Automated Differential Diagnosis in Medical Systems using Neural Networks, *k*NN and SOM

Rahul Isola
Department of Mechanical Engineering
D.B.I.T.
Mumbai, India
rahulisola@gmail.com

Rebeck Carvalho
Department of Computer Engineering
D.B.I.T.
Mumbai, India
maaask3@gmail.com

Mangala Iyer
School of Pharmacy and Technology
N.M.I.M.S.
Mumbai, India
prarthana.iyer@gmail.com

Amiya Kumar Tripathy
Department of Computer Engineering
D.B.I.T.
Mumbai, India
aktripathy@dbit.in

Abstract—**The amount of Medical data recorded in hospitals and its significance as an ever-growing source of information has been long known and proven. Though the importance of the information hidden in these records has never been doubted, this data has mostly been used only for clinical purposes. Only recently has this been properly mined for valuable information to be used for research and to develop systems that assist the medical fraternity. Mostly, the systems that make use of this information are domain specific systems that predict diseases restricted to their area of specialization (like heart, brain etc.). But these systems are limited and are not applicable to the whole medical dataset. Our system uses this vast storage of information so that diagnosis based on this historical data can be made. This system aids medical diagnosis in the whole dataset by computing the probability of occurrence of a particular ailment from the medical data. The system mines the data using a unique algorithm which increases accuracy of such diagnosis by combining Neural Networks and Differential Diagnosis all integrated into one single approach. The strengths of *k*NN, Hopfield algorithm, SOM and P2P Grid Architecture are used to make the system unique and effectively enhanced.**

Keywords-Data mining; Neural Network; Decision support system; KNN classification; SOM

I. Introduction

Technology has played an unquestionable role in helping the medical fraternity in various ways, from surgical imagery to x-ray photography. But when it comes to diagnosis, the process still requires a doctor's knowledge. Medical diagnosis involves processing various variables, ranging from medical history, climatic conditions, blood pressure, environment and lots more. The number of variables counts upto the total variables that are required to understand the complete working of nature itself, which no model has successfully analyzed yet. To overcome this problem, Medical decision support systems [1] [2] [3] are becoming more and more essential, which will assist the doctors in taking correct decisions.

The system, making use of various techniques including KNN classification and Neural Networks, gives the possible diagnosis along with the set of most probable diseases which have similar symptoms. Conventional methods completely overlook various variables involved such as prevailing conditions, the build-ups resulting in the symptoms etc., due to sheer magnitude of available unknown variables.

To solve this problem, doctors generally perform the process of Differential Diagnosis [4] i.e. they shortlist the most probable disease based on their experience, and when the diagnosis goes wrong, diseases that show symptoms similar to the one originally diagnosed are considered. This requires a lot of research and a lot of prior medical experience.

The process of Differential Diagnosis has been emulated in this system, thus making this rather tough task a lot easier. This method is further modified and enhanced to reduce the huge number of underlying variables to just one by finding the root disease, or the most probable disease, using smart pattern matching involving *k*NN Classification technique [5] and the next probable diseases by performing Differential Diagnosis, using the Hopfield Neural Networks Theory [6] and SOM (Self Organizing Map)[7]. Using these, and by utilizing a database having a comprehensive list of medical history at the disposal of this system, the probability of occurrence of a disease may be calculated, regardless of the various unknown variables. The algorithm will output the disease from the symptoms entered and also gives the next highly probable disease and thus, the most effective course of action to be performed can be determined.

Making a medical decision is highly specialized and challenging due to various factors, especially in the case of rare diseases or diseases that show similar symptoms. These factors may vary from inexperience of the doctor, misdiagnosis due to environmental factors (stress, fatigue etc), misdiagnosis due to inconclusive test reports etc. Also, the latest findings and developments by the researchers are not quickly spread to all doctors, which could delay diagnosis and treatment of patients. The process of Differential Diagnosis involves doctors narrowing down the diseases to the root cause by using their knowledge and experience and confirming it by performing various tests. The number of tests to be performed to reach the conclusion in case of rare diseases or diseases with similar symptoms involved is huge. In such cases, it may not always be feasible to perform so many tests due to the time and money spent. Especially in developing countries, the problem of lack of trained and experienced doctors and test centers leads to intensification of this problem [8]. To tackle this, medical decision support systems were introduced that can be accessed by anyone, anywhere. Thus, the aim of this system is to provide a centralized Medical Decision Support System accessible by all doctors anywhere.

