Medical Disease Diagnosis Using Structuring Text

Gomathi.P
(Student)
Computer science and engineering
K.S.R college of Engineering
Tiruchengode.
India.
gomathyp7@gmail.com

Mrs.N.S.Nithya
(Asst.prof)
Computer science and engineering
K.S.R college of Engineering
Tiruchengode.
India.
sachinnithya@yahoo.com

Abstract--- Medical diagnosis is an important domain of research which aids to identify the occurrence of a disease. The paper proposes a novel glide path to knowledge discovery in medical systems by acquiring relevant information from the data set. This approach makes diagnosis easier. Using naive bayes, the overall speed and accuracy of the algorithm increased and extract high quality data set from an unstructured text. The primary advantage of the scheme is that it can be used to whatever sort of dataset whether it is a predefined dataset or not.

Keywords-Medical diagnosis, knowledge discovery, naïve bayes.

I. INTRODUCTION

Medical data mining focuses on various data mining techniques used particularly in medical application. Various Malaysian medical data collected and stored in Medical Data Repository. These data are used for various techniques and tasks. Among techniques used are statistical techniques, Neural Network, Rough Set Theory and Hybrid techniques. Medical data repository is a complete collection of medical data stored systematically and accessible in various formats. The data repository serves as a platform for researchers to develop new data mining techniques. The collection of data mining techniques will produce an intelligent data miner to support medical users such as medical institutions, hospitals, research centers, medical specialist and officers, medical policy makers and government. Medical data mining is one of key issues to get useful clinical knowledge from medical databases. These algorithms either rely on medical knowledge or general data mining techniques. Further, it is often the case that finding the correct subset of predictive features is an important problem in its own right. For example, physician may make a decision based on the selected features whether a dangerous surgery is necessary for treatment or not.

II. Related work

There are many ongoing researches in the field of medical diagnosis. Bayesian networks (BN) plays an important role in medical diagnostics. It is used to represent conditional dependencies among the random variables. It computes the probability of the occurrence of various diseases when the symptoms are given . The KNN is the simplest of all classifiers and is used in predicting diseases. ANN also called Neural Network consists of a neurons that are interconnected. They represent relationships between inputs and outputs. Patient is assigned to one of the classes of diseases with this network. ANN is good in identifying diseases and does not need any details of how to recognize as it learns by example . It is easy to maintain and has good capacity. Has good computational power with good accuracy. Back propagation is used to train artificial neural networks. It is a supervised learning method.

III. Proposed methodology

In the proposed work user will search for the disease diagnosis (disease and treatment related information) by giving symptoms as a query in the search engine. These symptoms are preprocessed to make the further process easier to find the symptoms keyword which helps to identify the disease quickly. The symptoms which keyword is matched with the stored medical input database to identify the multiple diseases related to that keyword. Multiple diseases is identified, it will make the pattern matching about the multiple diseases and also find the probability of diseases. Then the disease will make a differential diagnosis to find the disease accuracy.

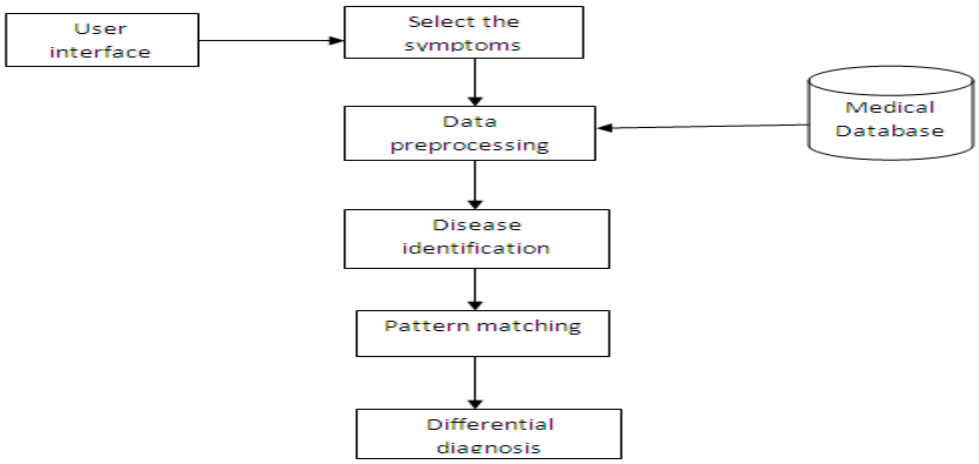


Fig 1. System architecture

The first step is user interface and select the symptoms. Next process database will preprocessing the dataset. The description dataset to improve the quality of data. So for that tokenization is done after that filtering is done to remove the nouns, propositions and so on. The ranking is generated according to the percentage match of the total number of symptoms entered. 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. This is especially the case of some diseases in an area, or some form of rare disease, or disease occurring due to various other factors. In such cases, it becomes very difficult to point at one disease using the symptom matching method. In such cases, recent medical historical data stored in the database of the proposed system is used.

