ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(22), DOI: 10.17485/ijst/2016/v9i22/93828, June 2016* ISSN (Online) : 0974-5645

Biomedical Text Mining for Diagnosing Diseases - A Review

R. Priya[1]* and R. Padmajavalli[2]

[1]Research and Development Centre, Bharathiar University, Coimbatore - 641046, Tamil Nadu, India; rpriyaphd@gmail.com

[2]Department of Computer Applications, Bhaktavatsalam Memorial College for Women, Chennai - 600080, Tamil Nadu, India; padmahari2002@yahoo.com

Abstract

Diagnosis of diseases is a difficult work that has to do in accurate manner. Text mining deals a great job in this field. A huge mass of data is available in biomedical field, using this data we can diagnosis many diseases by text mining techniques in efficient manner. Text mining methods are used to retrieve useful knowledge from large data. **Objective:** The aim of this paper is to review several text mining methods used in biomedical field. This survey is helpful to select a best text mining method for biomedical data. **Methods/Analysis:** In this paper, classification method is used to study the biomedical text mining for diagnosing diseases. In the field of biomedical, classification can be done on the basis of patient disease pattern to separate the patients into high risk or low risk The classification techniques have two methods they are Binary contains two classes and multilevel contains more than two classes. Classification method is widely used in biomedical text mining. In this paper different classification techniques can be applied to categorize the text they are SVM (Support Vector Machine) NN (Neural Network), K-NN (K-Nearest Neighbor), Bayesian Method and DT (Decision Tree). **Findings:** In this paper, different classification techniques were surveyed and their merits and limitations have been discussed. The various classification techniques were applied in medical data where useful patterns and knowledge were extracted. The important task is that to select the suitable data and classification method for disease diagnosis. The objective of this survey is that how the classification methods are applied in biomedical application and to select which method is suitable and efficient for diagnosis of a particular disease. **Novelty/Improvement:** The main advantage of the survey is that it can be applied to any kind of dataset, it is a description dataset or not. For future improvement, we will implement our proposed methodology on using some major chest diseases datasets and measured performance in terms of training time and accurate diagnosis.

Keywords: Biomedical Text Mining, Classification, Concept Linkage, IE (Information Extraction), Topic Tracking

1. Introduction

Biomedical research refers to the study of the medical issue and problems using biological methodologies, including basic medical research and clinical medical research[1]. Biomedical text mining can make easier to begin the process of discovery and to integrate the data present in the biomedical literature. Bioinformatics translator has been prominent with integrating biological and clinical data. The main view of this research is focused on biomedical text mining, aimed at correlating diseases and molecular entities[2].

Data mining is used to point out the hidden information in biomedical data and correct differentiate pathological from normal data. It can be used to extract hidden features of a group of patients and state of diseases that can aid in automated decision making**.** Data mining offers a clear examination in the field of biomedical[3].

2. Methods

2.1 Text Mining

Text Mining[4] is the task of discovering unknown information that it may be new or previous information, extracting automatically from various text documents.

* *Author for correspondence*

2.2 Text Mining Techniques

The text mining techniques are IE (Information Extraction), IR (Information Retrieval), Categorization or Classification, Topic Tracking, Clustering, Summarization and Concept Linkage. Classification includes SVM (Support Vector Machine), K-NN (K-Nearest Neighbor), NN (Neural Network), Bayesian Method and DT (Decision Tree). Clustering includes Partition Method, DB (Density Based) Clustering and Hierarchical clustering. These models are described in the following sections.

2.2.1 IE (Information Extraction)

IE (Information Extraction)[5] is that structured information is automatically extracted from unstructured / semi structured documents. This is mostly done by Natural Language Processing (NLP).

2.2.2 IR (Information Retrieval)

IR (Information retrieval) is nothing but finding and extracting information in documents. The documents may be web documents contain text or image[6].

2.2.3 Summarization

Summarizing of text in a compressed form of its input, which specifies human Consumption. The document can be either individual or group of document[7].