II. Similar Works

A simple search on any search engine gives every possible list of symptoms with even the required medication for any

978-0-7695-4593-6/11 $26.00 © 2011 IEEE
DOI 10.1109/DeSE.2011.20

disease along with various residential remedies etc, derived from anything ranging from regional superstition and beliefs to extracts from latest medical journals. Some sites give the feature to diagnose the disease based purely on the input of symptoms [9], [10]. This information, however authentic, may prove fatal if given in the hands of an untrained patient as it may not apply to him. Each patient is different, with different variables leading up to his present state (his medical history, background, local climate etc.).

Considering localized subset of medical datasets, algorithms have been formed [11]; and accurate results have been achieved by some of them [12]. But on a much larger generalized data set for every medical field, obtaining accurate results has yet been very difficult.

Iliad[13] and DXplain[14] are two software that come close to provide expert diagnosis. Iliad uses Bayesian reasoning to calculate the posterior probabilities of various diagnoses under consideration, given the findings present in a case. Similarly, DXplain acts on a set of clinical findings (signs, symptoms, laboratory data) to produce a ranked list of diagnoses which might explain (or be associated with) the clinical manifestations. DXplain takes advantage of a large data base of the crude probabilities than Iliad, but it still falls way short of actual number of possibilities actually possible. Also, new diseases are not considered at all, and the system leaves no scope to add them subsequently. The system of DXplain uses a modified form of Bayesian logic. The disadvantages of both these software are that both use only Bayesian Classification to perform differential diagnosis on a fixed data set. Hence, the software becomes redundant over time. They run on the assumption that the classes pre-generated in the systems is fixed, and will not change, but in real life, it is not so. The symptoms vary based on various conditions, including climatic conditions like season, average temperature, humidity and rainfall, to prevailing medical conditions that the patient is suffering from. These software have a static database which is not self-learning. Hence catching trends and adapting according to the conditions becomes impossible.

We have already proposed an algorithm that uses Hebbian Learning, Bayesian Classification and Back Propagation to diagnose diseases. This is further enhanced here by using *k*NN, Hopfield and SOM instead, to provide better results. The results of detailed comparisons are given in the Section VI.

III. FUNCTIONAL ASPECTS: RELEVANT DATA FETCHING (DATA MINING)

Existing medical systems focus on collecting and mining the entire data. The entire patient records are loaded and all factors are considered. The medical field cannot be easily analyzed because for generating a probabilistic rating, not only symptoms but factors like test results and external climate conditions are also required, which may or may not be present in the report. Existing system have failed to understand one of the most important attributes, Misdiagnosis, which interconnects and addresses all these issues. If we mine Misdiagnosis and store it as an attribute, it will solve the problem, because doctors' misdiagnosis only in case of presence of ambiguities, and in similar cases, there is a high chance of other doctors also performing the same misdiagnosis. The misdiagnosis attribute is very important in mining because it will directly lead to correct diagnosis, and in turn, it will eliminate all the underlying variables as they would already have been covered in the diagnosed ailment.

Retrieving relevant medical data for this system is a major task, with various issues rising about data confidentiality, data security, ambiguity etc. To tackle these issues, we came up with the idea of implementing a complete Hospital management system, with the decision support system as an integral part. This way, the issue of data confidentiality is addressed, as the system will be accessible only by the hospital representatives, but also have access to all the records generated by the doctors. Also, since the system will be hosted on the internet, the database will be accessed and updated by every user of the system from all the hospitals using the hospital management system, resulting in an updated and complete database. More users will result in a more complete and accurate database.

Once the medical records are obtained and are in place, the system uses NLP (natural language processing) [15] technique to determine the report results. MontyLingua [16], a NLP tool, was used in the system. Text mining using NLP obtains two results from the reports generated by the doctors, the correct and wrong diagnosis, and the symptoms resulting in the diagnosis. Other aspects like personal patient information, like name etc., or any information that can identify the patient in any way, are completely ignored, hence keeping the confidentiality intact. These results are stored in the database as: Disease Diagnosed & Actual Disease attributes. The symptoms will result in change in weightage of the symptoms for the corresponding disease (explained later in the algorithm section).

IV. FUNCTIONAL ASPECTS: ALGORITHM

Various algorithms utilized in this method are explained below. Fig. 1 shows a 3 tier workflow pattern of the system. The system uses a 3 tier workflow method for triple precision diagnosis. Each case is explained with the help of a practical case study given below.

