IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

A Review on Automated Disease Diagnosis Techniques

Sunena Rose M V[1], Dr. Sobhana N. V[2]

Department of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India[1]

Professor, Dept of Computer Science and Engineering, Rajiv Gandhi Institute of Technology, Kottayam, Kerala, India[2]

Abstract: The development of computer technologies and increased expenditure of healthcare are the reasons for innovation of automated disease inference system. Automated disease inference system will give the disease information which a person is facing on the basis of health related questions. Using this system health seeker will get immediate response as compared to the existing system. Diseases and symptoms are collected and used as the Question Answer pairs. In this paper, a survey of different techniques for automatic disease diagnosis is done. In shallow learning methods, which make use of patient details from hospital records with structured fields, they can focus on only a single or a few diseases. The automated scheme can deals with wide range of diseases which posses scalability and generalizability.

Keywords: Support Vector Machine; Neural Network; Deep Learning; Decision Tree Induction.

I. INTRODUCTION

Research and development in healthcare field is a very important part of science because no human being is immune to diseases. There are a lot of medical information is available from diverse sources like health related websites, hospital records and social media. One survey in [1] shows that 59 percent of U.S. adults using internet as diagnostic tool. Results from a national consumer survey conducted by Makovsky Health and Kelton [2] show the average U.S consumer spends nearly 52 hours looking for health information on the internet annually, and visit the doctor three times. These findings show the relevance of online health resources and underscore the importance of accuracy and accessibility of online health information which act as a springboard for patient-doctor communication and peer support.

The online resources which are currently prevailing can be broadly classified into two categories. One is reputable portals run by official sectors, renowned organizations or other professional health providers. These portals contain up-to-date health information on various topics in a well-structured manner. WebMD and MedlinePlus are typical examples. The category is the community-based health services, such as HealthTap and HaoDF. They provide interactive platforms in which health seekers can ask health related query while the doctors provide the knowledgeable answers. However there are some limitations for the community based health sevices. First of all, it is very time consuming for health seekers to get their posted questions resolved. It will take hours or days. Second, doctors have to cope with ever-expanding workload, which leads to decreased efficiency. Third, the quality of answers given by the doctor depends on the experience and time. Health seekers ask the questions for supplemental information of their diagnosed disease, preventive information of their disease and possible disease of their manifested signals.

II. LITERATURE REVIEW

Liqiang Nie et al. [3] proposed a disease inference system from health related questions via sparse deep learning. In this method, it first mines the medical signatures from the text data and applied to sparse deep learning network. The inference of diseases is based on the inter-relations developed from the adjacent layers in the network. The network consists of three hidden layers. It is an unsupervised learning method, which learns from a wide range of disease types. This study tried to categorize the queries asked by the health seekers. The model cannot identify discriminant features for each specific disease.

Extracting relevant informations about diseases and symptoms from a clinical record is a very important task related to an automatic disease inference system. Parikshit Sondhi et al. [4] constructed a framework for mining clinical notes through symptom relation graph which is known as SymGraph. Symptoms and their co-occurrence relations are collected from the clinical data of a number of patients and the symptoms become nodes and co-occurrence relationship between the symptoms become the edges of SymGraph. This work mainly focused on finding the relationship between different symptoms, to expand the set of symptoms and to update symptom relationships. This method only considers the explicitly present medical entities.

Data mining techniques in healthcare field is of very importance and it provides better classification, clustering and knowledge extraction from health related data. Data mining methods applied to extract hidden and potential

relationships and patterns from large datasets and very useful for early detection of diseases. Mai Shouman et al. [5] make use of decision trees for diagnosing heart disease patients. Type of decision trees and data discretization methods are focused in this work. Unsupervised discretization methods such as equal-width interval and equal frequency methods, supervised discretization methods such as chi merge and entropy are used in this work for finding better performance. Information Gain and Gain ratio are the different decision tree types used. Multiple classifiers voting technique and reduced error pruning was also used to produce accurate decision rules.

