Disease Inference System based on Health-Related Question Answer system

[1] Paresh Karande [2] Manoj Khairavkar, [3] Vinayak Lokhande, [4] Khemchand Mahajan [5] Rupali Umbare
[1][2][3][4][5] Department of Computer Engineering, SavitribaiPhule Pune University, Pune, India
[1] pareshkara888@gmail.com [2] manojk3535@gmail.com [3] vinayakslokhande@gmail.com
[4] mahajankhemchand@gmail.com [5] umbarerupali1@gmail.com

Abstract: - **Health is one of the expanding subjects used for assessing health condition among patients who suffer from specific ailment or diseases. We aims to model the relationship between the Health variables using integrated model of inference system and linear regression. Linguistic data were collected by a guided interview and fed into the deep sparse inference system to yield Health indices. We next propose a novel significant learning plan to surmise the conceivable maladies given the inquiries of well-being seekers. The proposed plan is embodied two key parts. The main part mines the discriminant therapeutic marks from crude elements. The second esteems the crude components and their marks as info hubs in one layer and concealed hubs in the consequent layer, individually. In the interim, it takes in the between relations between these two layers by means of pre-preparing with pseudo-marked information. Taking after that, the shrouded hubs serve as crude components for the more unique mark mining. With incremental and option rehashing of these two segments, our plan manufactures a scantily joined profound construction modelling with three shrouded layers.**

Keywords— **cloud, learning, neural network, back-propagation, privacy preserving, data classification, cipher text**

I. INTRODUCTION

Medical care and research are the most vital part of science for humans, as none of us are immune to physical ailments and biological deterioration. Today we are not able to give time for our health which we should. So because of this unconcern approach we are more prone to diseases. The rapidly increasing medical concern of the baby boomer generation is one major factor stressing the health care system. Many of us are surfing internet to get any disease related information but still they did not get the appropriate information they require so for them our system will give accurate information. Disease Inference system which will give the disease information which he/she is facing on the basis of health related questions. In a less amount of time he/she will get to know what he/she is facing and that to by sitting at the home. We are also providing them the nearest doctor suggestion which he can consult for his treatment. Using our system health seeker will get immediate response as compared to the existing system. Our approach is distinctly different in that we are trying to build a general predictive system which can utilize a less constrained feature space, i.e. taking into account all available demographics and previous medical history. Moreover, we rely primarily on predictions to account for the previous medical history, rather than specialized user inputs.

II. RELATED WORK

Deep learning methods have recently made notable advances in the tasks of classification and representation learning. Deep learning allows computational models that are composed of multiple processing layers to learn representations of data with multiple levels of abstraction. These methods have dramatically improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains such as drug discovery and genomics. T. C. Zhou, M. R. Lyu, and I. King [1] have proposed the concept of Inquiry Routing. There is a serious gap between the existing open questions and potential answerers. To bridge the gap, they present a new approach to Question Routing, which aims at routing open questions to suitable CQA users who may answer these questions. They developed and evaluated a variety of local features, including question features, user history features, and question-user relationship features. Developing several global features, and integrating them with local features to further enhance the classification performance. D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi in [2] proposed CARE, a Collaborative Assessment and Recommendation Engine, which relies only on a patient's medical history using ICD-9-CM codes in order to predict future diseases risks. They also describe an Iterative version, ICARE, which incorporates ensemble concepts for improved performance. There novel systems require no specialized information and provide predictions for medical conditions of all kinds in a single run. Existing group inquiry noting discussions normally star vide just printed answers so L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua in [3] proposed a method to generate queries from QA pairs for multimedia search and perform query-dependent re-ranking for image and video data obtained from search engines by analysing visual features. They investigate the prediction of appropriate answer medium. L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua in [4] they proposed a scheme that is able to enrich textual answers in cQA with appropriate media data. Shouman et al. [6] and Ghumbre et al. [7] have respectively explored decision tree and SVM in the inference of heart disease, which is the leading cause of death in the world over the past 10 years according to the report from world health organization. The existing work was monotonously conducted on retrieval Within the scope of this paper, we explain the classification techniques. The architecture provides the service to implement the classification of user provided query. The concept of Question Routing will help us in boosting users' adhesiveness and loyalty to the system and increases the efficiency of the system. Majorly it will help us in making the system self-learning, which is our ultimate aim.

