International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 03 Issue: 04 | Apr-2016 www.irjet.net p-ISSN: 2395-0072

DISEASE INFERENCE SYSTEM FROM HEALTH-RELATED QUESTIONS VIA LEARNING APPROACH

DEEPTHI M[1], POONGOTHI T[2]

[1]*M Tech Student, Department of Information Technology, K. S. R. College of Engineering, Tiruchengode, Tamil Nadu*
[2]*Associate Professor, Department of Information Technology, K. S. R. College of Engineering, Tiruchengode, Tamil Nadu*

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 training samples. In this paper, user will search for the disease diagnosis by giving symptoms as a query in the search engine. We next preprocess the query to find the symptoms keyword. The symptom keyword which is preprocessed is matched with diseases stored in the local database to identify the corresponding disease. In order to make the process time consuming, PCA-L1 classify was applied to classify disease into subgroups. If higher preference is given to that subgroup and searching in that new smaller subgroup thus reduces database access. At last, k-nearest neighbor algorithm is a method for classifying objects based on closest training examples in the feature space. Extensive experiments on a real-world dataset labeled by online doctors show the significant performance gains of our scheme.*

Keywords: Support vector Machine, OSC-NMF, PCA, Decision tree, Margin-based Censored Regression Approach, Novel Multi-Task learning techniques.

1.INTRODUCTION

The turning gray of society, heightening expenses of medicinal services furthermore, thriving PC advances are as one driving more shoppers to invest longer energy online to investigate health data. One study demonstrates that 59% of U.S. grown-ups have investigated the web as a symptomatic instrument in 2012. Another review reports that the normal U.S. shopper spends near 52 hour every year online to discover health learning, while just visits the specialists three times every year in 2013. The current programmed question answering procedures are relevant here.

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 input.

2. RELATEDWORK

Due to lack of data sets, they have undergone studies on machine learning based approaches to extract named entities. The types of medication

related named entities including medication names, dosage, mode, frequency, duration and reason from hospital discharge summaries. So many machine learning based systems have been developed and showed good performance. Those systems involve two steps: Frist, recognition ofmedication related entities and Second, determination of the relation between a medication name and its modifiers. A machine learning algorithms such as Conditional Random Field (CRF) and Maximum Entropy have been applied to the named entity recognition task. In this study, SVM-based NER system for medication related entities was developed. This system investigates different types of features and our results showed that by combining sematic features from a rule-based system, the SVM-based NER system could achieve the best F-score of 90.05% in recognizing medication related entities [11].

In this paper, it deals with how unsupervised feature learning can be used for cancer detection and cancer type analysis from gene expression data. The advantage of the proposed method is the possibility of applying data from various types of cancer to automatically form features which help to enhance the detection and diagnosis of a specific one. PCA is used to address the very high dimensionality of the initial raw feature space followed by sparse featurelearning techniques to construct discriminative and sparse features for the final classification step in this method. This method shows best performance and it need only very limited size data sets [12].

Alzheimer's disease is the most common neurodegenerative disorder associated with aging. In this paper, a novel multi-task learning technique to predict the disease progression measured by cognitive scores and select biomarkers predictive of the progression was developed. The prediction of cognitive scores at each time point is considered as a task in multi-task learning. A novel convex fused sparse group lasso formulation was proposed to allow the simultaneous selection of a common set of biomarkers for multiple time points and specific sets of biomarkers for different time points using the sparse group lasso penalty. The proposed two non-convex formulations, which are expected to reduce the shrinkage bias in the convex formulation. Weplan to extend our formulations to deal with missing data [8].

Extracting useful patterns from large collection of electronic clinical record is particularly challenging because it is longitudinal, sparse, and heterogeneous. In this paper, A nonnegative matrix factorization based framework using a convolutional approach for open-ended temporal pattern discovery over large collections of clinical records is developed. In addition, it uses an event matrix based representation that can encode quantitatively all key temporal concepts including order, concurrency, and synchronicity. The experimental results based on both synthetic and real world electronic patient data are presented to demonstrate the effectiveness of the proposed method [7].

Community based question and answering administrations have conveyed clients toanother time of information spread by permitting clients to make inquiries and to answer other clients questions. In any case, because of the quick expanding of posted questions and the absence of a compelling approach to discovering trigging questions, there is a genuine crevice between posted questions what's more, potential answers. This hole may debase a CQA administration's execution and additionally lessen clients' reliability to the framework. To conquer any hindrance, we show another way to deal with inquiry routing, which goes for directing inquiries to members why should likely give answers. We consider the issue of inquirysteering as an arrangement assignment, and add to an assortment of nearby and worldwide elements which catch distinctive parts of inquiries, clients, and their relations. Author likewise perform a systematicalcorrelation on how distinctive sorts of components add to the last results what's more; demonstrate that question-client relationship components play a key part in enhancing the general execution [13].