IV. IMPLEMENTATION

Commonly used as a preliminary data mining practice, data preprocessing transforms the data into a format that will be more easily and effectively processed for the purpose of the user. Searching a database of diseases and symptoms is time consuming there by requires large database access which decreases the accuracy of the system. The relevant information are retrieved from description database using tokenization, filtering and stemming. 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. 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. This feature has been identified as the most suitable for the present system.

V. EXPERIMENTAL RESULTS

In this system proposed using javascript and SQL was developed. The data have been obtained from limited scope. The system was run for a dataset of 4000 diseases. The dataset contains diseases along with their symptoms. Specific test cases were run, and the following results were obtained. The dataset used here is a description dataset. The quality and representation of data is important. So to reduce the dataset tokenization, filtering, and stemming is done to the dataset. It is done to the entire dataset.

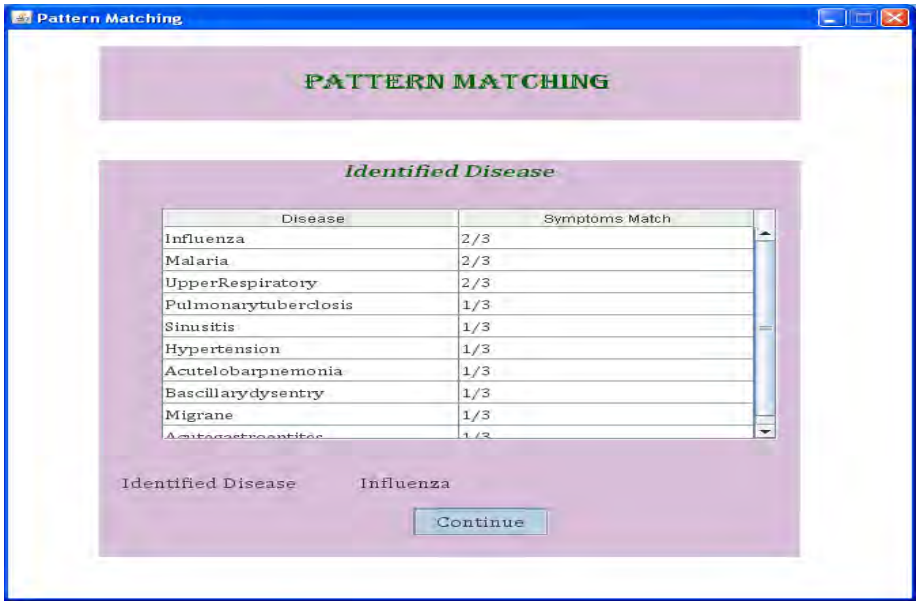
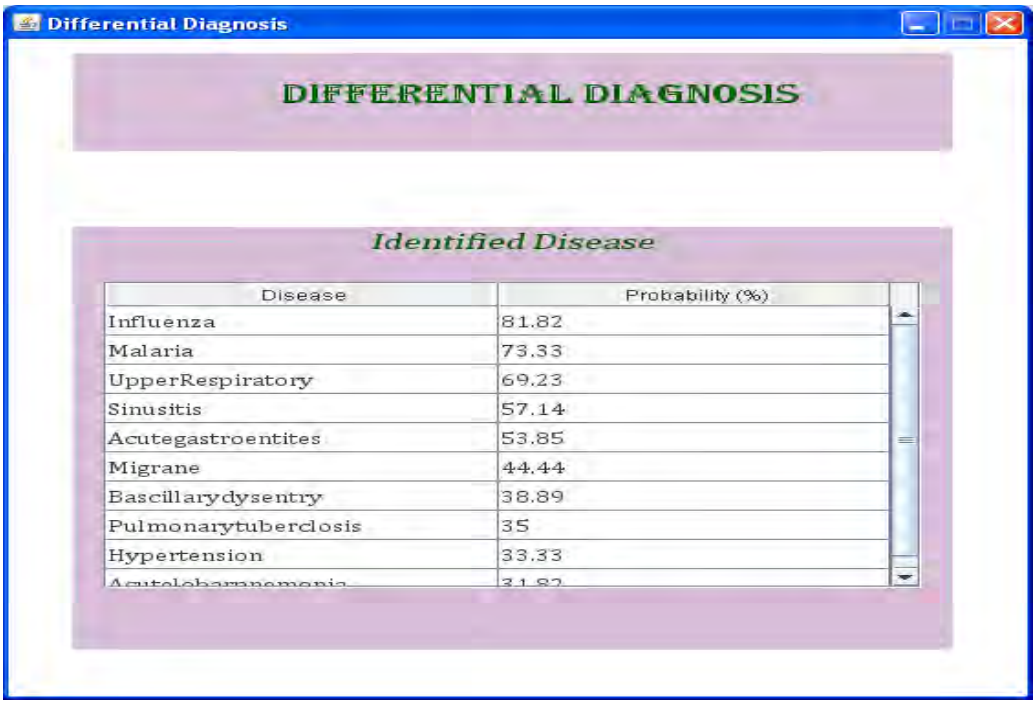


Fig.2 Pattern Matching

Fig. 2 shows the list of diseases found by matching symptoms and its probabilities of occurrence is calculated. This diagnosis matched very accurately with the patient's actual ailment.