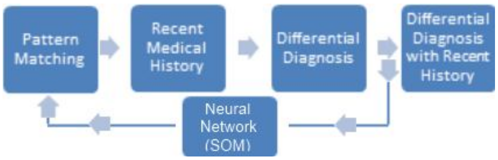


Fig. 1 Workflow Pattern

A. *Iterative pattern search*

Iterative pattern search utilizes data that is stored as given in table 1. 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. i.e. the new symptom list will contain symptoms of only those diseases that were obtained in the previous list. These related 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.

TABLE I
DATABASE FOR STEP 1: ITERATIVE PATTERN SEARCH SAMPLE DATASET

Diseases	Symptoms & Weight	Class Weight
Diabetes (D_1)	Headache (W_1), Increase in Blood Sugar (W_2), Insulin Low (W_3)	Endocrine (C_1)
Pericarditis (D_2)	Chest pain (W_4), Fever (W_5), Weakness (W_6), Malaria (W_7), Shortness of Breath (W_8), Syncope (W_9)	Cardiovascular (C_2)
Viral Fever (D_3)	Headache (W10), Cold (W_{11}), Fever (W12), Running Nose (W_{13}), Weakness (W_{14})	Parasitic (C_3)
Sinusitis (D_4)	Pain in the Sinuses (W15), Headache (W_{16}), Heavy Eyebrows (W_{17}), Blurry Vision (W_{18}), Fever (W_{19})	Respiratory (C_4)

After a few initial iterations, top diseases in the list gain highly in ranking, allowing one to identify the ailment. i.e. ranking of the disease varies at a much greater precision as more and more symptoms are given. e.g. on the database given in Table I, on entering Headache as a symptom, Diabetes, Viral fever and sinusitis will have 100% match, giving it equal ranking, whereas Pericarditis will be excluded. On entering the next symptom, e.g. Fever, Viral Fever will have both the symptoms matching, giving it 100% match, while the other three will have a match of only one out of the given two symptoms, thus 50% match, resulting in a drop in their ranking.

But this has to search a record database of more than 20000 diseases and even more symptoms, which is very time consuming, so we apply *k*NN Classification to classify diseases into subgroups and if a group of symptoms match we give higher preference to that subgroup, hence searching in that subgroup thus reducing database access.

In pattern recognition, the *k*-nearest neighbor algorithm (*k*NN) is a method for classifying objects based on closest training examples in the feature space. *k*NN is a type of instance-based learning, or lazy learning where the function is only approximated locally and all computation is deferred until classification. The *k*-nearest neighbor is classified by a majority vote of its neighbors, with the object being assigned to the class most common amongst its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of its nearest neighbor.

*k*NN has been modified to give faster processing as follows [17]. For the data given in Table I instead of using the Euclidean distance between the neighbors their weights will be considered. The logic of assigning weights is explained later. *(Ref. part D of this section).* Here weights are given to the individual symptoms corresponding to each disease, each individual disease and the subclass disease category. Common symptoms like Headache are assigned different weights in different diseases. This is done because if we take same weight for repeated symptoms in all disease it will lead to improper diagnosis. This happens because each symptom has a special role to the disease and the subclass it belongs to, for e.g., Headache may gain more weightage in Viral Fever or sinusitis as compared to the other diseases where Headache occurs, but is not that prevalent.

*k*NN will first sum the individual weights of each symptom, compare it first to the nearest subclass and then to all the diseases in that subclass resulting in faster accuracy. The choice of assigning K is given to the doctor, depending on how many comparison is desired, default is set to K=10.For the data in Table III, if Headache, Fever and Pain in the Sinuses is entered, then the weights W_{15}, W_{16} and W_{19} will be considered. Next all the weights will be added and compared to all sub classes C_1, C_2, C_3 and C_4 of which C_4 is most likely the answer depending on its weight. Lastly all the Diseases in class C_4 are considered and if Sinusitis (D_4) weight is closer to the sum of all the input symptoms weights then it is the possible diagnosis.

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. But if multiple diseases are found with nearest weights or same weights, then we proceed to point B.