2.2.4 Classification

Classification is one of the supervised techniques. It is the task of searching a model that explains and prominent data classes or concepts. These models are derived from the analysis of trained data. The model is used to conjecture the objects class label that is unknown[8].

Various classification techniques can be applied to categorize the text such as SVM (Support Vector Machine), K-NN (K-Nearest Neighbor), NN (Neural Network), Bayesian Method and DT (Decision Tree).

2.2.4.1 K-NN (K-Nearest Neighbor)

K-NN (K-Nearest Neighbor) is a technique of correlation learning, which is comparing the training tuples with the given test tuple, which are similar to it[9].

2.2.4.2 DT (Decision Tree)

DT (Decision tree) is like a tree structure, in which test attributes as internal node, test out come as branch node and class cable as leaf node. Root node is the topmost node in the tree. Decision tree has been used in operations research to find the conditional probabilities[10].

2.2.4.3 SVM (Support Vector Machine)

SVM (A Support Vector Machine) converts the training data, where it finds a hyper plane using support vectors. The hyper plane splits the data by class.

2.2.4.4 NN (Neural Network)

NN (Neural Network) is a huge number of individual neurons similar to processing nodes and a huge number of weights between these nodes[10].

2.2.4.5 Bayesian Method

In classification and probabilistic learning, Bayes theorem played an important role. Prior Knowledge and observed data is combined by the Probabilistic model. Naïve Bayes classification is one of the simplest Bayesian Algorithm. It has two phases they are learning phase and test phase[11].

2.2.5 Clustering

Clustering is the task of partitioning a dataset objects in to subsets. Every subset is called a cluster; the objects are similar to one another in one cluster and dissimilar in another cluster. Clustering Methods Are Partition Method, DB (Density Based) Clustering and Hierarchical clustering. These models are described in the following sections[12].

2.2.5.1 Hierarchical Clustering

This clustering is made by multiple levels. It can be categorized as either divisive or agglomerative, based on the decay is formed[13].

2.2.5.2 Partition Method

In this method dataset objects are partitioned into several clusters, Formally, given set S, number of objects as N and number of clusters as M to form a partitioning algorithm classifies the objects into M partitions (M<=N), where

each partition denotes a cluster[12].

2.2.5.3 DB (Density Based) Clustering

DB method is the process of partitioning a dataset objects into multiple or a hierarchy clusters. DB clustering can be stretched into subspace from full space clustering[12].

2.2.6 Topic Tracking

Topic tracking is used to track the user views, to find the changes in the IAI (Information Area of Interest) and regularly produces a summarized report of changes, this reveal the emerging topic in the particular information area[14].

2.2.7 Concept Linkage

Concept linkage is a technique to find the corresponding documents which share the same concepts[15]. Concept linkage is mainly used to promote browsing.

2.2.8 Question Answering

The question answering method is used to ask questions in World Wide Web and then it gets the related answer. This technique has allowed in many websites[16].

3. Related Works and Discussion

Microarray data were analysed by various classification method such as SVM, Decision tree, Bagging, Boosting and Random Forest. The data set obtained from Kent Ridge were comparatively analysed by 10-fold cross validation approach. Among all classification methods random forest shows accuracy result[17].

The early warning system of chronic disease was promoted by KNN and Linear Discriminate Analysis (LDA). The connection between the heart disease and hypertension were analysed and minimized the complication occurrences of the disease by constructing an early warning system[18].

Patient those who are having chronic disease are classified by their actions, using universal hybrid decision tree. They extended their research work to get more accuracy by classifying various activities of patients[19].

SVM classification method is used to classify many diseases. For diagnosing diseases, the combination of both SVM and K means clustering were applied to microarray data[20].

ANN was used to examine chest diseases, comparative analysis also done for chest diseases that was conducted by probabilistic neural network, generalized regression and multilayer neural networks[21].