Developers have been using several data mining techniques to help health care professionals in the diagnosis of heart disease and after study; Decision Tree is one of the successful data mining techniques used. But, most of the researcher applied J4.8 Decision Tree based on Gain Ratio and binary discretization. The other two successful type of Decision tree that are less used in the diagnosis of heart disease are Gini Index and Information Gain. Apart fromthis technique, voting method, and

reduced error pruning are known to produce more accurate Decision Trees [14].

In numerous text classification applications, it is engaging take each archive as a series of characters as opposed to a sack of words. Past examination thinks about around there for the most part concentrated on distinctive variations of generative Markov chain models. Albeit discriminative machine learning strategies like support vector machine have been very fruitful in content grouping with word highlights, it is not one or the other viable nor productive to apply them clearly taking all substring in the corpus as elements. In this paper, author proposes to parcel allsubstrings into measurable comparability gatherings, and afterward pick those gatherings which are essential as components for content grouping. Specially, author is proposing an addition tree based calculation that can remove such elements in direct time. Authors are examinations on English, Chinese and Greek datasets demonstrate that SVM with key-substring-gathering components can accomplish exceptional execution for different content grouping assignments [4].

3.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, service, 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 Everyone Healthy, WebMD, and MedlinePlus. They cover a wide range of diseases, including endocrine, urinary, neurological and other aspects with these disease concepts as queries.

4. EXISTING SYSTEM

Disease inference is a thinking outcome in view of the given inquiries; this undertaking is non-trifling because of taking after reasons. In this paper, Input is given in the form of questions and selects those that ask for possible disease of their manifested symptoms for further analytic. It consists of two key components. First, mines the medical signatures from raw features. Second,the raw features and their signatures as input nodes in one layer and hidden nodes in the subsequent layer. It finds the inter-relations between these two layers by using pre-training. By repeating the steps, sparsely connected deep architecture was formed. Finally, fine-tuning was used to fit exact disease from raw feature.

Disadvantage

The vocabulary gap between diverse health seekers makes the data more inconsistent, as compared to other formats of health data.

5. IMPLEMENTATION

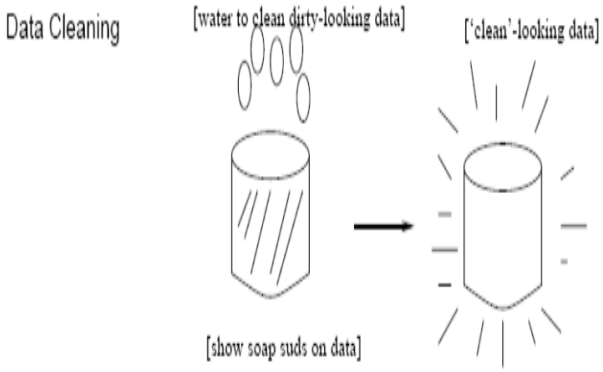
5.1 DATA PREPROCESSING

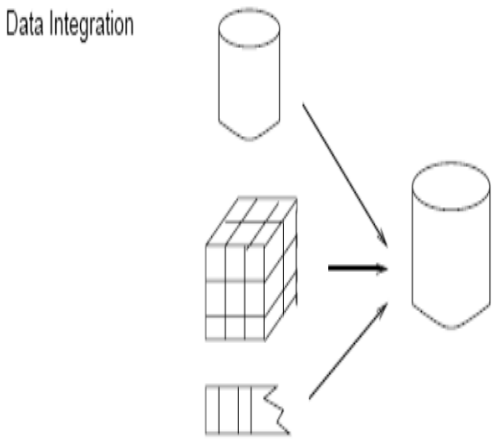
Data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. In this phase, noisy and irrelevant data are removed from extracted data. Then the stop words like a, an, the, is, was, etc.., are removed. After eliminating the human errors, unwanted words like filler words were removed. Followed by that, stemming is done, which is the process of removing morphological and in flexional ending words to their root words. Finally the semantic word extraction is performed and it is stored in the local database.

Data preprocessing is a data mining technique that involves transforming raw data into an understandable format. Real world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a proven method of resolving such issues. Data preprocessing prepare raw data for further processing. Data preprocessing is used database-driven applications such as customer relationship management and rule-based applications.

