ISSN 2320 - 2602
Volume 3, No.2, February 2014
International Journal of Advances in Computer Science and Technology
Available Online at http://warse.org/pdfs/2014/ijacst13322014.pdf

SYMPTOM'S BASED DISEASES PREDICTION IN MEDICAL SYSTEM BY USING K-MEANS ALGORITHM

[1]Sathyabama Balasubramanian, [2]Balaji Subramani,
[1] M.Tech Student, Department of Information Technology,
[2] Assistant Professor, Department of Information Technology,
V.S.B Engineering College, Karur, Tamilnadu, India
sathyadharshana@gmail.com
hereiambalaji@gmail.com

ABSTRACT

Medical diagnosis is an on-going research in medical trade. Here the prediction of various diseases like heart, lungs and various tumours supported the past data collected from the patients may be terribly troublesome task. These are not applicable for whole medical dataset. During this paper the diagnosis may be created and supported the historical knowledge. To compute the chance of prevalence of explicit unwellness from medical knowledge by using k-mean ,Large memory Storage and Retrieval(LAMSTAR) and Medical diagnosis methodology. The system uses service oriented architecture (SOA) whereby the system elements of diagnosis, data portal and alternative miscellaneous services are provided. This reduces the multiple diseases showing the similar symptoms problem and it will increase the accuracy of such diagnosis.

Key words*:* Data Mining, Knowledge discovery, Service oriented architecture (SOA) and differential diagnosis, LAMSTAR Network.

1. INTRODUCTION

Since the arrival of advanced computing, the doctors' still requires the technology in various possible ways like surgical representation process and x-ray photography, but the technology perceptually stayed behind. The method still requires the doctor's information and experience due to alternative factors starting from medical records to weather conditions, atmosphere, blood pressure and numerous alternative factors. The huge numbers of variables are consider as entire variables that are required to understand the complete working process itself, however no model has analyzed successfully. To tackle this drawback, Medical decision support systems must be used. This system is able to assist the doctors to make the correct decision.

Medical decision support system refers to both the process of attempting to determine or identify possible diseases or disorder and the opinion reached by this process. The diagnostic opinion in the sense, it indicates either degree of abnormality on a continuum or a kind of abnormality in a classification. It's influenced by non medical factors such as power ethics and financial incentives for patient or doctor. It can be a brief summation or an extensive formulation, even taking the form of story or metaphor. It might be a means of communication such as computer code through which it triggers payment, prescription, notification, information or advice. Indication of medical diagnostic includes knowledge of what is normal and measuring of patient's current condition. Automated decision support systems are rule based systems that are automatically providing solutions to repetitive management problems.

Medical decision could be extremely specialized and difficult job due to alternative factors or incase of rare diseases. The alternative factors include stress; tired misdiagnosis might vary from ignorance of doctors and incomplete information. Standard algorithm may go through the entire variables like prevailing conditions history of medical records, history of family records and various factors relating to the patient records, sheer magnitude of obtainable hidden factors.

Differential diagnosis methods can be used to identify the presence of an entity where multiple alternatives are possible and also refers to include the candidate alternatives. This method is needs a process of elimination or obtaining information that shrinks the probability of candidate conditions to negligible levels. It contains four steps: 1) The doctor gather all information about the patients and create a symptoms list.2) The doctor should make a list of all possible causes of symptoms.3) The doctor should prioritize the list by which is the most dangerous possible cause of symptoms put in the top of the list.4) The doctor should rule out or treat the possible causes beginning with the most urgently dangerous conditions."Rule Out" in the sense to use

the test method or other scientific method. If there will be no such diagnosis means removing the diagnosis from the list and using tests that should have distinct results, depends on which diagnosis is correct. This can be done based on the doctor's knowledge and experience. This method is very easy to implement.

