Automated Medical Diagnosis using K-Nearest Neighbor Classification

Zaheerabbas Punjani[1],
B.E Student, TCET
Mumbai,
Maharashtra, India

Ankush Deora[2],
B.E Student, TCET
Mumbai,
Maharashtra, India

Varun Mishra[3],
B.E Student, TCET
Mumbai,
Maharashtra, India

Dr. Rekha Sharma[4],
Associate Professor, TCET
Mumbai,
Maharashtra, India

ABSTRACT

Health care practices involve collecting all kinds of patient data which would help the doctor correctly diagnose the condition the subject is likely suffering from. This data could be everything from the simple symptoms observed by the subject, initial diagnosis by a physician or a detailed test result from a lab. Thus far this data is only utilized for subjective analysis by a doctor who then ascertains the disease in play using his/her personal medical expertise.

We posit that there is definite potential for application of data mining routines on this rich reserve of patient data. Employing apposite data mining techniques, useful patterns and conclusions could be drawn from the raw data at our disposal. These findings could in turn be utilized in a number of productive ways like to carry out automated diagnosis, equip doctors with a better understanding of the causes and factors in play behind a particular disease. Automated diagnosis in particular could prove very useful for the determination of non-critical diseases or in cases where a doctor may not be available to carry out diagnosis such as being in a remote location. By studying different data mining techniques, we found a suitable technique for carrying out medical diagnosis using K-NN and achieved accuracy above 84%. We incorporated it in an application MedEval which will provide diagnosis, storing of patient information and provide visualization of that data.

General Terms

Data Mining, Classification, K Nearest Neighbor algorithm, Medical Diagnosis, Knowledge base, Visualization of Data.

Keywords

Data Mining, Classification, K Nearest neighbor, Medical diagnosis, Diagnosis application.

1. Introduction

With the advance in technology and methodology people have begun automating tasks in order to save time and energy. Most of these tasks were repetitive or simple, but nowadays we have begun automating processes which seemed impossible to do so before. Techniques in data mining has led to the ease of extracting not only information, but patterns and conclusions from data which previously had to be done manually. By using a technique of our own we created a web application that will automate the process of medical diagnosis for users who are in need of it where it isn't possible to consult a doctor.

The application is named as MedEval (Medical Evaluation) which will act as the first line of diagnosis.

The primary objectives of this project are to provide diagnosis to the user with the help of easy questions to determine the symptoms and match them with the data in the knowledge base and also provide precautions to be taken and suggested medicine (if the disease is not critical). There are already some diagnosis tools available, but the problem with those systems is firstly, complex questions for determining the diagnosis and secondly the ambiguous nature of the answers. The secondary objectives are to provide user with their diagnosis history and to provide visualizations based on the diagnosis data collected. These visualizations will be in the form of graphs and will provide information about the trending diseases according to the location of the user and the year.

2. Literature Survey

Disease diagnosis often done based on doctors experience and personal opinion rather than the data hidden in the medical data base, could in turn sometimes lead to wrong diagnosis and affects the quality of services provided by hospitals to the patients.

There exists different methods for disease classification, for example, heart disease classification using K-nearest neighbor classifier with optimal feature subset selection [1]. Symmetrical uncertainty was used as a goodness measure to rank the attributes and based on the ranking least ranking attributes are pruned. Feature subset selection is a preprocessing commonly used in machine learning, where subset of features available from the data is selected for application of a learning algorithm. The subset contains the least no. of attributes which would contribute to accuracy. Remaining and unimportant attributes will be discarded. These evaluated attributes are given to KNN algorithm which helps in classification of heart disease.

K-Nearest-Neighbor (KNN) is one of the most widely used data mining techniques in pattern recognition and classification problems.

In some cases it was also investigated if integrating voting with KNN can enhance its accuracy in the diagnosis of heart disease patients. The results show that applying KNN achieved an accuracy of

97.4% which is higher than any other published findings on the benchmark dataset used and also show that applying voting in the classification could not enhance the KNN accuracy in the diagnosis of heart disease [2].