Data goes through a series of steps during preprocessing as shown in Fig 1:

- Data Cleaning: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.
- Data Integration: Data with different representations are put together and conflicts within the data are resolved.
- Data Transformation: Data is normalized, aggregated and generalized.
- Data Reduction: This step aims to present a reduced representation of the data in a data warehouse.
- Data Discretization: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.





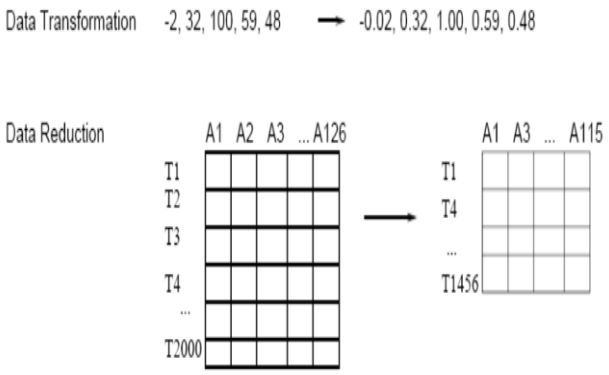


Figure 1: Forms Of Data Preprocessing

5.2 DISEASE GROUPING

In this module the keyword which is a preprocessed symptom is matched with the diseases stored in the local database to identify the corresponding disease related to those symptoms given by the user. This has to search a record database of more than 20000 diseases and even more symptoms, which is very time consuming, so PCA-L1 classification was applied to classify diseases into subgroups. If a group of symptoms match higher preference is given to that subgroup and searching in that new smaller subgroup thus reduces database access. In pattern recognition, the k-nearest neighbor algorithm is a method for classifying objects based on closest training examples in the feature space. PCA-L1 is a type of instance based learning or lazy learning where the function is only approximated locally and all computation is deferred until classification. PCA-L1 has been modified to give faster processing as follows. Here, weights have been given to the individual symptoms corresponding to each disease, and the subclass disease category.

In data analysis problem with a large number of input variables, dimensionality reduction methods are typically used to reduce the number of input variables to simplify the problems without degrading performances. Among them, the principal component analysis is one of the most popular methods. In PCA, one tries to find a set of projections that maximize the variance of given data.

From now on, we derive a new algorithm to solve. The optimization of this objective function is

difficult because it contains absolute value operation, which is nonlinear. In order to find the projection vector w that maximizes this L1 objective function, the following algorithm is presented. We refer to the algorithm as PCA-L1 to differentiate it from the L1-PCA as shown in Fig 2.

lgorithm:PCA-L1

1) Initialization: Pick any w(0). Set w(0) ←w(0)/||w(0)||2 and t = 0.

2) Polarity check: For all i Є {1,….,n}, if w T(t)x_i<0, p_i(t) = -1, otherwise p_i(t) = 1.

3) Flipping and maximization: Set t ← t + 1 and w(t) = p_ni = 1 p_i(t – 1)x_i. set w(t) ← w(t)/||w(t)||2.

4) Convergence check:

a. If w(t) 6 = w(t – 1), go to Step 2.

b. Else if there exists i such that w T(t)x_i = 0, set w(t) ← (w(t) + Δw)/||w(t) + Δw||2 and go to Step 2. Here, Δw is a small nonzero random vector.

c. Otherwise, set w* = w(t) and stop.

