variable values necessary for classification. NB is a very powerful technique in diagnosing diseases [14].It is used to provide efficient output with attributes independent to each other. The NB classifier needs a very large training set to obtain good results.

Medical data are unstructured text. So text mining is incorporated to retrieve relevant information from this unstructured text. Text represents information in complex, difficult and rich manner. The aim is to convert text into data with which the analysis can be done [25]. It contains phases such as preprocessing, information extraction, feature generation, feature weighting and operation on feature space [26]. Text mining can be used in wide variety of applications. It can be used in medical diagnosis for making the diagnosis of disease easier. So the time taken to search the entire text can be reduced by extracting only the important information from a text. For example extracting relevant information such as symptoms from the medical data and then a comparison is done between the extracted symptoms and entered symptoms to see whether a disease is present or not. Thus it makes diagnosis easier.

IV. Conclusion

Medical diagnosis is to determine the presence of a disease. The LAMSTAR network is fast in assigning weights when compared to back propagation. With LAMSTAR the speed and accuracy of medical diagnosis can be improved [6].Naive bayes classifier outperforms all other sophisticated algorithms and is a best tool for medical diagnosis [13]. Structuring of text is done to obtain relevant data from text. Medical diagnosis can be done in an efficient manner by combining LAMSTAR network, differential diagnosis, naive bayes classification and text mining. With this the diagnosis of disease is easier and also the time consumption can be reduced.

V. Acknowledgement

The authors would like to thank Mr. Arun Kumar for the expertise and medical domain guidance and would also like to convey special thanks for his constant support.

References

Journal Papers:

[1]. Geetika , *A Survey of Classification Methods and its Applications*, International Journal of Computer Applications *(0975 – 8887) Volume 53– No.17,* September 2012.

[3]. Ms. Aparna Raj, Mrs. Bincy G, Mrs. T.Mathu, *Survey on Common Data Mining Classification Techniques*, International Journal of Wisdom Based Computing, *Vol. 2(1),* April 2012.

[4]. Daniel Nikovski , *Constructing Bayesian Networks for Medical Diagnosis from Incomplete and Partially Correct Statistics*, IEEE Transactions on Knowledge and Data Engineering, *Vol. 12, No. 4,* July/August 2000.

[5]. Antonella Meloni, Andrea Ripoli, Vincenzo Positano, *Member, IEEE*, and Luigi Landini , *Mutual Information Preconditioning Improves Structure Learning of Bayesian NetworksFrom Medical Databases*, IEEE Transactions on Information Technology in Biomedicine, *Vol. 13, No. 6,* November 2009.

[6]. Rahul Isola, Student Member, IEEE, Rebeck Carvalho, Student Member, IEEE,and Amiya Kumar Tripathy, Member, IEEE, *Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and k-NN*, IEEE Transactions on Information Technology in Biomedicine, *Vol. 16, No. 6*, November 2012.

[7]. Hubert Kordylewski, Daniel Graupe, Fellow, IEEE, and Kai Liu , *A Novel Large-Memory Neural Network as an Aid in Medical Diagnosis Applications*, IEEE Transactions on Information Technology in Biomedicine, *Vol. 5, no. 3,* September 2001.

[8]. Nathan C. Schneider,Daniel Graupe , *A Modified LAMSTAR Neural Network and its Applications*, International Journal of Neural Systems, *Volume 18, Issue 04*, August 2008.

[9]. Arti Gupta, Maneesh Shreevastava , *Medical Diagnosis using Back propagation Algorithm*, International Journal of Emerging Technology and Advanced Engineering,*ISSN 2250-2459, Volume 1, Issue 1*, November 2011.

[10]. Qeethara Kadhim Al-Shayea , *Artificial Neural Networks in Medical Diagnosis*, IJCSI International Journal of Computer Science Issues, *Vol. 8, Issue 2*, March 2011.

[11]. Abraham, A. , *Meta-Learning Evolutionary Artificial Neural Networks*, Neurocomputing Journal, *Vol. 56c, Elsevier Science, Netherlands, (1–38),*2004.

[12]. H.R.A. Cardon, R.van Hoogstraten , *Key Issues for Successful Industrial Neural Network Applications:an Application in Geology*, Springer.

