Tracking patient disease through symptoms via sparse deep learning

1. Sonali C. Sethi, 2.Kulkarni P.R.

[1.] M.E. (Student), Department of computer science, AEC, Beed, Maharashtra, India. - 2. Assistant Professor, Department of Computer Science, AEC, Beed, Maharashtra, India.

ABSTRACT

Automatic disease inference is of significance to overcome any issues between what online health seekers with strange side effects need and what occupied human doctor with one-sided aptitude can offer. However, accurately and efficiently inferring diseases is non-trivial, especially for community-based health services due to the vocabulary gap, incomplete information, correlated medical concepts, and limited high quality tra ining samples. Here the sparse deep learning algorithm is used as the data mining technique. Deep learning is a branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using model architectures, with complex structures. The proposed scheme uses question-answering, deep learning as inferring methods. Some attributes used are raw features, medical attributes etc. The proposed scheme is comprised of two key components. The first globally mines the discriminate medical signatures from raw features. The second deems the raw features and their signatures as input nodes in one layer and hidden nodes in the subsequent layer, respectively. Meanwhile, it learns the inter-relations between these two layers via pre-training with pseudo- labeled data. . This paper present idea of deep learning architecture which is used in the health care domain for the diagnosis of diseases.

Keyword: *- data mining, disease inference, deep learning.*

1. Introduction

The graying of society, increasing costs of healthcare and burgeoning computer technologies tends to more consumers to spend longer time online to explore health information. The current prevalent online health resources can be categorized into two categories. One is the reputable portals run by official sectors, famous organizations, or other professional health providers. Second category is the community based health services. Both resources offer interactive platform, where users can ask their health related questions and doctors can provide trustworthy and satisfactory answers. There are number of online health portals are available but they have several limitations .First off all it is very time consuming for user to get their posted questions Take Health-Tap as an example, which is a question answering site for participants to ask and answer health-related questions. The questions are written by patients in narrative language. The same question may be described in substantially different ways by two individual health seekers. Deep Learning is a relatively recently developed set of generative machine learning techniques that autonomously generate high-level representations from raw data sources, and using these representations can perform typical machine learning tasks such as classification, regression and clustering. The prime intention o f deep learning comprises of two key components. The first globally mines the discriminate medical attributes from raw features. The raw features serve as input nodes in one layer and hidden nodes in the subsequent layer, respectively. The second learns the inter- relations between these two layers via pre-training. With incremental and alternative repeating of these two components, the scheme builds a sparsely connected deep learning architecture with three

hidden layers. Deep learning scheme is applied to infer the possible diseases using Question and Answer data values.

2. Literature survey

Many people spend longer time online to explore health information. One survey in [5] shows that 59% of U.S. adults have explored the internet as a diagnostic tool in 2012. Another survey in [6] reports that the average consumer spends close to 52 hours annually online to find wellness knowledge, while only visits the doctors three times per year in 2013. Accurate disease inference is one of the great challenges faced by the researchers and user as well. Many researchers developed different methods for disease inference for different diseases. This related work in the health care domain is as follows:

2.1 Decision Tree

Decision tree is one of the data mining techniques used for the diagnosis of the heart disease. Author Mai Shumen, Tim Turner, Rob Stocker used decision tree in the health domain for the diagnosis of heart diseases .Decision Tree that is based on gain ratio and binary discretization. This methodology involves systematically testing different discretization techniques, different Decision Trees type and multiple classifiers voting technique and in the diagnosis of heart disease patients. Different combinations of decision tree types and voting, discretization methods, are tested to identify which combination will provide the best performance in diagnosing heart disease patients. Decision tree has useful accuracy in the diagnosis of heart disease but it has several limitations hence its result may not be able to give accurate result.

2.2 Support Vector Machine

Support Vector Machine is a method of machine learning, classification and recognition, which is based on statistical theorems .The classification performance of SVM, is superior to traditional classification methods, especially the generalization. Support vector machine is widely used for the heart disease and cancer diagnosis. SVM is powerful classification algorithm which is used to determine the support vectors in a fast, iterative manner. An integer-coded genetic algorithm was applied to Cleveland heart disease database for selecting the important and relevant features and discarding the irrelevant and redundant ones. SVM gives maximum accuracy with increase in size of training data.