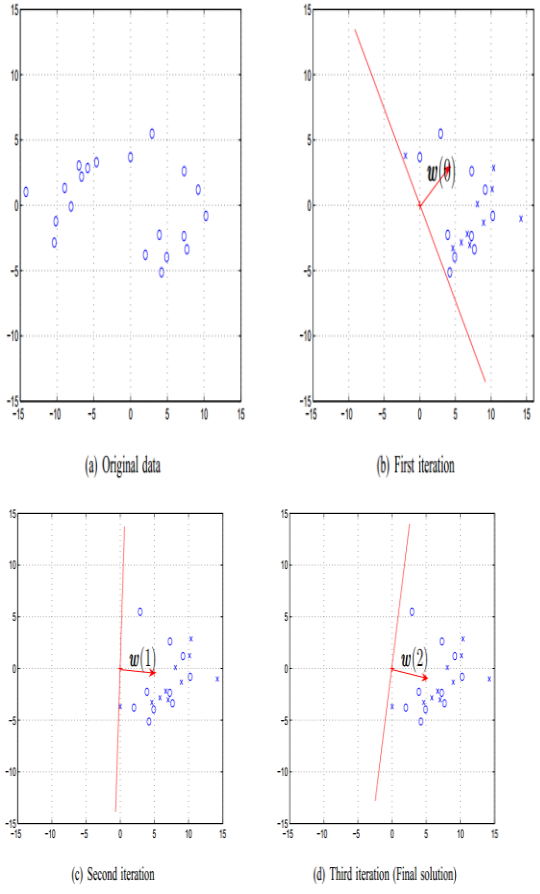


Fig 2: PCA-L1

5.3WEIGHT ASSIGNING

Multiple disease identification after grouping the nearest matched symptoms are found. The final report is then mined to obtain the correct symptoms. The correct symptom thus obtained is then compared with the original symptoms entered. This information is now fed to the LAMSTAR network for assigning weights. Information in the LAMSTAR network is encoded via correlation links between individual neurons in different self-organizing maps modules. This helps to avoid them need to consider a very large number of links, thus contributing to the network efficiency.

Initially, all weights are assigned the value zero. For each correct matched symptom the weight increases +1 and for each unmatched symptom the weights are kept constant. The reason for keeping weights constant for unwanted symptoms is that if the unmatched symptoms were assigned negative weights, then certain symptoms would be repeatedly degraded and when they would actually surface in some diagnosis, because of too much negative weight, the change in the ratio of weight of the symptom for that particular disease to the total weight of all symptoms for the disease will be more significant.

This will lead small changes in trend to result in bigger change in the ratio as compared to not subtracting negative weight. To keep the weight ratio to be as stable and precise as possible, the fluctuations should not be much. Hence, only positive weights are considered. Since all information in the LAMSTAR network is encoded in the correlation links, the LAMSTAR can be utilized as a data analysis tool. In this case the system provides analysis of input data such as evaluating the importance of input sub words, the strengths of correlation between categories, or the strengths of correlation of between individual neurons.

The system's analysis of the input data involves two phases:

- Training of the system
- Analysis of the values of correlation links created after the training.

Since the correlation links connecting clusters among categories are modified in the training phase, it is possible to single out the links with the highest values. Therefore, the clusters connected by

the links with the highest values determine the trends in the input data. In contrast to using data averaging methods, isolated cases of the input data will not affect the LAMSTAR results, noting its forgetting feature. Furthermore, the LAMSTAR structure makes it very robust to missing input sub words. After the training phase is completed, the LAMSTAR system finds the highest correlation links and reports messages associated with the clusters in SOM modules connected by these links.

5.4 PATTERN MATCHING

The pattern recognition aims to classify data based on either a priori knowledge or on statistical information extracted from the patterns. The patterns to be classified are usually groups of measurements or observations, defining points in an appropriate multidimensional space. A complete pattern recognition system consists of a sensor that gathers the observations to be classified or described; a feature extraction mechanism that computes numeric or symbolic information from the observations and a classification or description scheme that does the actual job of classifying or describing observations, relying on the extracted feature. The classification or description scheme is usually based on the availability of a set of patterns that have already been classified or described. This set of patterns is termed the training set and the resulting learning strategy is characterized as supervised. Learning can also be unsupervised, in the sense that the system is not given an a priori labeling of patterns, instead it establishes the classes itself based on the statistical regularities of the patterns.

Pattern matching using iterative search utilizes data that is stored. The first step of the algorithm involves selecting the symptoms shown by the patient. As an output, the algorithm gives the list of all possible diseases ranked according to the number of symptoms matched in the database. The list is generated after input of every symptom. After the first iteration, for the second iteration, the next list of symptoms will be shortlisted according to the disease list that was obtained in the previous iteration. The new symptom list will contain symptoms of only those diseases that were obtained in the previous list. Theserelated symptoms will then be shown to the user who shortlists another symptom from the new list. The new disease list will be listed, ranked according to the number of symptoms matched. The ranking is generated according to the percentage match of the total number of symptoms entered. This procedure goes on iteratively, with diseases being placed in the ranks according to its probabilities.

If a single disease in the given subset gains maximum weight above all other diseases, it is the interpreted by the system as the possible diagnosis. The databases and data warehouses become more and more popular and imply huge amount of data which need to be efficiency analyzed. Knowledge Discovery in Databases can be defined as the discovery of interesting, implicit, and previously unknown knowledge from large databases.