A heart disease diagnosis method is developed by Shashikant Ghumbre et al. [6] using Support Vector Machine. Radial Basis Function (RBF) Network is a type of Artificial Neural Network which posses a feed forward structure with single hidden layer. This work compares RBF method with SVM. Support Vector Machine provides more accuracy in pattern classification problems and become a good choice for automated diagnosis.

Fused Sparse Group Lasso is a type of modeling disease progression. Jiayu Zhou et al. [7] designed this model in the context of Alzheimers disease. In this work, cognitive scores of patients are measured at multiple times instant and the problem of predicting cognitive scores considered as multi-task regression. The temporal pattern of different features at each timepoint can reveal the state of disease progression. This work only considers the case of training data with no missing values.

In community based health services, vocabulary gap is a problem that arises between the health seeker and stored data. Liqiang Nie et al. [8] designed a joint local global approach for medical terminology mapping. Local mining map the question answers pair into the entries in external dictionary. It consist of noun phrase extractor, medical concept identifier and medical concept normalize. Global learning model enhance the result obtained from local learning. Liqiang Nie et al. [9] proposed a method to automatically code inconsistent data from community generated health documents and used Concept Entity Impurity (CEI) approach for local mining. Inter-expert and inter-terminology relationships are extracted for a graph-based global learning.

For extracting logical knowledge for deep question answering Zhao Liu et al. [10] proposed an unsupervised method which used semantic role labeling techniques to convert natural language expression to predicates in first logic. Markov logic networks are the inference engine used in this work. Aliaksei Severyn et al. [11] proposed a convolutional deep learning model for reranking short text pairs. Two distributional sentence models based on convolutional neural networks are also designed in this work. These models used to findout semantic similarity among queries and documents. Sentence matrix, convolutional and pooling layers are the components of neural network.

Md Moinul Hoque et al. [12] designed a semantic understanding based ontological framework for factoid question answering. The model extracts text and builds a knowledge base which can give answers to the factoid questions. Cognitive Entity Relationship Network presented in this paper can find the relation between entities. A text inference engine presented in this work can isolate different features and objects. An ontology mapping module is responsible for mapping ontology of questions with knowledge-base. This model cannot answer the question that doesn't have prestored facts.

K. Charalampous et al. [13] presented a sparse deep learning algorithm for recognition and categorization. In this model, hierarchical temporal memory (HTM) is the basic computational unit which can store and process the data. It makes use of machine learning and optimization methods and outperforming the predecessors in the field of categorization.

F. Wang et al. [14] proposed a Nonnegative Matrix Factorization (NMF) based framework using a convolutional approach for open ended temporal pattern discovery from clinical data. In this method, each patient is represented by an event matrix. NMF is a powerful tool for identifying underline structures in a matrix through regularized decomposition. The convolution only occurs along the time axis but not on event axis. So it is called one sided. The method can mine shift-invariant temporal patterns from heterogeneous events, handle sparsity and scalability problems

S. Doan and H. Xu [15] developed a Support Vector Machine (SVM) based Named Entity Recognition (NER) system for recognizing medication related entities in hospital discharge summaries. Performance of the SVM based NER system for different feature combinations are analysed in this work. The machine learning based NER system combined with the semantic features from a rule based system could achieve the best F-score of 90.05% in recognizing medication related entities.

J. Zhou et al. [16] proposed a multi-task learning formulation for predicting disease progression is to track the progress of Alzheimer's disease and to identify it in early stage. In this method the prediction is from the cognitive scores and selecting markers predictive of the progression. To analyze the intrinsic relatedness among different tasks at different time periods, a temporal group lasso regularizer uses a subset of features. This work only focuses on linear models.