B. *Mining medical records (Based on recent trends)*

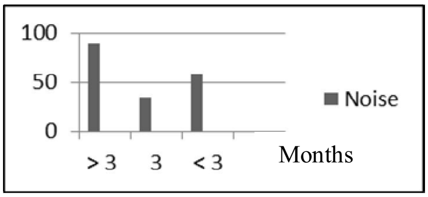


Fig. 2 Recent Medical Mining Trends

In the previous case, 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 is especially the case in case of some epidemic in the area, or some rare disease, or disease arising due to localized conditions etc. e.g. Swine flu epidemic initially showed the same symptoms as that of viral fever, resulting in rising the ranking of both viral fever and swine flu. In such cases, it becomes very difficult to point at one disease using the iterative pattern search method. In such cases, we use recent medical history with a time period of 3 months to rank the diseases on basis of the probability of their occurrence in the review period. 3 months provided accurate diagnosis with less noise (*Shown in Fig. 2*).

If the interpreted diagnosis is still vague, then we proceed to point C.

C. *Differential Diagnosis*

TABLE II
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

A Hopfield network is a form of recurrent artificial neural network invented by John Hopfield. Hopfield nets 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. Furthermore, it is through a Hopfield network that human memory can be further understood**.** The units in Hopfield nets are binary threshold units, i.e. the units only take on two different values for their states and the value is determined by whether or not the units' input exceeds their threshold. Hopfield nets can either have units that take on values of 1 or -1, or units that take on values of 1 or 0. So, the two possible definitions for unit *i*'s activation, a_i, are:

$$a_i \leftarrow \begin{cases} 1 & \text{if } \sum_j w_{ij}s_j > \theta_i, \\ -1 & \text{otherwise.} \end{cases}$$

$$a_i \leftarrow \begin{cases} 1 & \text{if } \sum_j w_{ij}s_j > \theta_i, \\ 0 & \text{otherwise.} \end{cases}$$

Where:

- w_{ij} is the strength of the connection weight from unit *j* to unit *i* (the weight of the connection).
- s_j is the state of unit *j*.
- θ_i is the threshold of unit *i*.

The connections in a Hopfield net typically have the following restrictions:

- $w_{ii} = 0, \forall i$ (no unit has a connection with itself)
- $w_{ij} = w_{ji}, \forall i, j$ (Connections are symmetric) ……IV

For the data given in Table III, if we apply Hopfield rule (equation IV) we find that Δw=Relative frequencies.
Therefore the individual weights are as follows:

$W(C_1)=5/10$, $W(C_2)=3/10$, $W(C_3)=2/10$

Where,

C_1= Diabetes, C_2 = Hypertension & C_3=Arthritis

So the Differential Diagnosis would be:

Diabetes => Hypertension =>Arthritis

D. *Weight assigning using SOM*

Using Step C, the correct disease shortlisted by the doctor is obtained who confirms it by taking the necessary tests. The final report is then mined using NLP processor Montylingua and the correct symptoms are compared with the original symptoms entered. This information is now fed to the SOM system for assigning weights. A SOM or self-organizing feature map (SOFM) is a type of artificial neural network that is trained using unsupervised learning to produce a low-dimensional (typically two-dimensional), discretized representation of the input space of the training samples, and called a map [18].

The training utilizes competitive learning. When a training example is fed to the network, its Euclidean distance to all weight vectors is computed. The neuron with weight vector most similar to the input is called the best matching unit (BMU). The weights of the BMU and neurons close to it in the SOM lattice are adjusted towards the input vector. The magnitude of the change decreases with time and with distance from the BMU. The update formula for a neuron with weight vector Wv(t) is

$$Wv(t + 1) = Wv(t) + \Theta\ (v, t)\ \alpha(t)(D(t) - Wv(t)), \ldots\ldots\ldots V$$

Where α(t) is a monotonically decreasing learning coefficient and D(t) is the input vector. The neighbourhood function Θ (v, t) depends on the lattice distance between the BMU and neuron v. During mapping, there will be one single winning neuron: the neuron whose weight vector lies closest to the input vector. This can be simply determined by calculating the Euclidean distance between input vector and weight vector. 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 unmatched 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.

The weights assigned here have been found out from the rigorous three tier process. It very important in pattern matching in the sense that they incorporate the misdiagnosis factor and the doctor may get 100 % result in the first step itself. SOM as compared to back propagation is faster in assigning of weight and gave more priority to recent weight due to its topology preservation and fast convergence capability.