To reduce the large number of variables and find the most probable diseases by using the K-Means algorithm. This algorithm is more suitable to cluster the more number of diseases. K-Mean is one of the unsupervised learning algorithms which are used to solve the clustering problem. The main idea is to determine the k centroids, one for each cluster. Different tests performed on the patients will served as a attributes for clustering. By using this algorithm it reduce the number of iterations, boundries of clusters are well define without overlapping, to produce the accurate result for each and every diagnosis. This system uses Service oriented architecture (SOA), anyone can access with internet connections and LAMSTAR Network can be used to calculate the weight, to increase the accuracy of algorithm, overall speed test and produce the better result.

2. RELATED WORK

In this section the previous related studies are reviewed. Here the lists of symptoms are required medication for every possible disease consists of both accurate and inaccurate. This paper describes the features of such diagnosis and related symptoms for such diseases [2].But this site may not give total information about symptoms/diagnosis that is not relevant to the patients (i.e.) family records or some other factors.

Iliad is an expert diagnostic system which is used to explain the relationships for finding the diseases. This system uses the Bayesian classification to compute the probability for possible diagnosis [3]. DXplain is a medical decision support system [4]; it generates the ranking for list of diagnosis which is the mostly likely diseases yielding the lowest rank. Using stored information, each disease prevalence and significance, the system differentiates the common diseases and rare diseases. This system also serves as a clinician reference with a searchable database of diseases and clinical manifestations.

Clinical decision support system is used to determine the diagnosis of patient records [5]. It contains three broad categories: 1) Improve the patient safety.2) Improve the quality of care.3) Improve the efficiency in health care delivery. Patient safety in the sense to reduce the errors and improve the medication. Second category describes to improve the clinical documentation and patient satisfaction. Third category describes to reduce the cost and list of duplications, decrease the adverse of events.

To differentiate the features of all the datasets here use novel classifier based on the bayes discriminate function [6].Hybrid algorithm is used to extract the salient features from the huge biological datasets. Machine learning algorithm is used for the training set. [7] The main objective is to discover the relationship between the attributes which is useful to make the decision. This method avoids the several problems in medical data such as missing values, sparse information and temporal data. Machine learning algorithm is suitable for this kind of data. Two kinds of experiments: 1) To discover association between the attributes.2) Test prediction for future disorder. The result shows that some methods predict some disorders better than others, so interesting to use all the algorithms at a time.

[8] In this paper the data mining framework propose two stages namely clustering and classification. First stage generates two clusters such as cluster-0 and cluster-1.In cluster-0 do not have any disease symptoms and cluster-2 has symptoms. This cluster group is referred to the association of class labels in original dataset. After comparing with original dataset mismatch instances are removed and estimate the accuracy, sensitivity and specificity measures for remaining instances. This will reduce the iteration and increase the accuracy.

The SOM (Self-Organizing Map) is a toolbox which is used to visualize the dataset and mapping the data from higher dimensional input space into lower dimensional space [9].The main goal of SOM is to cause the different parts of the network to respond similarly to certain inputs. The data request and response are in xml format. Here K-Means algorithm is proposed. [10] There are several important properties in SOM .They are stability, compared to the total percentage of system change, isolates consumers in the development of implementation process. Reuse avoids the cost of re-implementation or modification functionality of the encapsulated services.

This paper reviews the principles and several different applications of the LAMSTAR Network [11].The LAMSTAR was specially developed for applications to problems involving very large memory that relates to many

different categories, where some of the data is exact while other data are fuzzy and where for a given problem, some data categories may be totally missing. Consequently, Network has been successfully applied to make decision, diagnosis and recognition problems in various fields.

[12] The knowledge base of the system contains a mathematical extract of a series of cases with known outcome inputted to the training phase. In medical diagnosis situations, the LAMSTAR system can be used as a: 1) teaching aid; 2) diagnosis aid; 3) tool for data analysis; 4) classification tools; and 5) prediction tools. The LAMSTAR network provides multidimensional analysis of input variables and this system does all this without reprogramming per each diagnostic problem. Thus the LAMSTAR network can be very effective in problems where the training domain is not well defined, and where is difficult to create reliable training sets, which exactly the situation one faces in case of medical diagnosis.

