ISSN:0975-9646

Analytical Approach to Predictive Disease Diagnosis using K-nn and Kstar

Rashmi Madhukar Jadhav[#1] Ms. Roshani Ade[#2]

[#1]*Department of Computer Engineering,*
Dr. D. Y Patil School of Engineering and Technology, Lohegaon, SPPU,
Pune, Maharashtra, India

[#2]*AssistantProfessor, Department of Computer Engineering,*
Dr. D. Y Patil School of Engineering and Technology, Lohegaon, SPPU,
Pune, Maharashtra, India

Abstract— **Finding the exact cause of disease in a patient requires Doctors expertise. Doctors use their knowledge to cure a particular ailment based on the symptoms shown by the patient. However, this information may sometimes not be to the threshold level so as to result in possible disease suffered by the patient. This paper focuses on differential diagnosis using K-nn and Kstar algorithms to build a predictive system for symptoms based search for estimating the probable disease. Performance accuracy of the two algorithms is measured to give the best fit for the prediction system.**

Keywords— **K-nn, Kstar, Predictive system, Disease Diagnosis.**

I. INTRODUCTION

Medical data is an enormous quantity of data that can be used for a variety of purposes. Data mining is the field that is proving applicability in medical research also. There are large numbers of datasets available in many hospitals. This data can be used for effective disease diagnosis. Diseases showing similar symptoms are hard to predict. Factors leading to misdiagnosis may differ with inexperience of doctors. Doctors may also be in a restless state of mind which may affect their diagnosis. In some cases doctors may diagnose a particular symptom depending on their habitual and repeated diagnosis and also factors due to misinterpretations, ambiguities in symptoms, and inadequate information. Research scholars have been striving hard to diagnose this mislead factor so as to assist doctors in treatment of diseases.

Classification of diseases is based on differential diagnosis method. This is done by the doctors by chopping down the differential process in steps from root cause of the disease to its treatment. A list of similar symptoms is searched through to get the exact matching disease to the input symptom. If only one symptom is input then the algorithm returns in lesser time giving output as per the input of symptom. But if the symptoms from the patient are large in number then the complexity of the algorithm increases. Experienced doctors use classifiers to reach to the root level of the disease. This is accomplished by knowledge of doctors and their previous experience in curing the disease.

However this needs skill of doctors. The problem gets severe if the doctor is new and has inadequate training. This scenario is prevalent in developing countries. We propose an inventive methodology in assisting doctors and making treatment results better, thus making tougher task easier to a large extent. Smart pattern matching technique is used by including k-NN classifier and the next probable diseases by performing differential diagnosis.

The rest of this paper is organized as follows. Section 2 describes the review of literature. Section 3 discusses the K-nn and the K-star used for classification. Section 4 discusses on the existing system architecture for diagnosis. Section 5 focuses on the proposed methodology for the disease diagnosis. The datasets are then described in section 6. Experimental results are analyzed in Section 7 and Conclusions are given in Section 8. In the second half of the paper, the simulation study and its results are presented before the paper ends with a discussion and conclusion.

II. RELATED WORK

The authors believe that the medical decision making covers important tasks such as diagnosis, therapy planning, interacting with patients, identifying medical errors etc. Medical diagnosis is a process aiming at identifying diseases based on findings, such as symptoms and lab reports. The development of Medical (Clinical) Diagnosis Decision Support Systems (MDDS or CDSS) dates back to 1950s. Such developments, particularly in diagnosis decision support, have high complexity. A limited number of systems are adopted for practical use in the clinical environment. In the diagnosis process, an appropriate representation scheme is necessary for both problem interpretation and knowledge retrieval [1].

The authors feel that a good medical diagnosis system requires a structured knowledge representation component (model) that reflects most of the existing medical relations. It also needs employing efficient reasoning methods that closely follow medical cognition. Organizing these relations obviously needs one or more specific forms of knowledge representation, and computational artifacts that can manipulate them. None of these tasks is easy. Furthermore, physician-like users usually lack knowledge

of how the framework works. In the worst case, a common bottleneck in knowledge-based systems is the knowledge acquisition, because of the acquisition mechanism is not transparent to the experts, or tacit knowledge is hardly complete enough to be usable [2]. In this paper the authors explore the case-based reasoning (CBR) methodology, as CBR not only utilizes the actual data (cases), but also works similarly to how human solves problems by recalling most relevant experiences and propose a framework for medical diagnosis decision making. This framework incorporates the disease-symptom ontology, and case-based reasoning (CBR) coupling with semantic similarity calculation [3].