.

Fig.3 Differential Diagnosis

Fig. 3 shows the list of diseases and their probabilities of occurring calculated on the basis of differential diagnosis technique. The graph shows a graphical representation of the same.In this system, an additional step, where differential diagnosis has been combined with recent medical history, to get more accurate results, has been implemented. This step has given accurate results to catch recent trends.

VI. CONCLUSION

Medical diagnosis is an important area of research which helps to identify the occurrence of a disease. Medical data is an ever growing source of information. The system, making use of various techniques mentioned, will in turn display the root disease along with the set of most probable diseases which have similar symptoms. The database used is a description database so to reduce the dataset tokenization, filtering and stemming is done. In this system, by using differential diagnosis, LAMSTAR, and Naive bayes, an attempt has been made to assist the doctors to perform diagnosis in accurate way. The main advantage of the system is that it can be applied to any kind of dataset whether it is a description dataset or not.

REFERENCES

[1] Rahul Isola , Rebeck Carvalho, and Amiya Kumar Tripathy, 2012, ' Knowledge Discovery in Medical Systems Using Differential Diagnosis, LAMSTAR, and KNN', IEEE Transactions on Information Technology in Biomedicine, Vol. 16, No. 6.

[2] Coiera .E, 2003,' The Guide to Health Informatics', 2nd ed. London , U.K.: Arnold, pp. 101123.

[3] Garima Singh, Vijay Kumar2,2013,' An Efficient Clustering and Distance Based Approach for Outlier Detection' ,International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 7.

[4] Warner .H.R and Bouhaddou .O, 2005,' Innovation review: IliadA medical diagnostic support program, Top Health Inf. Manage', vol. 14, no. 4,pp. 5158.

[5] Hubert Kordylewski and Daniel Graupe, 2001,'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.

[6] Han.J and Kamber.M, 2011,' Data Mining Concepts and Techniques'.

[7] Escudero.J, Zajicek .J.P and Ifeachor .E, 2011, 'Early Detection and Characterization of Alzheimer's Disease in Clinical Scenarios Using Bioprofile Concepts and K-Means', 33rd Annual International Conference of the IEEE EMBS Boston, Massachusetts USA.

[8] Li .L, Jing .L, and Huang.D, 2007,' Protein-protein interaction extraction from biomedical literatures based on modified SVM-KNN', in Nat. Lang. Process.Knowl. Engineer, pp. 17.

[9] Celebi M.E, Aslandogan Y.A and Bergstresser R.P, 2005,'Mining Biomedical Images with Density-based Clustering', Proceedings of the International Conference on Information Technology: Coding and Computing.

[10] Mai Shouman, Tim Turner, and Rob Stocker, 2012,' Applying k-Nearest Neighbourin Diagnosing Heart Disease Patients ',International Journal of Information and Education Technology, Vol. 2, No. 3.

[11] Berlingerio .M, Giannotti F.B.F and Turini .F, 2007,' Mining clinical data with a temporal dimension: A case study', in Proc. IEEE Int. Conf. Bioinf.Biomed., pp. 429436.

[12] Qeethara Kadhim Al-Shayea, 2011,' Artificial Neural Networks in Medical Diagnosis', International Journal of Computer Science Issues, Vol. 8, Issue 2.

[13] Belciug . S, 2009, ' Patients length of stay grouping using the hierarchical clustering algorithm', Annals of University of Craiova, Math. Comp. Sci. Ser., ISSN: 1223-6934, vol. 36, no. 2, pp. 79-84.

[14] Belciug .S, Gorunescu F, Salem .Aand Gorunescu .M, 2010,'Clustering-based approach for detecting breast cancer recurrence', 10th International Conference on Intelligent Systems Design and Applications, vol.45,pp.5342.

[15] Balasubramanian. T and Umarani .R, 2012, 'An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique', Proceedings of the International Conference on Pattern Recognition, Informatics and Medical Engineering, pp. 21-23.

[16] Siegenthaler .W, 2011,' Differential Diagnosis in Internal Medicine: From Symptom to Diagnosis'. New York: Thieme Medical Publishers.

[17] Liu.Z, T. Sokka, Maas. K, Olsen .N.J and Aune T.M, 2009, 'Prediction of Disease Severity in Patients with Early Rheumatoid Arthritis by Gene Expression Profiling', Human Genomics and Proteomics.