A. Khosla et al. [17] designed an integrated machine learning approach to stroke prediction. Support Vector Machine and Margin-based Censored Regression (MCR) are the learning algorithms used in this method. This work provides an extensive comparison of machine learning methods with the Cox proportional hazards model and found that machine learning methods outperform the Cox model. Conservative mean heuristic method is used for feature selection. This work motivates the application of machine learning methods in healthcare analysis.

Nut Limsopathen et al. [18] proposed medical records search by combining bag-of-words representation and bag-of-concepts representation. The method linearly merges

the relevance score for each query and improves the performance of retrieval.

Liqiang Nie et al. [19] developed a method for comprehensive vertical search for healthcare domain, named Wenzher. It is a recommender system and social platform which organizes the healthcare data into a single view for the given query. The answer will be from heterogeneous data sources. Vertical search over the question answer pairs in the database eliminates duplication problem.

Deepthi Vadicherla et al. [20] studied the classification of Heart Disease using Support Vector Machine and Artificial Neural Network (ANN). Comparison among the methods shows that SVM machine with Sequential Minimal Optimization (SMO) is efficient than ANN.

A formwork to find related information about a disease and treatment in a medical domain repository is developed by Oana Frunza et al. [21]. A study of different machine learning algorithms along with different textual representation are done in classifying short texts and finding the semantic relations between disease and treatments. It is found that probabilistic model with rich representation of texts gives better results.

Long Wan and Wenxing Bao [22] proposed a model of animal disease intelligent diagnosis using Support Vector Machine. In this model, separate Support Vector Machines are used for each class of disease and classified SVM are combined to get a SVM classification group. This grouping improves accuracy and performance of the model.

S. N Deepa and B. Aruna Devi [23] conducted a study on Neural Networks and Sequential Minimal Optimization (SMO) based classification for Brain Tumor. The method identifies normal and abnormal brain images using a feature set, texture features are extracted from the images and classifying the images based on these texture features. SMO classifier performs better than Neural Networks. Selection and evaluation of features are difficult due to the minimum optimization approach used in this model.

Ammar Asjad RajaIrfan-ul-Haq et al. [24] developed Syncope disease prediction framework using data mining-ensemble technique. Support Vector Machine, Naïve Bayes and Decision Tree Induction based on Gini index are the different algorithms used in this method. Each classifier is fed with the testing data and their output is used as the input to voting system in the ensemble method. Result obtained from the ensemble method is more accurate than the results from individual classifiers.

III. CONCLUSION

Automatic disease inference system gives the disease information which a person facing and that by sitting at the home. While these systems are constructed, the data used needs to be classified appropriately. In this study, we tried to give a brief introduction to the broad field of automatic disease inference and presented a brief overview of currently available techniques. At last we conclude that, Deep Learning is the more advanced scheme that is able to infer the possible diseases for the given question of health seekers. It is generalizable and scalable as compared to previous disease inference using other learning approaches. The accuracy and efficiency of the inference can be further improved by using ensemble methods. It combines the predictions of several base estimators and produces more accurate solutions than a single model. Combining the predictions from Deep Learning, Support Vector Machine classifier and Decision tree classifier, it will produce more reliable outcomes.