5.5 DISEASE DIAGNOSIS

To perform differential diagnosis, the system uses a Hopfield network. They serve as content-addressable memory systems with binary threshold units. They are guaranteed to converge to a local minimum, but convergence to one of the stored patterns is not guaranteed. The value is determined by whether or not the units' input exceeds their threshold. They are also rigorous but suffer from the curse of dimensionality regarding number of categories they can handle efficiently, and are sensitive to incomplete data sets. Hopfield nets can either have units that take on values of 0, 1, or -1. Hopfield networks are constructed from artificial neurons. These artificial neurons have N inputs. With each input i there is a weight $w_i$associated.

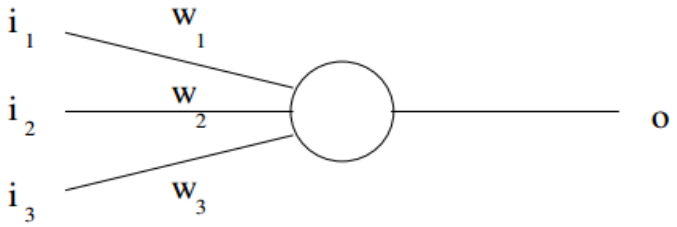


Figure 3: An artificial neuron as used in a Hopfield network.

They also have an output. The state of the output is maintained, until the neuron is updated.

6. CONCLUSION

In this paper, we proposed a method to detect disease inference from queries given by health seeker and with the help of data set collected from various source. The proposed method, which uses PCA-L1 classifier to address the very high dimensionality of the initial raw feature and it limit the size of data set. After reduction of data set, KNN algorithm was used to detect the exact disease by verifying queries given by user.

REFERENCES

[1] 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.

[2] AditysKhosla, Yu Cao, Cliff Chiung-Yu Lin, Hsu-Kuang Chiu, Junling Hu, &Honglak Lee (2010), 'An Integrated Machine Learning Approach to Stroke Prediction', ACM, Pp. 183-192.

[3] CeyhunBurakAkgul, DevrimUnay&AhmetEkin (2009), 'Automated Diagnosis of Alzheimer's Disease Using Image Similarity and User Feedback', ACM, Pp. 34.

[4] Dell Zhang & Wee Sun Lee (2006), 'Extracting Key-Substring Group Features for Text Classification', ACM, Pp. 474-483.

[5] D.A. Davis, N. V. Chawla, N. Blumm, N. Christakis &A. L. Barabasi (2008), 'Predicting Individual Disease Risk Based on Medical History', Proc. 13thInt. Cont. Inf. Knowl. Manage., Pp. 769-778.

[6] F. Wang, N. Lee, J. Hu, J. Sun, S. Ebadollahi &A. Laine (2013), ' A Framework for Mining Signature from Event Sequences and its Application in Healthcare Data', IEEE Trans. Pattern Anal. Mach. Intell., Vol. 35, No. 2, Pp. 272-285.

[7] Fei Wang, Noah Lee, Jianying Hu, Jimeng Sun, &Shahram Ebadollahi (2012), 'Towards Heterogeneous Temporal Clinical Event Pattern Discovery: A Convolutional Approach', ACM, Pp. 453-462.

[8] Jiayu Zhou, Jun Liu, VaibhayA. Narayan &Jieping Ye (2012), 'Modeling Disease Progression Via Fused Sparse Group Lasso', Acm, Pp. 1095-1103.

[9] L. Nie, M. Wang, Z. Zha, G. Li & T. S. Chua (2011), 'Multimedia Answering: Enriching Text with Media Information",Proc.Int.AcmSigir Conf. Res.Develop.Inf. Retrieval, Pp. 695-704

[10] K. Nie, Y. L. Zhao, X. Wang, J. Shen&T. S. Chua (2014), 'Learning to Recommend Descriptive Tags for Questions in Social Forums', ACM Trans. Inf. Syst., Vol. 32, No. 1, Pp. 5.

[11] Son Doan & Hua Xu (2010), 'Recognizing Medication Related Entities in Hospital Discharge Summaries Using Support Vector Machine',, Coling, pp. 259 – 266.

[12] Rasool Fakoor, Faisal Ladhak, Azade Nazi, Manfred Huber (2013), 'Using Deep Learning to Enhance Cancer Diagnosis and Classification', Proceeding of the 30th International Conference on Machine Learning', Vol. 28.

[13] Tom Chao Zhou, Michael R. Lyu, Irwin King (2012), 'A Classification based Approach to Question Routing in Community Question Answering', IW3C2, ACM, Vol.04.

[14] Mai Shouman, Tim Turner, Rob Stocker (2011), 'Using Decision Tree for Diagnosing Heart Disease Patients', CRPIT, Vol. 121.