3. ARCHITECTURAL MODEL

In architectural model it contains two databases: Patient Records database and Disease/Symptoms database. Four web services are used to implement the SOA. They are Pattern matching, recent trends, differential diagnosis and recent differential diagnosis. The patient Record database contains all the patient information from all the hospitals in the network. Diseases/Symptoms database is a centralized database.

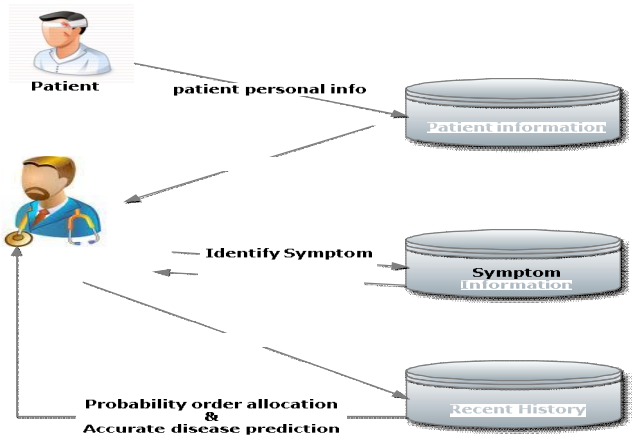

Figure 1: Architectural Model

It contains the list of existing known diseases and their corresponding symptoms along with their weights. These databases are replicated across various servers and these replicated servers are used to achieve the fault tolerance with concurrency protocols to achieve atomic transactions.

First the doctor retrieves the symptoms from the patient record database. After retrieving the symptoms, the doctor identify whether any symptom related diseases contains in the Diseases/Symptoms database. Here the pattern matching service is activated. If any diseases match with Symptoms means list out all the possible matched symptoms and presents the result to the doctors. If the doctors not satisfied with results, compare to recent history and recent trend service must be activated. This service makes use of the Diseases/symptoms database and Patient Record database and the result obtained from pattern matching service to get results.

After comparing the diseases to the recent history, cluster the shortlisted diseases. This list is used to compute the probability of each occurrence of particular diseases from the medical data. The probability may be computed based on the distance vector. The highest priority cluster produces the accurate result. Finally, to avoid the vagueness in decisions, the doctor use differential diagnosis and recent diagnosis features use Diseases/symptoms database and Patient record database and result acquired from recent trend services to gain the results. Since the large medical data, using simple client server architecture would not produce the effective aforesaid services and would increase the response time of the system.

Finally we conclude that SOA was well suited to apply this system because it improve the delivery of important information and sharing of data across the community of healthcare professionals more practical in cost, security and risk deployment. In various existing EHRs, SOA is more essential for data providers to this system, are already using this very successful and efficient architecture. The system enforced as various services in the existing SOA, result in easy implementation, integration and scalability with existing EHRs.SOA also handles the related issues to data security and patient confidentiality.

4. MODULE DESCRIPTION

4.1 Collecting the medical dataset

Electronic Medical Record (EMRs) or Electronic Health record (EHRs) database contains all the Patient related information's. Here the admin allow accessing all the patient related information from the medical database. These

collections of medical data's are used for further processing. Patient details also maintaining the module that means like patient personal details, symptoms and blood pressure level, blood group always maintain this module.

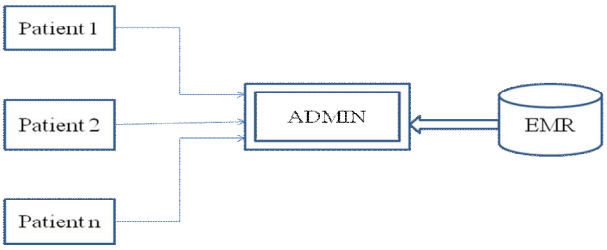

Figure 2: Collecting the medical dataset.