V. IMPLEMENTATION

P2P systems have been shown to be scalable and robust. Hence, a large scale telemedicine solution calls for a P2P grid based architecture [19] integrating hospitals as centralized schemes cannot scale to such huge number of requests. The distributed nature of peer to peer overlay network increases robustness in case of failures whereas a central server becomes a single point of failure. The formation of a grid of hospital nodes with internet as the backbone will also enable cost-effective utilization of existing hardware resources. The lack of infrastructure particularly in the rural areas and the fact that more than 80 percent of the world population live in areas with mobile phone coverage, make the provision of mobility both at the patient and doctor side necessary. Support for mobility at the patient's end will improve the coverage of the telemedicine solution whereas mobility at the doctor's end would enable a doctor located anywhere to access the patient reports via internet and provide the required advice. The next important requirement in a large-scale distributed telemedicine solution is the context-aware scheduling of patient requests or in other words the dynamic discovery of relevant resources (doctors) to serve the requests. Considering the context parameters like location, patient history, severity of ailment etc. while scheduling would enable the requests to be served effectively.

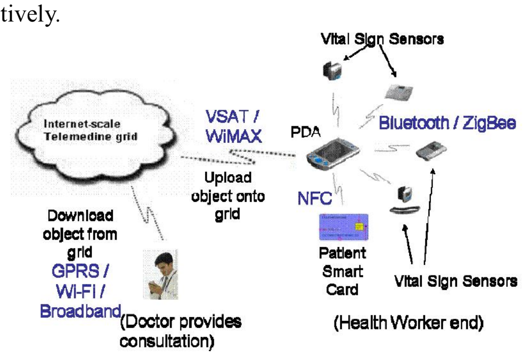
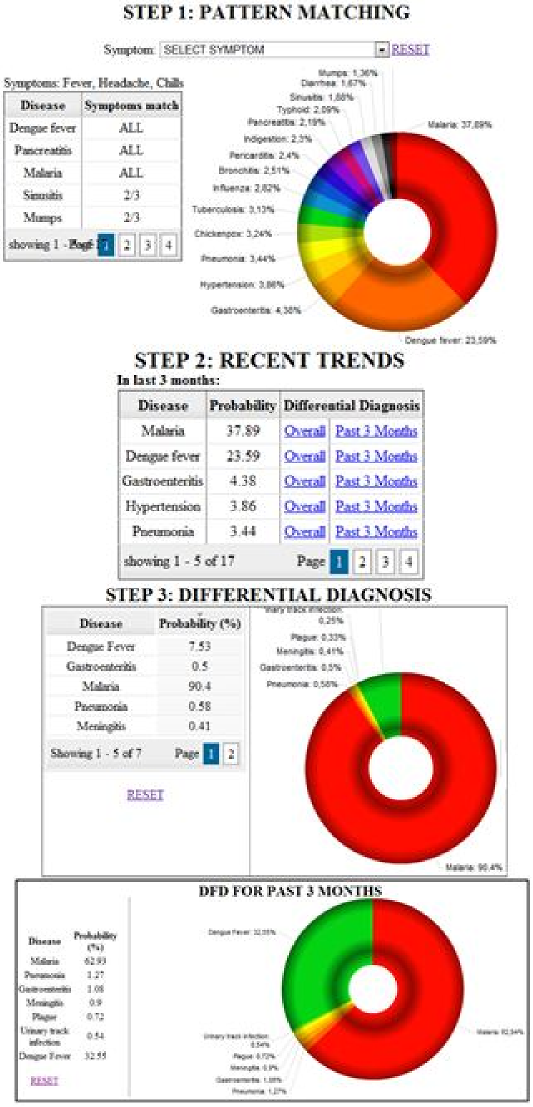


Fig. 3 Implementation of our system on P2P Grid Architecture

The health workers visit the rural areas equipped with vital parameter measuring devices (referred as sensors in figure 3) like Blood Pressure meter, Blood Sugar meter, ECG jacket etc. forming a ZigBee / Bluetooth network with a PDA. They measure the vital parameters and then upload them onto the data grid. The grid scheduler matches the request context to a nearest available doctor and notifies him via email / SMS. The doctor can then look at the request on his PDA and give consultation. The proposed system provides a persistent object space abstraction for the storage of medical data on the grid (formed by the nodes contributed by the participating hospitals). An object space provides an easy-to program abstraction for building applications while the distributed nature of storage (on the grid) makes it highly available as well as resilient to failures. The patient requests are scheduled to appropriate doctors by a distributed context-aware scheduler that considers context parameters like proximity, patient history, severity of ailment etc. The scheduler first tries to schedule the request by using its local resource information (about the available doctors). If no resource is free, it inserts the request into a location-wise tuple space thereby enabling nearby hospitals to pick it up. After a timeout the request is made

available globally by using a global tuple space. Thus, the solution emphasizes on timely fulfillment of patient requests as well as proximity of service. The proximity aware overlay structure and the scheduling mechanism minimize the data movement cost and improve the efficiency of the system. The solution uses the existing internet infrastructure and supports mobility at doctor and patient ends.