In this paper the author has focused on the enormous quantities of medical data are utilized only for clinical and short term use. MediQuery uses this vast storage of information so that diagnosis based on this historical data can be made. There are systems to predict diseases of the heart, brain and lungs based on past data collected from the patients. The focus is computing the probability of occurrence of a particular ailment from the medical data by mining it using a unique algorithm which increases accuracy of such diagnosis by combining Neural Networks, Bayesian Classification and Differential Diagnosis all integrated into one single approach. The system uses a Service Oriented Architecture (SOA) wherein the sys-tem components of diagnosis, information portal and other miscellaneous services provided are coupled [4].

The author has focused on the availability of huge amounts of medical data leads to the need for powerful data analysis tools to extract useful knowledge. Researchers have long been concerned with applying statistical and data mining tools to improve data analysis on large data sets. Disease diagnosis is one of the applications where data mining tools are proving successful results. Heart disease is the leading cause of death all over the world in the past ten years. Several researchers are using statistical and data mining tools to help health care professionals in the diagnosis of heart disease. Using single data mining technique in the diagnosis of heart disease has been comprehensively investigated showing acceptable levels of accuracy. Recently, researchers have been investigating the effect of hybridizing more than one technique showing enhanced results in the diagnosis of heart disease.

But, using data mining techniques to identify a suitable treatment for heart disease patients has received less attention. This paper identifies gaps in the research on heart disease diagnosis and treatment and proposes a model to systematically close those gaps to discover if applying data mining techniques to heart disease treatment data can provide as reliable performance as that achieved in diagnosing heart disease. Motivated by the world-wide increasing mortality of heart disease patients each year and the availability of huge amount of patient's data from which to extract useful knowledge, researchers have been using data mining techniques to help health care professionals in the diagnosis of heart disease. Developing a tool to be embedded in the hospitals management system to help and give advice to the healthcare professionals in diagnosing and providing suitable treatment for heart disease patients is important. Several data mining techniques are used in the diagnosis of heart disease such as Nave Bayes, Decision Tree, neural network, kernel density, automatically defined groups, bagging algorithm, and support vector machine showing different levels of accuracies.

The authors have proposed that applying data mining techniques in identifying suitable treatments for heart disease patients is fruitful and needs further investigation. To evaluate if applying data mining techniques to heart disease treatment can provide as reliable performance as achieved in heart disease diagnosis [5].

Data mining is the exploration of large datasets to extract hidden and previously unknown patterns, relationships and knowledge that are difficult to detect with traditional statistical methods [6-10]. Data mining is rapidly growing successful in a wide range of applications such as analysis of organic compounds, financial forecasting, healthcare and weather forecasting [11]. Data mining in healthcare is an emerging field of high importance for providing prognosis and a deeper understanding of medical data [12]. Data mining applications in healthcare include analysis of health care centers for better health policy-making and prevention of hospital errors, early detection, prevention of diseases and preventable hospital deaths, more value for money and cost savings, and detection of fraudulent insurance claims [13].

III. ALGORITHMS

- K-NN

The KNN is the simplest of all classifiers and is used in predicting diseases. For classification majority vote is considered. Value is assigned to class with highest match. As number of classes increases the performance of KNN increases .The number of neighbors is obtained with value of k [14]. Implementation of KNN mechanism is easy and the debugging process is very faster. As value of k decreases noise points in training set increases. As value of k increases it becomes expensive [15].

- K-STAR

K Star: Instance based classifier that uses similarity function from the training set to classify test set. Missing values are averaged by column entropy curves and global blending parameter is set [16].

IV. EXISTING SYSTEM

Iliad is an expert system program using Bayesian reasoning to calculate the posterior probabilities of various diagnoses when input is given as the findings. Also, DXplain is a decision support system which acts on a set of clinical findings to produce a ranked list of diagnoses which may give the clinical indicators. DXplain also provides justification for why each of these diseases might be considered, suggests what further clinical information would be useful to collect for each disease. DXplain formulates this list using a large database.

V. PROPOSED ARCHITECTURE

Figure 1 shows the proposed system design.

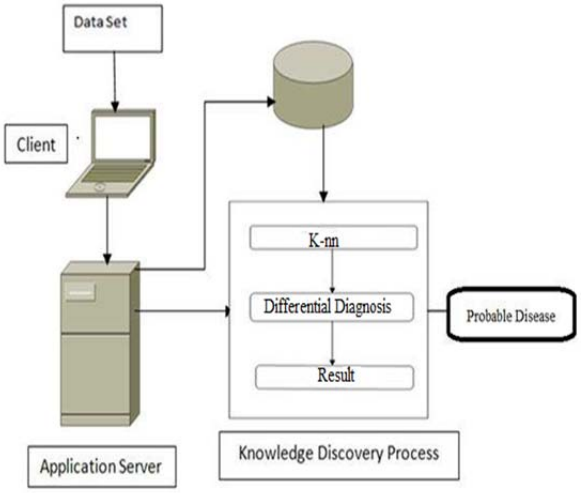

Figure 1: System Design

The flow of the system is as shown in the figure. The input symptoms are fed to the software which then undergoes knowledge discovery process and gives the probable list of diseases.