4.2 Symptoms comparing using iterative search

Symptoms' comparing using iterative search employs data that stored in Table 1.

Table 1: Sample Database: Iterative Pattern Search:

Diseases	Symptoms and Weight	Class Weight
Diabetes (D_1)	Headache (W_1), Increase in blood sugar (W_2), Insulin low (W_3)	Endocrine (C_1)
Pericar-ditis(D_2)	Chest pain (W_4), Fever (W_5), weakness (W_6), Malaria (W_7), Shortness of breath (W_8), Syncope (W_9)	Cardiovas-cular (C_2)
Viral Fever (D_3)	Headache (W_{10}), Cold (W_{11}), Fever (W_{12}), Running Nose (W_{13}), Weakness (W_{14})	Parasitic (C_3)
Sinusitis (D_4)	Pain in the sinuses(W_{15}), Headache (W_{16}), Heavy eyebrows (W_{17}), Blur-ry vision (W_{18}), Fever (W_{19})	Respirato-ry (C_4)

In this module symptom matching using iterative search utilize data that is stored. The first step of the algorithm involves selecting the symptoms shown by the patient. 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. From the data in Table I, if *headache*, *fever* and *pain* in the *sinuses* are entered, then the weights W15, W16 and W19 will be considered. Next all the weights will be added and compared to all subclasses C1, C2, C3 and C4 is most likely the answer depending on its weight. Finally all the diseases in class C4 are considered, and if sinusitis (D4) weight is closer to the sum of all the input symptoms weights, then it is possible diagnosis.

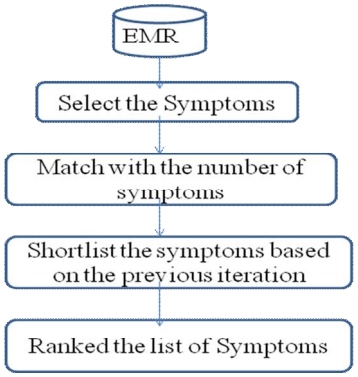

Figure 3: Iterative Search

4.3 Mining Medical Records

In this module if multiple diseases are found with similar ranking, it becomes difficult to pinpoint to one of them, when no more symptom is unique to any single disease affecting its ranking. This especially the case in case of some epidemic in the area, or some rare disease, or disease arising due to localized conditions and various other factors.

4.4 Identifies the Differential Diagnosis

In this module is to perform differential diagnosis, the system uses a Hopfield network. They function as content-addressable memory systems with binary threshold units. They are bound to converge to a local minimum; however convergence to one of the stored patterns is not secured. Hopfield networks can either have units that take on values of 0, 1,-1.There is two restrictions in Hopfield Network: They are,

$W_{ii} = 0$ (no unit has a connection with itself) (1)

$W_{ij} = W_{ji}$ (connections are symmetric) (2)

Table 2: Sample Sorted Database for Differential Diagnosis

Patient	Disease Diagnosed	Actual Disease
A	Diabetes	Diabetes
B	Diabetes	Diabetes
C	Diabetes	Diabetes
D	Diabetes	Diabetes
E	Diabetes	Diabetes
F	Diabetes	Hypertension
G	Diabetes	Hypertension
H	Diabetes	Hypertension
I	Diabetes	Arthritis
J	Diabetes	Arthritis

In Table 2 applied Hopfield rule (2) to find the relative frequencies (ΔW) and calculate the individual weights.