III. DATA COLLECTION

Open Government Data (OGD) Platform India - data.gov.in - is a platform for supporting Open Data initiative of Government of India. The portal is intended to be used by Government of India Ministries/ Departments their organizations to publish datasets, documents, services, tools and applications collected by them for public use. It intends to increase transparency in the functioning of Government and also open avenues for many more innovative uses of Government Data to give different perspective.

We collected more than 900 popular disease concepts from EveryoneHealthy5, WebMD and Medline-Plus. They cover a wide range of diseases, including endocrine, urinary, neurological and other aspects. With these disease concepts as queries.

IV. DATA UPLOAD

Data transferring action is finished transmission of a document from Laboratory framework to Server. From a system client's perspective, to transfer a record is to send it to server that is set up to get it. Individuals who offer information with others on transfer administrations (US) transfer records. The File Transfer Protocol (FTP) is the Internet facility for downloading and uploading files.

V. 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-labelled data and further strengthened by fine-tuning with online doctor labelled data.

VI. SYSTEM ARCHITECTURE

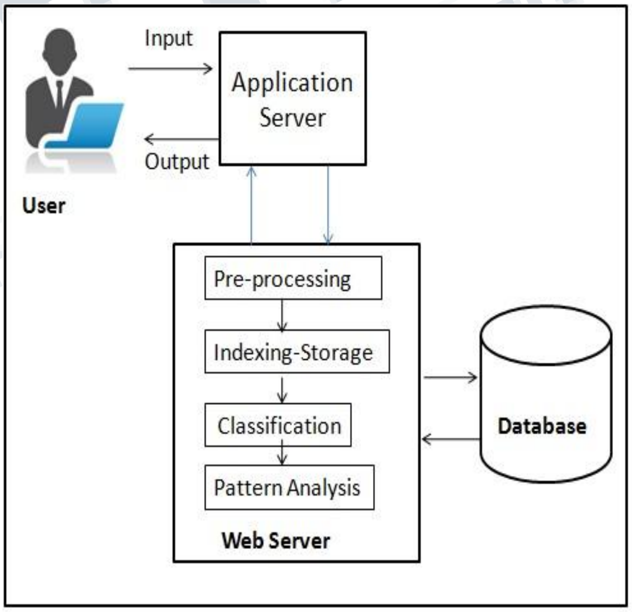

Figure 1: System Architecture

1) In this architecture the four main components are the Health Seeker, Application Server, Web Server, 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 Authentication of user, creation of new user, deletion of user, 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. Basically application server maintains the GUI of the System. All Processing on Query is done in Web server.

3) Query's passed to Web server is processed first for stop word elimination, for that Natural Language Processing (NLP) is used. 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 using Decision Tree. All the Relevant Data we got is then passed further for its Classification.

5) In Classification Block we use SVM (Support Vector Machine) for classifying the Data available. Here we Divided data into different classes. As SVM is the Best Algorithm till now which gives guaranteed classification of data given, we use it with KNN (K-Nearest Neighbor).We proposed to use K-NN with SVM. SVM is best for Classification and further these classified data is given to K-NN for Pattern analysis. Layer by layer elimination of entries is done by SVM. 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 through Application server in the form of the inferred possible disease.

VII. PROPOSED SYSTEM

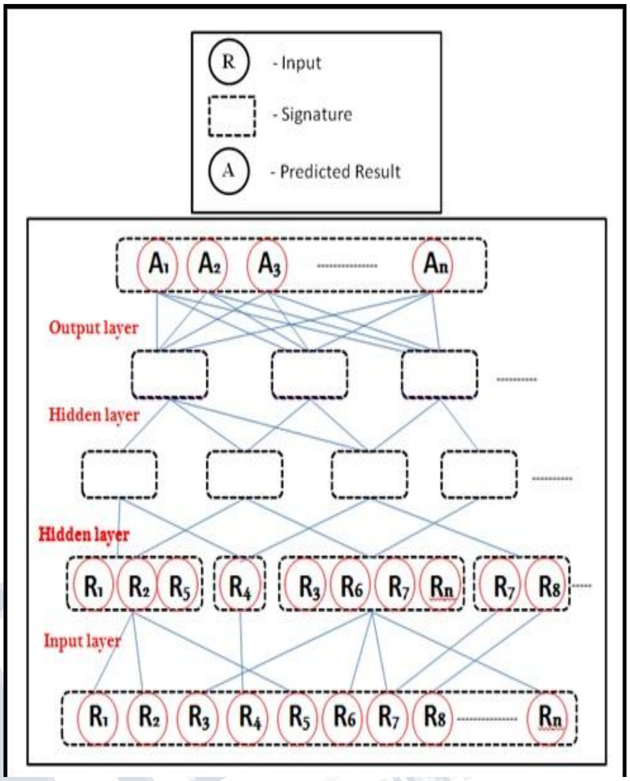

Figure 2: Hidden Layer Architecture.