Neural networks have also been used along with clustering algorithms to provide diagnosis, for example using Hopfield network, LAMSTAR Network and K-Means algorithm to assist the doctors to perform differential diagnosis along with the possible implementation using SOA technique [3]. 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 results. It reduces the effects of misdiagnosis, especially practitioners and students can also easily identify the diseases.

3. Proposed System

The proposed system MedEval consists of the following modules

3.1 Modules

1) Registration

The Registration module is used for registration of new users. The users can choose to avoid registration but will miss on the many perks of registering such as, individual diagnosis history, suggested medicine and also the access to the visualization feature which would allow users to view various graphs to get an idea about the prevalent disease in the area of their choice.

2) Diagnosis

This module is the brains of the system which is responsible for identifying the diseases based on the symptoms. It takes input from the user with the help of a simple form and uses the knowledgebase containing information about the diseases and their symptoms and finally provides the result using the K-nearest neighbor algorithm.

3) Diagnosis History

This module is available only to the registered users and provides user with all their previous diagnosis information along with the suggested medicine (if disease wasn’t critical). The information is pulled from the public database containing the diagnosis history.

4) Visualization Module

This module again is available only to the registered users. The module provides visualizations based on the diagnosis data collected. These visualizations will be in the form of graphs and will provide information about the trending diseases for a particular location from a list of locations.

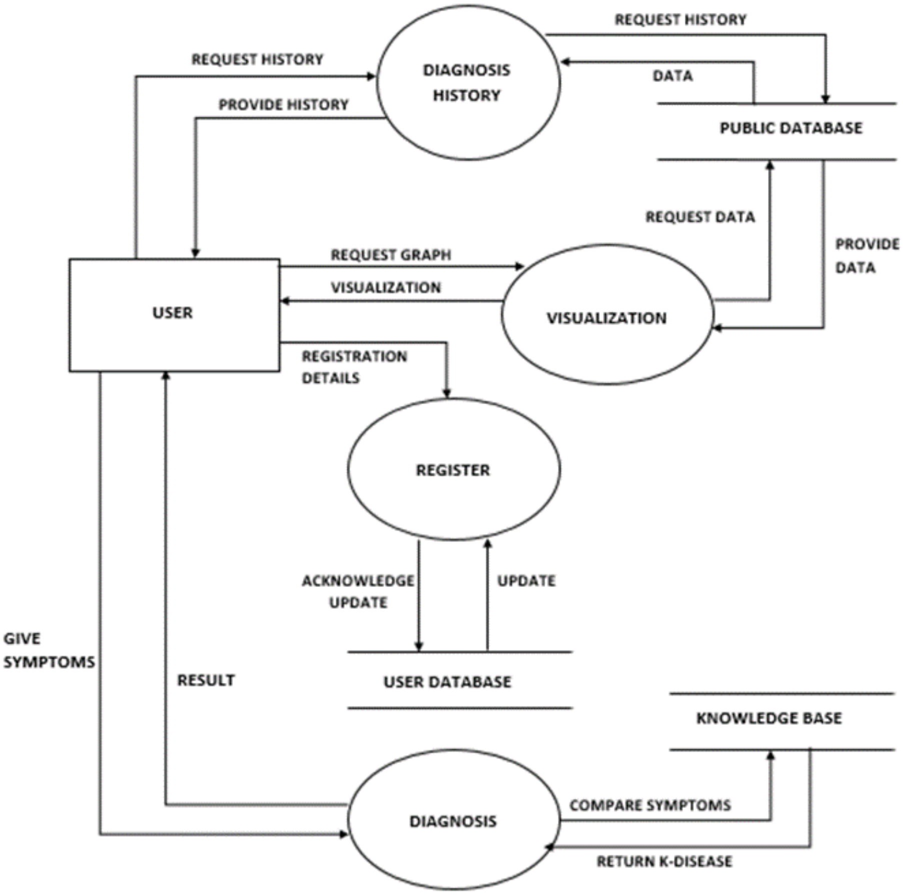


Fig 1. Fig 1. Level 1 DFD for MedEval Diagnosis System

3.2 Technique Used

K-NN is the method that would help us in the diagnosis process of the application. The K-NN has to be modified to suit our data set as it is not for a particular disease, but for multiple diseases.