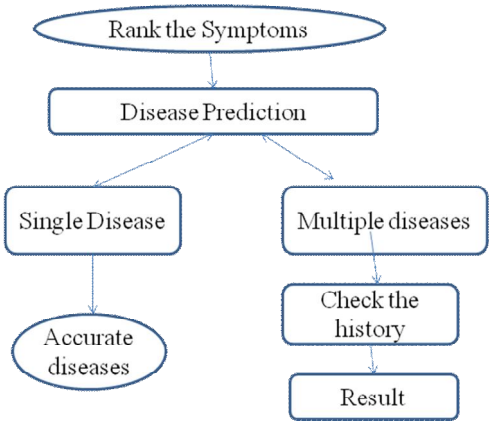

Figure 4: Differential Diagnosis

4.5 Diseases Shortlisted Using LAMSTAR Network

In this Module the correct disease shortlisted by the doctor is obtained who confirms it by taking the necessary tests. 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. If any stored pattern matches the input sub word within a present tolerance, the system updates weights according to the following procedure:

$$W_{i,m}(t+1) = W_{i,m}(t) + \acute{a}_i\,(X_i(t) - W_{i,m}(t)), \text{for } m: €_{min} < €_{(const).}$$

Where

$W_{i,m}(t+1)$ = Modified weights in module I for neuron m;

$\acute{A}_i$ = Learning coefficient for module I;

$€_{min}$ = Minimum error of all weights vectors W_i in module.

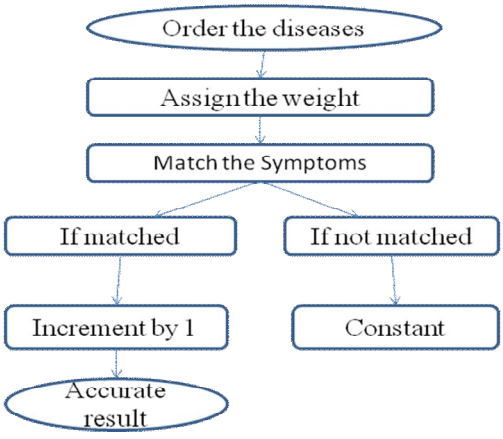

Figure 5: Diseases Shortlisted

5. THE K-MEAN CLUSTERING ALGORITHM

This part briefly describes the K-Means algorithm. K-Means algorithm is a typical clustering algorithm in data mining and which is widely used for clustering the large set of datas.In 1967, Macqueen was firstly proposed the K-Means algorithm, it was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem. This method is used to classify given data objects into K different clusters through the iterative, converging to a local minimum. So the results of created clusters are compact and independent.

The algorithm consists of two separate phases. In first phase, selects K-centers randomly, where the value K is fixed in advance. The next phase is to take each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data objects and the cluster centers when all the data objects are included in some clustes.The first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process continues repeatedly until the criterion function becomes the minimum. Supposing that the target object is X, Xi indicates the average of cluster Ci, criterion function is defined as follows.

$$E = \sum_{i=1}^{k} \sum_{x € c_i} \left| x - x_t \right|^2$$

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance which is used for determining the nearest distance between each data objects and cluster center. The Euclidean distance between one vector X=(X1, X2 ….Xn) and another vector Y= (Y1, Y2 ……Yn) .The Euclidean distance d (Xi, Yi) can be obtained as follow:

$$d(x_i, y_i) = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2}$$

The process of K-Means algorithm as follows:

INPUT: Number of desired clusters K and a database D= {d1, d2… dn} Containing n data objects.

OUTPUT: A set of K clusters.

STEPS:

1) Randomly select K data objects from dataset D as initial cluster centers.

2) Repeat;

3) Calculate the distance between each data object di (1<=i<=n) and all K cluster centers cj (1<=j<=k) and assign data object di to the nearest cluster.
4) For each cluster j (1<=j<=k), recalculate the cluster center.
5) Until no changing in the center of clusters.
The K-Means clustering algorithm always converges to local mininmum.Before the K-Means algorithm converges, calculation of distance and cluster centers are done whereas the loops are executed a additional range of times, where the positive integer t is known as the number of K-Means iterations. The precise value of t varies looking on the initial starting cluster centers. The distribution of data points encompasses a relationship with the new clustering center. So the computational time complexity of the K-Means algorithm is o (nkt). N is the number of all data objects, k is the range of clusters, t is the iterations of algorithm, typically requiring k<<n and t<<n.