[13]. Abid Sarwar ,Vinod Sharma , *Intelligent Naïve Bayes Approach to Diagnose Diabetes Type-2*, Special Issue of International Journal of Computer Applications (0975 – 8887) on Issues and Challenges in Networking, Intelligence and Computing Technologies – ICNICT 2012, November 2012.

[15]. R. Bhuvaneswari,K. Kalaiselvi , *Naive Bayesian Classification Approach in Healthcare Applications*, International Journal of Computer Science and Telecommunications,*Volume 3, Issue 1*, January 2012.

[17]. Shadab Adam Pattekari,Asma Parveen , *Prediction System For Heart Disease using Naive Bayes*, International Journal of Advanced Computer and Mathematical Sciences ISSN 2230-9624. *Vol 3, Issue 3,* 2012, pp 290-294.

[18]. Bekir KARLIK , *Hepatitis Disease Diagnosis Using Backpropagation and the Naive Bayes Classifiers,* Journal of Science and Technology,*Volume : 1 , Number : 1* , Year :2011.

[23]. Shaikh Abdul Hannan,V. D. Bhagile, R. R. Manza, R. J. Ramteke , *Diagnosis and Medical Prescription of Heart Disease Using Support Vector Machine and Feedforward Backpropagation Technique*, (IJCSE) International Journal on Computer Science and Engineering *Vol. 02, No. 06,* 2010, 2150-2159.

[24]. Alexander Statnikov, Constantin F. Aliferis, Ioannis Tsamardinos,Douglas Hardin, Shawn Levy ,*A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis, Vol. 21 no. 5 2005, pages 631–643.*

[25]. Andreas Holzinger,Regina Geierhofer,Felix Mödritscher,Roland Tatzl Semantic, *Information in Medical Information Systems:Utilization of Text Mining Techniques toAnalyze Medical Diagnoses*, Journal of Universal Computer Science, *vol. 14, no. 22 (2008), 3781-3795.*

[26]. J.Sreemathy, P. S. Balamurugan , *An Efficient Text Classification Using KNN AND Naive Bayesian*, International Journal on Computer Science and Engineering (IJCSE), *ISSN : 0975-3397, Vol. 4 No. 03* March 2012.

Proceedings Papers:

[2]. Thair Nu Phyu , *Survey of Classification Techniques in Data Mining*, Proceedings of the International Multiconference of Engineers and Computer Scientists, march 18 - 20, 2009, Hong Kong.

[14]. Hardik Maniya,Mosin I. Hasan,Komal P. Patel ,*Comparative study of Naïve Bayes Classifier and KNN for Tuberculosis*, International Conference on Web Services Computing (ICWSC) Proceedings published by International Journal of Computer Applications® (IJCA) 2011.

[16]. Diana Dumitru ,*Prediction of recurrent events in breast cancer using the Naive Bayesian classification*, Annals of University of Craiova, Math. Comp. Sci. Ser.*Volume 36(2)*, 2009, Pages 92-96,*ISSN: 1223-6934.*

[19]. Zhenghao Shi,Lifeng He ,*Application of Neural Networks in Medical Image Processing*, *Proceedings* of the Second International Symposium on Networking and Network Security (ISNNS '10) Jinggangshan, P. R. China, 2-4, April. 2010, pp. 023-026.

[20]. Yu-Len Huang,Kao-Lun Wang,Dar-Ren Chen,*Diagnosis of breast tumors with ultrasonic texture analysis using support vector machines*, Neural Comput & Applic,2006 *15: 164–169.*

[21]. Shashikant Ghumbre, Chetan Patil, Ashok Ghatol ,*Heart Disease Diagnosis using Support Vector Machine*,International Conference on Computer Science and Information Technology (ICCSIT'2011) Pattaya Dec. 2011.

[22]. Sumit Bhatia, Praveen Prakash,G.N. Pillai ,*SVM Based Decision Support System for Heart Disease Classification with Integer-Coded Genetic Algorithm to Select Critical Features*, *P*roceedings of the World Congress on Engineering and Computer Science WCECS 2008, October 22 - 24, 2008, San Francisco, USA.