2.3 Radial Basis Function Network

RBFs are feed forward neural network that have three layers, namely input layer, output layer and hidden layer. Hidden unit in the middle layer implements a radial activated function. Radial basis function network having advantages over the other methodologies are less computational time and its accuracy. Input layer consist of number of features in the form of medical condition and in the middle layer radial basis function is implemented and at the end disease is diagnosed in the output layer. Output layer contains M number of unit which is the possible no of output. The result of each output unit is the sum of the output of the middle layer. Multi-layer perceptron is combined with radial basis function network to show the absence or presence of disease.

2.4 Multilayer Perceptron Networks

MPN is the most common architecture of neural networks. Multilayer perception networks mainly used in diagnosis of heart disease .MPN is a feed forward neural network model that maps sets of input data onto a set of appropriate output. It consists of three layers input layer, one hidden layer and output layer. The input and output layers consist of input and output neurons respectively, whose numbers are set in task specification whereas number of neurons in hidden layer that will lead to successful classification results is not trivial and is usually task dependent. The number of hidden neurons is important for accuracy. Training of MPN is based on error correction back propagation algorithm. This algorithm consist of two phase, which is called as feed forward pass and back propagation. In the first phase supervised learning is used where the network is trained using the data for which inputs and the desired outputs are known. Every node in the hidden layer of the network as well as the nodes in the output layers calculates the activation values. The differences between actual output and the desired output are used to calculate the error. During the back propagation, the error is propagated from the output layer to input layer.

3. Context Analysis Using Question Answering Deep Learning

a) This is the first work on automatic disease inference in the community-based health services. Distinguished from the conventional sporadic efforts that generally focus on only a single or a few diseases based on the hospital generated records with structured fields, our scheme benefits from the volume of unstructured community generated data and it is capable of handling various kinds of diseases effectively.

b) It investigates and categorizes the information needs of health seekers in the community-based health services and mines the signatures of their generated data.

c) It proposes a sparsely connected deep learning scheme to infer various kinds of diseases. This scheme is pre-trained with pseudo-labeled data and further strengthened by fine-tuning with online doctor labeled data.

4. System Architecture

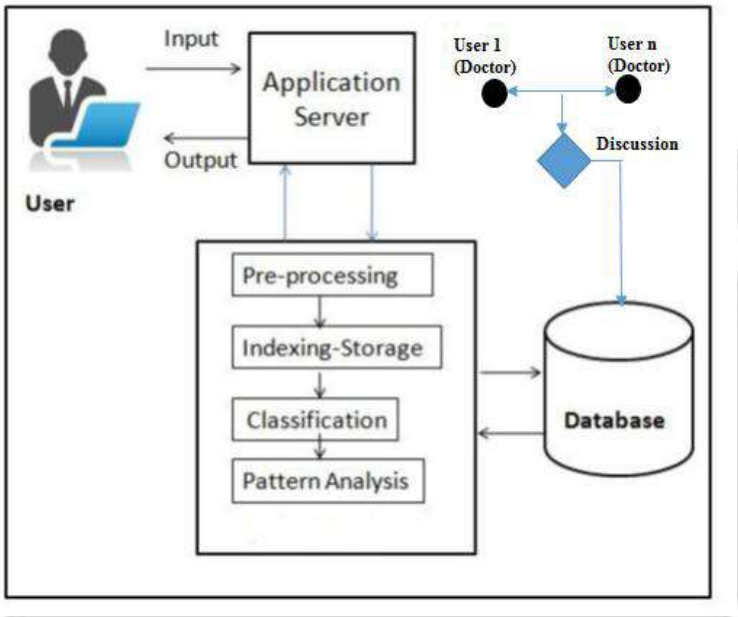

Fig -1: system architecture

1) In this architecture the four main components are the Health Seeker, Application Server, Doctors in community and the Dataset.