6. CONCLUSION

This paper uses Hopfield network, LAMSTAR Network and K-Means algorithm to assist the doctors to perform differential diagnosis along with the possible implementation using SOA technique. By using these techniques, it improves the overall speed and increase the accuracy of algorithm. Especially in large datasets, LAMSTAR network gave faster and better result. It reduces the effects of misdiagnosis, especially practioners and students can also easily identify the diseases. It will also help the medical fraternity in the long run by helping them in getting accurate diagnosis and sharing of medical practices which will facilitate faster research and save many lives.

REFERENCES

1. Rahul Isola,Rebeck Carvalho and Amiya Kumar Tripathy.**Knowledge discovery in Medical system by using Differential Diagnosis, AMSTAR and K-NN,** *IEEE Transaction on Information Technology in Biomedicine*, Vol.16, No.6, November 2012.
2. **Web MD: Better Information, Better Health**.[online]. Available at http: // Symptoms. Webmd. com/ Symptomchecker, June 10, 2012.
3. H.R.Warner and O.Bouhaddou. **Innovation Review; Iliad A Medical Diagnostic Support Program**. Top health Inf.Manage., Vol.14, No.4, 1994.
4. **Department of Medicine Massachusetts Hospital, Boston, Dxplain System**. Available at http://dxplain.org/dxpdemopp/dxpdemobrief files/frame.htm, 2011.
5. E.Coiera. ***The Guide to Health Informatics.*** 2nd ed.London, U.K.: Arnold, October 2003.
6. Michael L.Raymer, Travis E.Doom, Leslie A. Kuhn and William F.Punch. **Knowledge Discovery in Medical and Biological Datasets Using a Hybrid Bayes Classifier/Evolutionary Algorithm.** *IEEE Transaction on Systems, Manx and Cybernetics*, Vol.33, Issue 5, October 2003.
7. J.Serrano, M.Tomeckova, J.Zvarova. **Machine Learning methods for Knowledge Discovery in Medical data on Atherosclerosis,** Department of Medical Informatics, Institute of Computer Science ASCR, Prague, Czech Republic.
8. B.M.Patil, Ramesh C.Joshi and Durga Toshniwal **Effective framework for Prediction of Diseases outcome using medical Datasets clustering and classification** *Published in International Journal of computational Intelligence* studies, Vol 1, Issue 3, pages 273-290, August 2010.
9. M.H.Valipour, B.A.Zafari, K.N.Maleki and N.Daneshpour. **A brief survey of Software Architecture concepts and Service Oriented Architecture,** in *Proc.2nd IEEE Int.Conf.Comput.Sci.Inf.Technol*, pp.34-38, August 8 2009.
10. Vikram Singh and Sapna Nagpal.**Guided Clustering technique for Knowledge Discovery-A case study of Liver disorder dataset**. Department of Computer science and Applications.
11. Daniel Graupe. ***The LAMSTAR Neural Network: A brief review***. Department of Electrical and Computer Engineering, University of Illinois, Chicago.
12. Hubert Kordylewski and Daniel Graupe. **A novel large memory neural network as an aid in medical diagnosis Applications,** *IEEE Trans.Inf.Technol.Biomed.*, Vol.5, No.3, pp 202-209, Sep 2001.

Ms.B.Sathyabama received the B.Tech (Information Technology) in V.S.B Engineering College, Anna University Chennai. Areas of Interest include Data warehousing and Data Mining and Networks.

Mr.S.Balaji Subramani has received the M.Tech (IT) in K.S.Rangasamy College of Technology. He is currently working as an assistant professor in V S B Engineering College. He has total number of 19 publications, 3 papers in International Journals, 7 papers in International conferences and 8 papers in national seminars/conferences and participated in various symposiums and workshops held at different places. His area of interest includes Word sense disambiguation, web mining, Information retrieval and social network analysis.