In this paper we proposed to develop a system which is Efficient, Self-Learning, a flexible and accurate system by the use of SVM with K-NN. This work good in lesser Time complexity means Health seeker will get his answer to his query within a small amount of time and that to accurate to large extend. We are making this system self-learning so that if same query is repeated again with same parameters then he/she will get answer directly. Beyond extraction, constructed entity graphs by exploring their co-occurrence relations and studied how to leverage such graphs to convert raw entity mentions into more useful knowledge, which is helpful for feature expansion. These efforts only consider the explicitly present medical entities, while they overlook the temporal aspect of data as well as the latent discriminative patterns across patient records .To deal with these two problems, proposed a nonnegative matrix factorization based framework to mine common and individual shift-invariant temporal patterns from different events over different patient groups, which is able to handle sparseness and scalability problems.

VIII. ALGORITHMIC DETAILS

8.1 Support Vector Machine (SVM) & KNN

Train an SVM on each of its leaves, using samples that flow to each leaf as training data. A tree and all SVMs associated with its leaves constitute a DTSVM classifier. In the training phase, all the SVMs in a DTSVM classifier are trained with the same parameter values. In the validation or the testing phase, we input a given data point x to the tree. If x reaches a homogeneous leaf, we classify x as the common class type of those samples; otherwise, we classify it with the SVM associated with that leaf. When a learning data set is given, we divide a given learning data set into a training and validation constituent. We then build a DTSVM classifier on the training constituent and determine its optimal parameter values with the help of the validation constituent.

8.2 Support Vector Machine (SVM) Algorithm

SVM is used to classify and analysis the data.It is used for machine learning.SVM is developed by COLT in 1992 by Boser,guyon, and vapnic. SVM training algorithm builds a model that assigns new examples into one category or the other, making it a non-probabilistic binary linear classifier. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible.The SVM problem (primal) is to find the decision surface that maximizes the margin between the data points of the two classes.SVM is supervise learning algorithm.It uses training data set do perform operation. It is mostly used in Image processing, medical field, cryptography.

8.3 K-Nearest Neighbors

The K-Nearest neighbour(KNN) algorithm measure the distance between a query and a set of scenarios in the data set. A predefine similarly metric is used to find the K most. The neighbors are taken from a set of objects for which the class (for k-NN classification) or the object property value (for k-NN regression) is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

IX. CONCLUSION

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. We also describe a novel constraint-based approach to high dimensional data analysis in medical field. We use it to analyze the effect of parameter choices on data transformations. Our method is able to learn important representations and detect latent relations in health and medical data.

REFERENCES

1) T. C. Zhou, M. R. Lyu, and I. King, “A classification-based approach to question routing in community question answering,” in The International World Wide Web Conference, 2012

2) D. A. Davis, N. V. Chawla, N. Blumm, N. Christakis, and A.-L. Barabasi, “Predicting individual disease risk based on medical history,” in The International Conference on Information and Knowledge Management, 2008.

3) L. Nie, M. Wang, Z. Zha, G. Li, and T.-S. Chua, “Multimedia answering: Enriching text qa with media information,” in Proceedings of the International ACM SIGIR Conference, 2011.

4) L. Nie, M. Wang, Y. Gao, Z.-J. Zha, and T.-S. Chua, “Beyond text QA: Multimedia answer generation by harvesting web information,” Multimedia, IEEE Transactions on, 2013.

5) 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.

6) M. Shouman, T. Turner, and R. Stocker, “Using decision tree for diagnosing heart disease patients,” in Proceedings of the Australasian Data Mining Conference, 2011.

7) S. Ghumbre, C. Patil, and A. Ghatol, “Heart disease diagnosis using support vector machine,” in Proceedings of the International Conference on Computer Science and Information Technology, 2011.

8) S. Fox and M. Duggan, Health online 2013, Pew Research Centre, Survey,2013.

9) Online health research eclipsing patient-doctor conversations, Makovsky Health and Kelton, Survey, 2013.

10) 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.