KNN is told to be one of the most simple & straight forward lazy supervised learning data mining technique. It's also called as memory based classification as the training samples are required to be in the memory at run time. KNN became popular due to its simplicity and relatively high convergence speed. KNN is called lazy learning as it does not have any training phase. In the classification step, we will be given an instance S; whose attributes we will refer to as S.Ai and we wish to know instance class. KNN classification has two stages

1) Find the k instances in the data set that are closest to S

2) These k instances then vote to determine the class of S

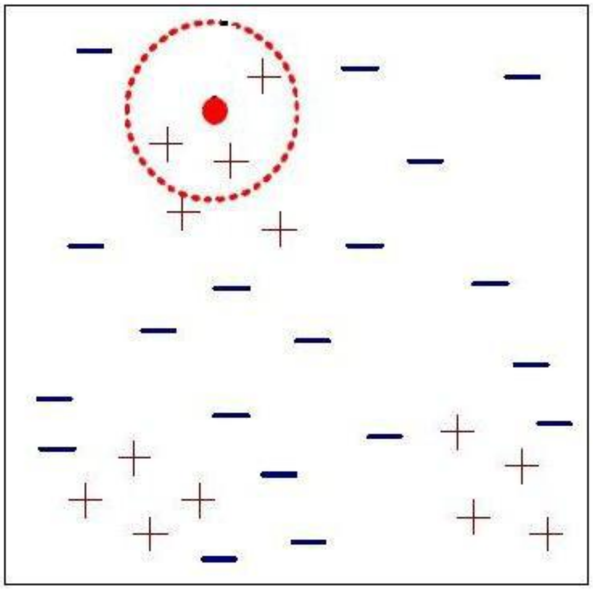


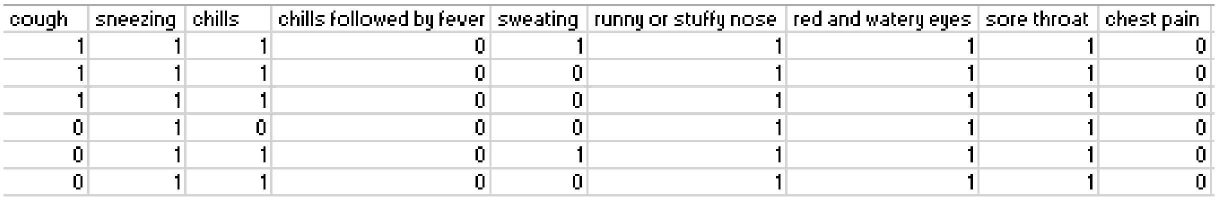
Fig 2. K Nearest neighbor classification

Assume that we have a training data set T made up of (xi) ε [1, |T|] training examples. The samples are described by a set of features F and numeric features are normalized in the range [0, 1].Each training sample is labeled with a class label Cj ε C. We have to classify an unknown sample S. From [1] for each Xi T we can calculate the distance between S and Xi using Euclidean distance or any other distance formula. Assume if the first instance is (a1, a2, ---an) and the second instance is (b1, b2---bn), the distance between first and second instance is calculated by

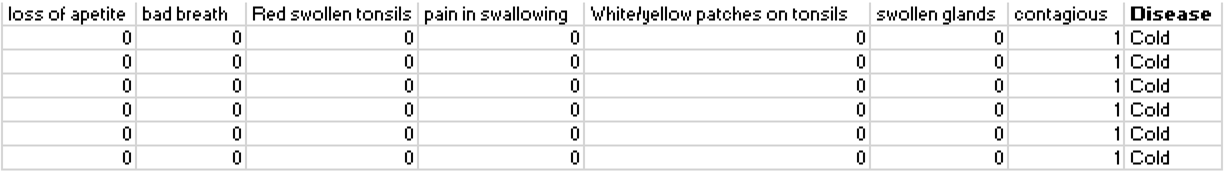
$$\sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \ldots\ldots(a_n - b_n)^2}$$

We'll be using a knowledge base that will be based on the possible symptoms for a particular disease. The reason is that the symptoms change for the same disease, later stages could be severe than before and these conditions have to be added to the knowledge base as well.