VI. Datasets Description

In order to compare the data mining classification techniques, computer files can be collected from the system hard disk and a data set is used from uci repository data mining tool is used for analyzing the performance of the classification algorithms. The dataset represents 10 years (1999-2008) of clinical care at 130 US hospitals and integrated delivery networks. It includes over 50 features representing patient and hospital outcomes. Information was extracted from the database for encounters that satisfied the following criteria:

(1) It is an inpatient encounter (a hospital admission).
(2) It is a diabetic encounter one during which any kind of diabetes was entered to the system as a diagnosis.
(3) The length of stay was at least 1 day and at most 14 days.
(4) Laboratory tests were performed during the encounter.
(5) Medications were administered during the encounter.

The data contains such attributes as patient number,

race, gender, age, admission type, time in hospital, medical specialty of admitting physician, number of lab test performed, HbA1c test result, diagnosis, number of medication, diabetic medications, number of outpatient, inpatient, and emergency visits in the year before the hospitalization, etc.

VII. Results

Figure 2 shows the graph for training datasets and time required.

Figure 2: Training datasets versus Time

As the number of instances increases the time needed also increases .The performance of algorithms is affected by the training data sets.

VIII. Conclusion

This system proposed has been aimed to provide essential medical services with clinical precision which needs high accuracy. Even though the system is to be used by doctors only, and the doctors have the final decision to make, the accuracy of the system is promising and will help the practitioners in their verdict. To verify this, the results obtained by this system were com-pared with the differential diagnosis provided by various other medical systems, including the information that is available at various online medical portals, and these were also verified by a panel of experts, consisting five reputed doctors at local level. The results obtained matched up to the doctors expectations, and since the system is self-learning, with time, as the database grows, the accuracy of the system improves.

Acknowledgment

We would like to thank the researchers as well as publishers for making their resources available for their guidance. We are also thankful to the reviewers for their valuable suggestions and also thank the college authorities for providing the required infrastructure and support.

References

[1] Margaret H. Dunham, ”Data Mining-Introductory and advanced topics”,Pearson Educa-tion,2013.
[2] Ian H. Witten,Eibe Frank,Mark A. Hall,”Data Mining”,Third Edition,Elsevier,2012.
[3] Hsien Tseng Wang,tansel,”Composite Ontology-based Medical Diagnosis Decision Sup-port System Framework”,Communications of the IIMA,Volume 13 Issue 2,2013.
[4] R. Carvalho, R. Isola, and A. Tripathy,”MediQueryAn automated decision support sys-tem”,in Proc. 24th Int. Symp. Comput.-Based Med. Syst.,Jun. 2730, 2011, pp. 16.
[5] Shouman, Mai, Tim Turner and Rob Stocker. ”Using data mining techniques in heart dis-ease diagnosis and treatment” Electronics, Communications and Computers (JEC-ECC), 2012 Japan-Egypt Conference on IEEE, 2012.
[6] J. Han and M. Kamber, ”Data Mining Concepts and Techniques.SanMateo, CA: Morgan Kaufmann, 2011.

[7] Lee, I.-N., S.-C. Liao, and M. Embrechts,"Data mining techniques applied to medical information",Med. inform, 2000.
[8] Obenshain, M.K.,"Application of Data Mining Techniques to Healthcare Data",Infection Control and Hospital Epidemiology,2004.
[9] Disease Diagnosis By Using Data Mining Algorithms
[10] Sandhya,J.,et.al,"Classification of Neurodegenerative disorders based on major risk fac-tors employong machine learning techniques",International Journal of Engineering and Technology.Vol.2,no.4,2010.
[11] Thuraisingham,B.,"A primer for understanding and applying data mining",IT profes-sional,IEEE,2000.
[12] Ashby, D. and A. Smith, The Best Medicine? Plus Magazine - Living Mathematics., 2005.
[13] Liao, S.-C. and I.-N. Lee, Appropriate medical data categorization for data mining classi-fication techniques. MED. INFORM., 2002, Vol. 27, no. 1, 5967.
[14] Ruben, D.C.J., Data Mining in Healthcare: Current Applications and Issues. 2009.
[15] Porter, T. and B. Green, Identifying Diabetic Patients: A Data Mining Approach. Ameri-cas Conference on Information Systems, 2009.
[16] Panzarasa, S., et al., Data mining techniques for analyzing stroke care processes. Proceed-ings of the 13th World Congress on Medical Informatics, 2010.