VI. RESULTS


STEP 1: PATTERN MATCHING

Symptom: SELECT SYMPTOM RESET

Symptoms: Fever, Headache, Chills

Disease	Symptoms match
Dengue fever	ALL
Pancreatitis	ALL
Malaria	ALL
Sinusitis	2/3
Mumps	2/3

showing 1 - Page 1 2 3 4

STEP 2: RECENT TRENDS

In last 3 months:

Disease	Probability	Differential Diagnosis	
Malaria	37.89	Overall	Past 3 Months
Dengue fever	23.59	Overall	Past 3 Months
Gastroenteritis	4.38	Overall	Past 3 Months
Hypertension	3.86	Overall	Past 3 Months
Pneumonia	3.44	Overall	Past 3 Months

showing 1 - 5 of 17 Page 1 2 3 4

STEP 3: DIFFERENTIAL DIAGNOSIS

Disease	Probability (%)
Dengue Fever	7.53
Gastroenteritis	0.5
Malaria	90.4
Pneumonia	0.58
Meningitis	0.41

Showing 1 - 5 of 7 Page 1 2

RESET

DFD FOR PAST 3 MONTHS

Disease	Probability (%)
Malaria	62.93
Pneumonia	1.27
Gastroenteritis	1.08
Meningitis	0.9
Plague	0.72
Urinary track infection	0.54
Dengue Fever	32.55

RESET

Fig. 4 Results and Comparisons

We have developed a sample system with limited database to test the above theory using PHP, MySQL and AJAX. The data has been obtained from the following sites medicine.net, wrongdiagnosis.com& webmd.com. We have applied our system on a sample dataset for malaria. Specific test cases were run, and the following results were obtained.

The 1st chart in fig. 4 depicts the number of symptoms matching and their probabilities considering the ranking adjusted according to the weights. The first step resulted in accurate prediction of diseases based on the symptoms entered. The table in fig.4 shows the list of diseases found by matching symptoms and its probabilities of occurring calculated based on its occurrence in the past three months. The second step resulted in very accurate prediction of diseases based on the recent trends. E.g. it accurately caught Malaria and Dengue for the given symptoms, during the monsoon season, where mosquitoes were a menace in the test locality. This matched very accurately with the patient's actual ailment. The last chart in fig. 4 shows the diseases similar to the root disease selected and probabilities of the disease occurring calculated based on the differential diagnosis technique. The data has been obtained from the following sites medicine.net, wrongdiagnosis.com & webmd.com. On comparison of our results with these sites, an average accuracy rate of 95 % was obtained (Fig. 7).

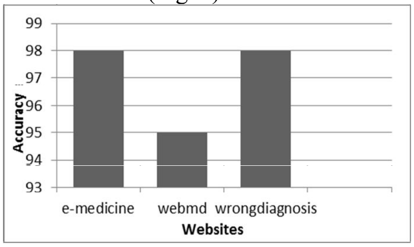


Fig. 7 Comparison of result with various sites

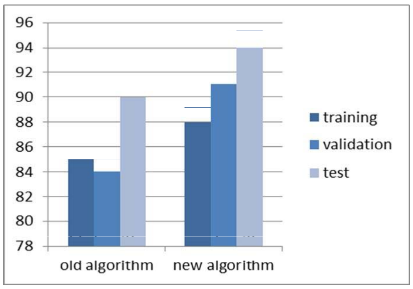


Fig. 8 Comparison of Old Algorithm [1] v/s New Algorithm

VII. Application

The system uses a centralized database. This ensures availability of abundant data for use by multiple franchisees who subscribe to the system. By using our system, many essential results can be obtained, reducing the need of human intervention, thus preventing misdiagnosis to a great extent. The system can be used in solving a few common problems, which includes diagnosis of multiple diseases showing similar symptoms, diagnosis of a person fighting multiple diseases, faster second opinion, identifying trends in medical records faster.