Table 1. Knowledge base example



cough	sneezing	chills	chills followed by fever	sweating	runny or stuffy nose	red and watery eyes	sore throat	chest pain
1	1	1	0	1	1	1	1	0
1	1	1	0	0	1	1	1	0
1	1	1	0	0	1	1	1	0
0	1	0	0	0	1	1	1	0
0	1	1	0	1	1	1	1	0
0	1	1	0	0	1	1	1	0



loss of apetite	bad breath	Red swollen tonsils	pain in swallowing	White/yellow patches on tonsils	swollen glands	contagious	**Disease**
0	0	0	0	0	0	1	Cold
0	0	0	0	0	0	1	Cold
0	0	0	0	0	0	1	Cold
0	0	0	0	0	0	1	Cold
0	0	0	0	0	0	1	Cold
0	0	0	0	0	0	1	Cold

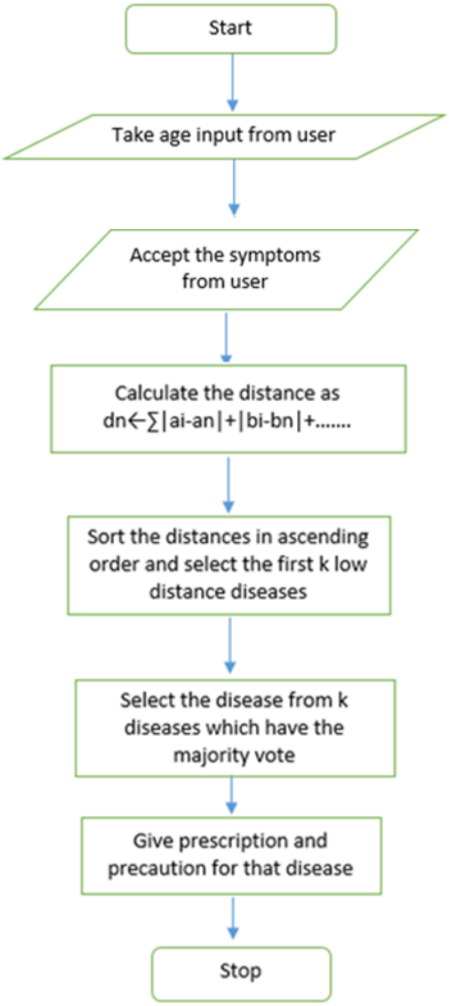

Start
Take age input from user
Accept the symptoms from user
Calculate the distance as dn←∑|ai-an|+|bi-bn|+.......
Sort the distances in ascending order and select the first k low distance diseases
Select the disease from k diseases which have the majority vote
Give prescription and precaution for that disease
Stop

Fig 3. Flow chat of algorithm

As we weren't able to get access to patient records, we started wondering as to how we will obtain medical data without the records. We finally used a knowledge base that contains information based on the knowledge of a doctor and medical websites [4], [5], [6] & [7]. The data was obtained on the basis of the theory that the diseases have one or more linchpin (primary) symptoms that are always present along with some symptoms that may or may not be present. The knowledge base also considers some anomalies such as extra unrelated symptoms that are visible. Also the knowledge base has combined information for adults and children.

The symptoms in the knowledge base are assigned weights for the degree of symptom denoted as an, bn,….

The input symptoms will also have their weights ai, bi,… and are compared with knowledge base entries to determine the disease

This distance is done by calculating the distance dn

The above Flowchart is used to calculate the distance of the input with each row in the database or rather the knowledgebase, using that distance the system finds the disease that matches closely to the input provided.

Some other methods are available for classification such as Neural Networks, Decision trees, Naïve bayes, etc. But the Dataset taken into consideration for classification is built in such a way that K-NN seems to be a better choice due to its simplicity and good accuracy [1], [2].

4. Methodology used

The project follows the software development life cycle and uses the spiral model. The spiral model consists of the phases shown in the figure.