2) Health Seeker is the person who gives input in the form of Query to the Application server, Application Server is responsible for taking input from user and passing it to further block for its processing and in return it also gives back result to the health seeker.

3) Query's passed is processed first and different pre- processing operations are applied on the given query and finally divided in a form of Tokens using String-Tokenizer.

4) This Tokens are the passed in further block for Indexing Storage. Where we mine each and every data related to the tokens we got from our Data Set. Each Data is assigned with an Index value. All the Relevant Data we got is then passed further for its Classification.

5) In Classification Block the data are divided into different classes. First step in text classification is transforming text which is in string format into format suitable for learning algorithm.

6) In Pattern analysis Boyer-Moore-Horspool algorithm is used. Layer by layer elimination of entries is done by using Deep learning algorithm. Recursive Process is done here in classification and pattern analysis block till the result is not obtained by Eliminating the Entries from Generalized data to Specified data i.e., from huge data to a smaller data and hence the result is obtained here.

6) Finally, the Result obtained through Classification and Pattern analysis is then passed to user.

5. Sparse Deep Learning

Deep learning is also called as hierarchical learning or structure learning .It is the branch of machine learning based on a set of algorithms that attempt to model high-level abstractions in data by using multiple processing layers with complex structures. Deep learning presents this idea of hierarchical explanatory factors where higher level, more abstract concepts being learned from the lower level ones. These architectures are often constructed with a layer-by-layer method. Sparsely connected deep learning method contains no of layers depending on the health care application.

This technique mainly contains input and output layers. The nodes in the input layer contain number of raw features and nodes in the output layer contain result which denotes ultimate disease type. Remaining layers are constructed incrementally alternating between sub graph mining and pre-training. Each node in the hidden layer is corresponding to a signature obtained by sub graph mining from a large graph, where the large no of raw features are assumed as nodes and edges, respectively. This deep learning method employs with the help of signature mining which help to find interdependent medical attributes from the large dataset.

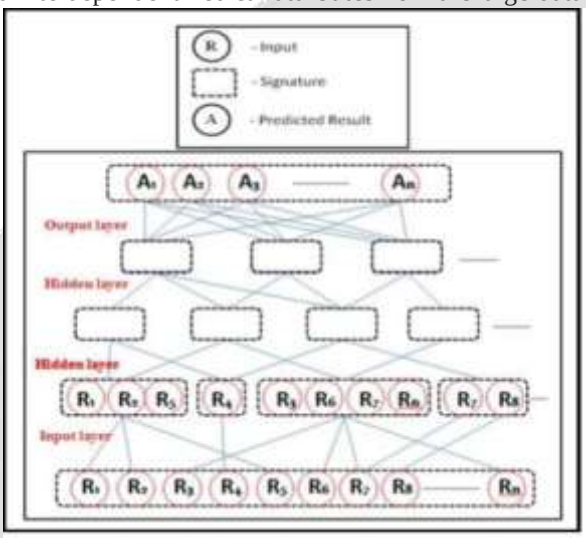

Fig -2: Hidden Layer Architecture

6. Proposed system

To build a disease inference scheme that is able to automatically infer the possible diseases of the given questions in community-based health services. Analyze and categorize the information needs of health seekers. As a byproduct, differentiate questions of this kind that require disease inference from other kinds. It is worth emphasizing that large-scale data often leads to explosion of feature space in the lights of n-gram representations, especially for the community generated inconsistent data. The system distinguished from the conventional sporadic efforts that generally focus on only a single or a few diseases based on the hospital generated records with structured fields, proposed scheme benefits from the volume of unstructured community generated data and it is capable of handling various kinds of diseases effectively. It investigates and categorizes the information needs of health seekers in the community-based health services and mines the signatures of their generated data. It proposes a sparsely connected deep learning scheme to infer various kinds of diseases. This scheme is pre-trained with pseudo-labeled data and further strengthened by fine-tuning with online doctor labeled data. This scheme builds a novel deep learning model, comprising two components. The first globally mines the latent medical signatures. The raw features and signatures respectively serve as input nodes in one layer and hidden nodes in the subsequent layer.