Bayesian method has an outstanding performance in diagnosis of psychiatric disease. The dataset of psychiatric patient was taken from Lugoj municipal hospital[22].

Genetic support vector machines (GSVM) was performed better analysis for heart valve diseases. GSVM classifies the ultrasound signal of heart valve and also extracts important features. In this work the automatic system examines heart valve diseases from 215 samples. After evaluation of samples the result was effectually find the Doppler heart sounds[23]. The comparison of different classification methods is shown in Table 1.

4. Conclusion

The various text mining methods in biomedical field were

Table 1. Comparison of different classification methods

Author	Algorithm	Working Mode	Advantages	Limitations
Hu et al[17].	Random Forest	i) Construct with many trees. ii) After each tree is built, all of the data are run down the tree. ii) Proximities are computed for each pair of cases.	i) It is unexcelled in accuracy among current algorithms. ii) It runs efficiently on large databases. iii) It can handle thousands number of input variables without variable deletion	i) Random forests have been observed to over fit for some datasets with noisy Classification/ Regression tasks. ii) It is not reliable for categorical variables with different number of levels.
Jen et al[18].	K-Nearest Neighbor	i) Find out the unidentified data point using the previously known data points (nearest neighbor).	i) It is easy to implement. ii) Training is done in a faster manner.	i) Testing is slow. ii) It requires a large storage area. iii) Sensitive to noise.

Chief et al[19].	Decision Tree(DT)	i) Search based on the topic or previously viewed by the user. ii) The topic is conjectured by the interest of the user.	i) Simple to understand and interpret. ii) There are no requirements of domain knowledge in the construction of decision trees. iii). It reduces the ambiguity of complicated decisions and assigns exact values to outcomes of various actions. iv).Performs well with large datasets	i)It is restricted to one output attribute.ii) It generates categorical output. iii) It is an unstable classifier i.e. performance of classifier is depend upon the type of dataset.
Soliman et al[20].	Support Vector Machine(SVM)	i) First select the main sentences and paragraphs. ii) Join them into an abstract form. iii) Perceive main concepts of the text. iv) Convey main concepts in natural language.	i) Better Accuracy as compare to other classifier. ii) Easily handle by complicated nonlinear data points. iii) Over fitting problem is not as much as other methods.	i) Computational is very expensive. ii) The main drawback is to choose a right kernel function. For each dataset different kernel function shows various results.
Er et al[21].	Neural Network	i) Used to perform non-linear statistical modeling ii) Uses gradient descent method. iii) Based on neurons. iv) Having multiple interconnected processing elements known as neurons.	i) Easily find the complex relationships between dependent and independent variables. ii) Able to handle noisy data.	i) Local minima. ii) Over-fitting. iii) The processing of ANN network is difficult to interpret and require high processing time if there are large neural networks.
Curiac et al[22].	Bayesian Method	i)Based on Bayes theory. ii) Concerted on prior, posterior and discrete probability distributions of data items.	i) It makes the computation process easier. ii) Have better speed and accuracy for large datasets.	i) It does not give exact results in some cases where there exists dependency among variables.
E. Avci et al[23].	Genetic support vector machines (GSVM)	i) Promote solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection and crossover.	i) Prediction accuracy is generally high. ii) Robust, works when training examples contain errors. iii) Fast evaluation of the learned target function.	i) Long training time. ii) Difficult to understand the learned function (weights). iii) Not easy to incorporate domain knowledge.

analysed in this survey. Their merits and limitations have been discussed. The purpose of the study is that how the classification methods are applied in biomedical field and to select a method which is suitable for diagnosing a particular disease. According to the performance the classification technique is well suitable for diagnosing diseases. Classification can be done on the basis of patient disease pattern to separate the patients into high risk or low risk. Classifying patients related information were diagnosed that leads to good result.

