Iterative Pattern Matching using K- nn and Lazy Bayesian Rule

Rashmi Madhukar Jadhav
Student
Department of Computer Engineering
Dr. D. Y. Patil School of Engineering and Technology,
Lohegaon, SPPU, Pune
Maharashtra, India

Roshani Ade
Assistant Professor
Department of Computer Engineering
Dr. D. Y. Patil School of Engineering and Technology,
Lohegaon, SPPU, Pune
Maharashtra, India

ABSTRACT

Medical Data is an abundant source of information for research. Enormous quantity of data is available but is put to use in merely diagnostic procedure. Researchers are now spotlighting their interest in the medical field. This paper focuses on the misdiagnosis attribute that is often ignored but may lead to exact cause of disease and thus result in fine diagnosis in the patient.

Keywords

Misdiagnosis, LBR, K-nn

1. INTRODUCTION

Although the purpose of a medical education many years ago might have been to provide a physician with all the information needed for a lifetime of informed medical decision making, that time is long gone. Lack of sufficient information when making medical decisions can negatively affect patient outcomes. Having so much information that it becomes confusing and overwhelming can have the same effect [1]. In response to information overload, unanswered questions, and medical errors and adverse events, computer-based clinical decision support systems have been developed.

Technology has made a rapid development in the field of diagnosis in medical sciences. This has also helped doctors in different ways from surgery images to X-ray photography. But technology is still lagging behind for diagnosis of variables that constitute different factors [2]. These factors involve medical history to geographical conditions. No medical model has been successful yet to analyze such variables. Medical decision support systems are coming into picture to help doctors assist in decision making. Since primeval times, man has attempted to explain natural phenomena using models.

Classification of diseases is based on differential diagnosis method. This is done by the doctors by slandering down the differential process in steps that range from root cause of the disease to its treatment. A list of similar symptoms is traversed through to get the exact matching disease to the input symptom [3]. 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 ground 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 to some extent. The problem gets intensified if the doctor is new and lacks training. The scenario comes into focus in developing countries. The algorithm outputs the disease from the symptoms entered and gives the next probability of disease so that the necessary treatment can be decided upon.

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

2. RELATED WORK

Hatice Cataloluk et al. have developed a software tool to obtain correct diagnosis of patients. The tool uses two variants of K-nn algorithm namely basic K-nn and Weighted K-nn. Majority voting is employed to decide classification whereas Weighted K-nn is the method based on the principle that the closer neighbors to the new record have more weighted effects than remote neighbors. The tool is mainly focused on erythematosquamous disease. The data sets Dermatology Data sets from the UCI repository that consists of 366 records and six classes of erythematosquamous diseases. The working principle calculates the distance between test data and set of classes. The paper deals with the comparison between the two methods in terms of their performance. The authors derive a conclusion that weighted K-nn gives better results [4].

Nikita Chavan et al have laid stress on detection and classification of brain tumors. Tumor can destroy all healthy brain cells. Detection and classification of tumor is in benign stage is attempted in this paper. Proposed work consists of two stages as feature extraction and classification. The tool used to extract the MRI images is GLCM (Gray Level Co-occurrence Matrix).The images are then classified using the K-nn classifier. The classification results in normal and abnormal images.

Noise gets added to the MR images due to imaging devices. Hence to reduce noise Gaussian filters are used that suppress the noise and thereby improve the quality of noise. In the training phase of classification the data points are assigned labels with class. The testing phase has unlabeled points and the algorithm generates list of K- nearest data point [5].

Gomathi.P et al. have used Naive Bayes to increase the overall speed and accuracy of the knowledge discovery process. The data set can be predefined or as available. The proposed method starts with the user searching for disease diagnosis by entering the symptoms. The symptoms are preprocessed and keyword matching is done that identifies multiple diseases. Finally pattern matching is done to find the accurate disease [6].

3. ALGORITHMS

3.1 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. Implementation of KNN mechanism is easy and the debugging process is very faster [7]. As value of k decreases noise points in training set increases. As value of k increases it becomes expensive.

3.2 LBR

For any given unlabelled test example, LBR- Meta starts from the global Bayesian rule where the antecedent is empty and the consequent is trained on the whole training set *D*. At each step, an attribute-value pair is selected and added into the antecedent to reduce the current instance subspace, and hence reduce the corresponding local training subset. There are two key problems to be solved in the LBR-Meta algorithm. The first key problem is how to select an attribute-value pair. After an attribute-value is added to the antecedent and the subspace is refined to a further smaller subspace, the second problem appears: how to determine which local classifier is best suited for this reduced subspace [8].