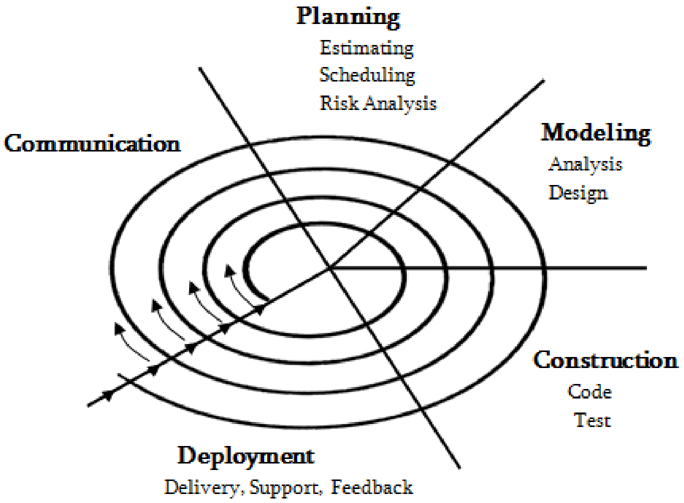


Fig 4. Spiral model

The spiral model is similar to the incremental model, with more emphasis placed on risk analysis. The spiral model has five phases: Communication, Planning, Modelling, Construction and deployment. A software project repeatedly passes through these phases in iterations (called Spirals in this model). The baseline spiral, starting in the planning phase, requirements are gathered and risk is assessed. Each subsequent spirals builds on the baseline spiral. In case of MedEval, the aim of each spiral was to refine both the algorithm and the knowledge base until it was in the satisfactory accuracy range and also improve the quality of the graphs in the visualization module.

5. Results

For testing the application for accuracy we collected data from different sources such as surveys conducted on people to ask for their symptoms, data collected from the doctor and symptoms mentioned by people on various medical forums [8], [9]. The data was collected for 6 diseases.

The value of K selected in the K nearest neighbor algorithm was selected as 5.

Table 2. Accuracy testing results

Disease	Accuracy
Cold/Viral fever	97.4 %
Tonsillitis	96 %
Typhoid	84 %
Pneumonia	94.42 %
Malaria	97.77%
Food poisoning	100 %

The minimum accuracy was for Typhoid which was 84% and the highest accuracy was for Food Poisoning which was 100%.

6. Conclusion

The MedEval Diagnosis Tool will help the users get information about their condition from anywhere and help them get an idea about the diseases trending in their city. MedEval not only provides diagnosis, but also provides the registered users with their diagnosis history and various graphs for visualization of all the diseases in their area, so that they can better prepare themselves against various diseases.

The accuracy testing with minimum accuracy being 84% and maximum being 100% makes sure that with enough innovation and data from the experts, medical diagnosis tools will one day become more reliable and help users from the comfort of their homes and in times of epidemics where doctors can focus on the immediate threat and leave the diagnosis of common diseases to the diagnosis tools.

7. Future Scope

The MedEval system will continue to grow by incorporating more diseases. The algorithm and knowledge base will undergo continuous refinement in order to provide users with flawless diagnosis.

We also plan to divide the users based on the age, so that the data collected from the diagnosis can be further used for data mining applications such as drawing conclusion on the occurrences of particular disease in a certain age group/groups and graph creation based on the age group to make the user more aware, for example, the increase of Cold/Viral fever cases in children can help the parents take more care of their children. The Knowledge base can be broken into different parts based on the type of disease such as Stomach disorders, skin infection, etc. This will help simplify the questions asked for the symptoms by reducing the amount of questions and branching them on the basis of some preliminary questions.

8. References

[1] M. Akhil Jabbar, B. L. Deekshatulu and Priti Chandra, "Heart disease classification using nearest neighbor classifier with feature subset selection", Annals. Computer Science Series, 11th Tome 1st Fasc, 2013

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

[3] Sathyabama Balasubramanian and Balaji Subramani, "Symptoms based diseases prediction in medical system by using k-means algorithm", International Journal of Advances in Computer Science and Technology, 3(2), February 2014

[4] Mayo Clinic health information, http://www.mayoclinic.org/

[5] WebMd health information, http://www.webmd.com/

[6] National health service UK Cold symptoms, http://www.nhs.uk/Conditions/Cold-common/Pages/Symptoms.aspx

[7] Healthline medical information, http://www.healthline.com/health/

[8] Patient health discussion forum, http://patient.info/forums/

[9] Topix Food poisoning forum, http://www.topix.com/forum/health/food-poisoning