REFERENCES

[1] "Online health research eclipsing patient-doctor conversations,"Makovsky Health and Kelton, Survey, 2013.
[2] S. Fox and M. Duggan, "Health online 2013," Pew Research Center, Survey, 2013.
[3] LiqiangNie, Men Gang, Luming Zhang & Shuicheng Yan (2015), "Disease Inference from Health –Related Questions Via Sparse Deep Learning", IEEE Transaction on Knowledge and Data Engineering, Vol.27, No.8.
[4] Sondhi, J. Sun, H. Tong, and C. Zhai, "Sympgraph: A framework for mining clinical notes through symptom relation graphs," in Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining, P.2012, pp. 1167–1175.
[5] M. Shouman, T. Turner, and R. Stocker, "Using decision tree for diagnosing heart disease patients," in Proc. 9th Australasian Data Mining Conf., 2011, pp. 23–30.
[6] S. Ghumbre, C. Patil, and A. Ghatol, "Heart disease diagnosis using support vector machine," in Proc. Int. Conf. Comput. Sci. Inf. Technol., 2011, pp. 84–88.
[7] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, "Modeling disease progression via fused sparse group lasso," in Proc. ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012, pp. 1095–1103.
[8] L. Nie, M. Akbari, T. Li, and T.-S. Chua, "A joint local-global approach for medical terminology assignment," in Proc. Int. ACM SIGIR Workshop, 2014, pp. 24–27.
[9] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua, "Bridging the vocabulary gap between health seekers and healthcare knowledge," IEEE Trans. Knowl. Data Eng., vol. 27, no. 2, pp. 396–409, un. 2014.
[10] Zhao Liu, Xipeng Qiu, Ling Cao, and Xuanjing Huang, "Discovering Logical Knowledge for Deep Question Answering," in Proc. CIKM'12,USA October 29–November 2, 2012
[11] Aliaksei Severyn and Alessandro Moschitti, "Learning to Rank Short Text Pairs with Convolutional Deep Neural Networks," in Proc. Int. ACM SIGIR'15, August 09 - 13, 2015.
[12] Md Moinul Hoque and Paulo Quaresma, "SEMONTOQA - A Semantic Understanding-Based Ontological Framework for Factoid Question Answering," in Proc. Int. ACM FIRE '14, December 05-07, 2014.
[13] K. Charalampous, I. Kostavelis, A. Amanatiadis and Gasteratos, "Sparse deep-learning algorithm for recognition and categorization," ELECTRONICS LETTERS 27th September 2012 Vol. 48, No. 20.
[14] F. Wang, N. Lee, J. Hu, J. Sun, and S. Ebadollahi, "Towards heterogeneous temporal clinical event pattern discovery: A convolutional approach," in Proc. 18th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2012, pp. 453–461.
[15] S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine," in Proc. Int. Conf. Comput. Linguistics, 2010, pp. 259–266.
[16] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in Proc. 17th ACM SIGKDD Conf. Knowl. Discovery Data Mining, 2011, pp. 814–822.
[17] A. Khosla, Y. Cao, C. C.-Y. Lin, H.-K. Chiu, J. Hu, and H. Lee, "An integrated machine learning approach to stroke prediction," in

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 3, March 2017

Proc. 16th ACM SIGKDD Conf. Knowl. Discovery Data Mining,2010, pp. 183–192.

[18] N. Limsopatham, C. Macdonald, and I. Ounis, "Learning to combine representations for medical records search," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2013, pp. 833–836.

[19] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua, "Wenzher: Comprehensive vertical search for healthcare domain," in Proc. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval, 2014, pp. 1245–1246.

[20] Deepti Vadicherla and Sheetal Sonawane. "Classification of Heart Disease Using SVM and ANN," in Proc. International Journal of Research in Computer and Communication Technology, 2013, Vol 2, Issue 9.

[21] Oana Frunza, Diana Inkpen, and Thomas Tran, "A Machine Learning Approach for Identifying Disease-Treatment Relations in Short Texts," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, 2011, VOL. 23, NO. 6.

[22] Long Wan, Wenxing Bao, "Research and Application of Animal Disease Intelligent Diagnosis Based on Support Vector Machine" IEEE International Conference on Computational Intelligence and Security, 2009, Pages 66-70.

[23] S.N. Deepa, B. Aruna Devi, "Neural Networks and SMO based Classification for Brain Tumor", IEEE World Congress on Information and Communication Technologies, 2010, Pages 1032-1037.

[24] Ammar Asjad Raja, Irfan-ul-Haq, Madiha GuftarTamim Ahmed Khan and Dominik Greibl, "Intelligent Syncope Disease Prediction Framework using DM-Ensemble Techniques", Future Technologies Conference (FTC) 2016 IEEE, 6-7 December 2016, San Francisco, United States .