4. PROPOSED SYSTEM

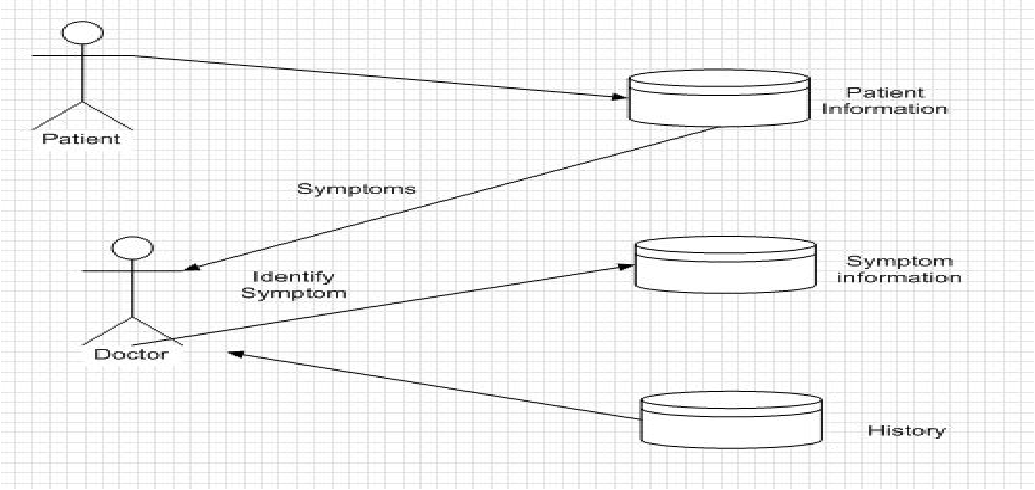

Figure 1. Architectural view of system

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

5. 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.

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.

6. RESULTS

Figure 2 shows the graph for training datasets and timerequired.

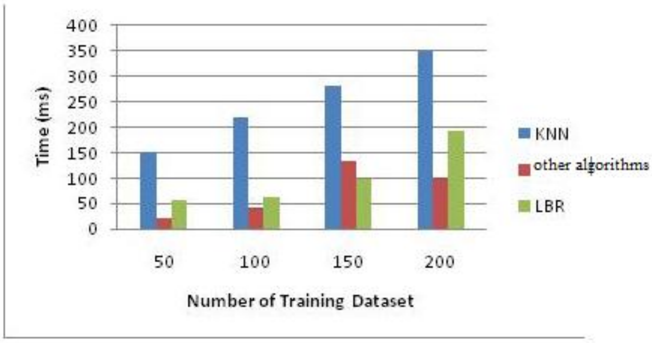

Figure 2: Training datasets versus Algorithm Accuracy

The performance of LBR shows maximum results as compared to the other algorithms.

7. 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.

8. ACKNOWLEDGMENTS

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.

9. REFERENCES

[1] Ian H. Witten, Eibe Frank, Mark A. Hall, "Data Mining", Third Edition, Elsevier, 2012.

[2] Margaret H. Dunham, "Data Mining-Introductory and advanced topics", Pearson Education, 2013.

[3] Rahul Isola, Rebeck Carvalho et al., "Automated Differential Diagnosis in Medical Systems using Neural Networks, k-NN and SOM",IEEE 2011.

[4] Hatice Catalouk and Metin KESLERA, "Diagnostic Software Tool for Skin Diseases with Basic and Weighted K-NN", IEEE 2012.

[5] Nikita V. Chavan, B.D. Jadhav et a.l, "Detection and Classification of Brain Tumors" ,International Journal of Computer Applications, Volume 112 – No. 8, February 2015.

[6] Gomathi.P et al., "Medical Disease Diagnosis using structuring text", IJCSET, Vol.5, No.05, May 2014.

[7] Ayse Cufoglu, Mahi Lohi and Kambiz Madani," Classification Accuracy Performance of Naïve Bayesian (NB), Bayesian Networks (BN), Lazy Learning of Bayesian Rules (LBR) and Instance-Based Learner (IB1) – Comparative Study", IEEE.

[8] LBR-Meta: An Efficient Algorithm for Lazy Bayesian Rules, Zhipeng Xie.

[9] David A. McMurrey, Joanne Buckley, "Handbook for Technical writing", Cenage learning, 2013.

[10] Lee, I.-N., S.-C. Liao, and M. Embrechts,"Data mining techniques applied to medical information", Med. inform, 2000.