VIII. Conclusion

With the support of various medicinal practitioners and hospitals, higher probability of getting the diagnosis right can be obtained. The system cannot be used as a substitute or a shortcut to diagnosis. But it can definitely complement the doctors' knowledge and assist them to reach a conclusion. The doctor always has the upper hand to decide whether to use the diagnosis given by the algorithm or not. After sufficient self-learning, with an extensive database of medical records to mine from, this can be used to build formidable medical assistance software that can be of great use to all doctors, and specially the new practitioners and students. It will 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] Rebeck Carvalho, Rahul Isola, Amiya Tripathy, "Automated Medical Decision Support System", CBMS 2011 Symposium., 24th IEEE International Conference on Computer Based Medical Systems, June 27th - 30th, University of the West of England, Bristol, UK.

[2] Kensaku Kawamoto, Caitlin A Houlihan, E Andrew Balas and David F Lobach, "Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success", BMJ 330 : 765 doi: 10.1136/bmj.38398.500764.8F (Published 14 March 2005)

[3] Randolph A Miller, "Medical Diagnostic Decision Support Systems—Past, Present, And Future - A Threaded Bibliography and Brief Commentary", JAMIA 1994;1:8-27 doi:10.1136/jamia.1994.95236141

[4] Walter Siegenthaler, Differential diagnosis in internal medicine: from symptom to diagnosis, 2011 Edition, APPL, aprinta druck, Wemding, Germany.

[5] Jiawei Han, Micheline Kamber, Data Mining Concepts and Techniques, 2011 edition, Morgan Kaufmann Publications.

[6] Polyn, S.M., & Kahana, M.J. (2008). Memory search and the neural representation of context. Trends in Cognitive Sciences, 12, 24-30.

[7] Kohonen, T. and Honkela, T. (2007). "Kohonen network". Scholarpedia.

[8] Susan F. Murray, Stephen C. Pearson, "Maternity referral systems in developing countries: Current knowledge and future research needs", Social Science & Medicine 62 (2006) 2205–2215

[9] Misdiagnosis: Symptom and heath diagnosis checker.Available at: http://www.misdiagnosis.com. (4th Feb 2011,7.30pm GMT)

[10] WebMD: Better Information, Better Health. Available at http://symptoms.webmd.com/symptomchecker. (4th Feb 2011,7.30pm GMT)

[11] Lishuang Li ; Linmei Jing ; Degen Huang ; "Protein-protein interaction extraction from biomedical literatures based on modified SVM-*k*NN", NLP-KE 2009

[12] Michele Berlingerio, Francesco Bonchi Fosca Giannotti, Franco Turini, "Mining Clinical Data with a Temporal Dimension: a Case Study", 2007 IEEE International Conference on Bioinformatics and Biomedicine

[13] Warner HR, Bouhaddou O (1994). "Innovation review: Iliad--a medical diagnostic support program.". Top Health Inf Manage 14 (4): 51–8. PMID 10134761.

[14] Department of Medicine Massachusetts Hospital, Boston, DXplain System (2011). Available at: http://dxplain.org/dxpdemopp/dxpdemo-brief_files/frame.htm

[15] Yucong Duan, Christophe Cruz (2011), "Formalizing Semantic of Natural Language through Conceptualization from Existence",. International Journal of Innovation, Management and Technology(2011) 2 (1), pp. 37-42

[16] Maurice HT Ling, "An Anthological Review of Research Utilizing MontyLingua, a Python-Based End-to-End Text Processor.", The Python Papers, Vol 1, No 1 (2006)

[17] Renqiang Min ; Stanley, D.A. ; Zineng Yuan ; Bonner, A. ; Zhaolei Zhang ; " A Deep Non-linear Feature Mapping for Large-Margin kNN Classification", Data Mining, 2009. ICDM '09. Ninth IEEE International Conference.

[18] Chien-Pen Chuang ; Shiunn-Shin Lee ; Jia-Shiunn Tsai ; Tai-Jung Kuo ; "Detecting mammography of breast microcalcification with SOL-based self-organization neural network", Natural Computation (ICNC), 2010 Sixth International Conference.

[19] C. O. Rolim et al., "Towards a Grid of Sensors for Telemedicine," Proceedings of the 19th IEEE Symposium on Computer-Based Medical Systems (2006): 485-490.