5. References

1. Jeon M-S, Kim H-J. Awareness levels of biomedical ethics in undergraduates. Indian Journal of Science and Technology. 2015 April; 8(S8); 149–53. DOI: 10.17485/ijst/2015/v8iS8/71493.
2. Biomedical [Internet]. [Cited 2016 Jan 07]. Available from: http://home.iitj.ac.in/~bagler/research/biomed.html.
3. Suganya P, Sumathi CP. A novel meta heuristic data mining algorithm for the detection and classification of parkinson disease. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–9. DOI: 10.17485/ijst/2015/v8i14/72685.
4. Michael WB. Automatic discovery of similar words in survey of text mining: clustering, classification and retrieval. Springer Verlag, New York, LLC; 2004.
5. Information extraction [Internet]. [Cited 2016 Jan 14]. Available from: https://en.wikipedia.org/wiki/Informa-

tion_extraction.
6. Jhanjil D, Garg P. Text mining; 2014
7. Agrawal R, Batra M. Detailed study on text mining techniques; 2013.
8. Jinshu S, Zhang B, Xin X. Advances in machine learning based text categorization. Journal of Software. 2006; 17(9):1848–59.
9. Han E-H, Karypis G, Kumar V. Text categorization using weight adjusted k-nearest neighbor classification. Army HPC Research Center University of Minnesota.
10. Goharian and Grossman. Data Mining Classification, Illinois Institute of Technology [Internet]. [Cited 2016 Jan 25]. Available from: http://ir.iit.edu/~nazli/cs422/CS422-Slides/DM-Classification.pdf .
11. Abraham R, Simha JB, Iyengar SS. Effective discretization and hybrid feature selection using naïve bayesian classifier for medical data mining. International Journal of Computational Intelligence Research. 2008.
12. Han J, Kamber M, Pei J. Data mining concepts and techniques. Cluster Analysis, 3rd edn; 2012.
13. Hierarchical clustering [Internet]. [Cited 2016 Jan 26]. Available from: https://en.wikipedia.org/wiki/Hierarchical_clustering.
14. Gupta V, Lehala GS. Survey of text mining techniques and applications. Journal of Emerging Technologies in Web Intelligence. 2009 Aug; 1(1):125–33.
15. Chau R, Tsoi AC, Hagenbuchner M, Lee VCS. A concept link graph for text structure mining. Wellington. New Zealand; 2009 Jan.
16. Chali Y, Joty SR, Hasan SA. Complex question answering: unsupervised learning approaches and experiments. Journal of Artificial Intelligence Research. 2009; 35(1):1–47.
17. Hu H, Li J, Plank A, Wang H, Daggard G. A comparative study of classification methods for microarray data analysis. Proceedings of Fifth Australasian Data Mining Conference (AusDM2006), Sydney, Australia. CRPIT, ACS; 2006. p. 33–7.
18. Jena CH, Wang CC, Jiangc BC, Chub YH, Chen MS. Application of classification techniques on development an early-warning system for chronic illnesses. Expert Systems with Application. 2012; 39:8852–58.
19. Chien C, Pottie GJ. A universal hybrid decision tree classifier design for human activity classification. Proceedings of 34th Annual International Conference of the IEEE EMBS San Diego, California: USA; 2012 Aug 28–Sep 1.
20. Soliman THA, Sewissy A, Latif HA. A gene selection approach for classifying diseases based on microarray datasets. Proceedings of 2nd International Conference on Computer Technology and Development (lCCTD 2010); 2010.
21. Er O, Yumusakc N, Temurtas F. Chest diseases diagnosis using artificial neural networks. Expert Systems with Applications. 2010; 37:7648–55.
22. Curiac DI, Vasile G, Banias O, Volosencu C, Albu A. Bayesian network model for diagnosis of psychiatric diseases. Proceedings of the ITI 2009 31st Int. Conf. on Information Technology Interfaces, Cavtat: Croatia; 2009 Jun 22–25.
23. Avci. A new intelligent diagnosis system for the heart valve diseases by using genetic-SVM classifier. Expert Systems with Applications. 2009; 36:10618–26.