7. Existing System

Little research has been dedicated to disease inference in the community-based health services. Information extraction from medical text is the basis for other higher-order analytics, such as representation, classification, and clustering. SVM is used to recognize the medication related entities in hospital discharge summaries, and classified these atomic elements into pre-defined categories, such as treatments and conditions. Machine learning techniques used to assist health professionals in the diagnosis of diseases.

7.1Disadvantages

1. These approaches are not applicable to online health data.

2. From the perspective of data property, they have different data structure, quality and number of training samples.

3. From the point of techniques, most of the previous efforts are unable to take advantages of other data types beyond the targeted ones, and hence are not scalable or generalizable.

8. Mathematical Model

System Description: Let S be the required System,
A System 'S' is defined as a set such that:
S = {s, e, X, Y, f, DD, NDD, success, failure}
Where,

s: Initial/Start state
e: End state/Final state
I: Set of input
O: Set of output
f: Functions .
DD: Deterministic data
NDD: Non Deterministic data

Success: Desired output of system is generated.
Failure: Desired output of system is not generated.

1. **Start States :**
   The Start State of the System where Authentication of the user is done.
   Users: {UR, UN }
   Where,
   UR: Set of Registered Users
   UN: Set of Un-Registered Users

2. Input Set 'I': {I1 , I2 }
(a) Phase 1: Registration
Set I1: {i11, i12,i13,i14,i15,i16 }
where,

i11: username
i12: Address
i13: Pin code
i14: Mobile no
i15: Email id
i16: Images

(b) Phase 2: Communication
Set I2: { i21, i22 }
Where, i21: UserID
i22: Context

3 Output Set 'O' : { O1, O2, O3 }

(a) Phase 1: Registration
Set O1 :{ o11, o12 }
Where, o11 : userID ; o12 : Password

(b) Phase 2: Communication
Set O2 :{ o21 }
where, o21 : Context Classification

(c) Phase 3: Result
Set O3 : { o31, o32 }
Wher1e,
o31 : DR-Statistic
o32 : DR-Result

4. Function Set 'P' : {P1, P2, P3 }

(a) Phase 1: Registration
Set P1: {p11}
where,
p11 : User registration

(b) Phase 2: Communication
Set P2:{ p21, p22, p23, p24 }
where, p21 : Storage
p22 : Stop Word elimination
p23 : Dataset Learning
p24 : Data analysis

(c) Phase 3: Result
Set P3:{ p31, p32 }
where,
p31 : SR Statistic
p32 : SR Result

5. Success Conditions: If the systemis able to give proper inference of the Disease for the given user Queries.

6. Failure Conditions: If system fails to give proper inference of Disease.

9. CONCLUSIONS

Deep learning methods have recently made notable advances in the tasks of classification and representation learning. In this work we demonstrate our results (and feasible parameter ranges) in application of deep learning methods to structural and functional information. Thus the proposed system involves an efficient machine learning approach for mining health related data. The hidden layers between the input and output layers are incrementally increased based on the accuracy. Present study reveals the community-based health services. This proposed system presents a sparsely connected deep learning scheme that is able to infer the possible diseases given the questions of health seekers. In future it can be used in clinic for diseases differencing according to users symptoms.

9. REFERENCES

[1]. 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.

[2]. 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.

[3]. S. Doan and H. Xu, "Recognizing medication related entities in hospital discharge summaries using support vector machine," in *Proceedings of the International Conference on Computational Linguistics*, 2010

[4]. LiqiangNie, Men Gang, Luming Zhang &Shuicheng Yan (2015), 'Disease Inference from Halth-Related Questions Via Sparse Deep Learning', IEEE Transaction on Knowledge and Data Engineering, Vol.27, No.8.

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

[6]. L. Nie, Y.-L. Zhao, X.Wang, J. Shen, and T.-S. Chua, Learning to recommend descriptive tags for questions in social forums, Acm Transactions on Information System, 2014.