A Novel Method for Disease Recognition and Cure Time Prediction Based On Symptoms

Mani Shankar*, Mayank Pahadia†, Divyang Srivastava‡, Ashwin T S§, G. Ram Mohana Reddy¶
Department of Information Technology
National Institute of Technology Karnataka
Email: *mas.11it42, †mkp.11it45@nitk.edu.in, ‡divyang.srivastava007, §ashwindixit9@gmail.com, ¶profgrmreddy@nitk.ac.in

***Abstract*—Healthcare is a sector where decisions usually have very high-risk and high-cost associated with them. One bad choice can cost a person's life. With diseases like Swine Flu on the rise, which have symptoms quite similar to common cold, it's very difficult for people to differentiate between medical conditions. We propose a novel method for recognition of diseases and prediction of their cure time based on the symptoms. We do this by assigning different coefficients to each symptom of a disease, and filtering the dataset with the severity score assigned to each symptom by the user. The diseases are identified based on a numerical value calculated in the fashion mentioned above. For predicting the cure time of a disease, we use reinforcement learning. Our algorithm takes into account the similarity between the condition of the current user and other users who have suffered from the same disease, and uses the similarity scores as weights in prediction of cure time. We also predict the current medical condition of user relative to people who have suffered from same disease.**

I. Introduction

The use of computer systems in decision making, prediction and recommendation has been a trending topic of research for more than a decade. The recent advances in medical science can be attributed to advances in computer technology. But, the prediction of medical behavior is still a very challenging task which is done with the help of a medical professional. The occurrence of every disease shows a pattern based on its symptoms. The main focus of this paper is to propose a system to exploit these patterns for predicting the associated diseases and the time that might be spent on their treatment. The core idea behind this was that every symptom of a disease has a unique impact on severity and recovery time. Our system tries to quantify this.

By prediction we mean to forecast an occurrence of a condition based on some mathematical calculation. For implementing this prediction, we need a recommender system[1]. A recommender system[2] is a system which reads an input, finds a pattern in it which is based on the dataset given to train the system. Based on the pattern it figures out a solution for the problem.

A naïve solution can be to create a database of every possible disease and its symptoms and predicting diseases based on that. The biggest drawback about this solution is that the efficiency and speed of this solution are very less and the size of this dataset would be very large.

The solution that we suggest is, of using the symptoms with the ratings given by the patient, to predict the possible diseases and the possible cure time of these diseases. Our solution is novel and better because we predict the diseases based on the severity of the patients symptoms and the cure time prediction is based on real-life data given by other patients. For an accurate prediction we give different coefficients to all the symptoms possible for a disease. We have gathered the data related to diseases and their symptoms from two sources - Wikipedia [3] and WebMD [4].

The rest of the paper is organized as follows. In Section 2, we discuss about the related research done in this field. In Section 3, we give a thorough methodology of our system. The experimental setup and the results obtained by us are explained in Section 4. Finally the conclusions and future work are given in Section 5.

II. Background and related works

Wasan et. al. [5] have proposed application of various data mining techniques as diagnostic tools to identify patterns in medical data. They have identifed knowledge discovery in hospital management systems as a potential field which can benefit from application of such techniques. Data mining is very useful in discovering patterns in large amounts of data. Medical diagnostics is one such field where pattern discovery can be of immense use. The applications of such techniques have been presented by Scales et. al. in [6]. Durairaj and Ranjani [7] have performed a comparative study of data mining algorithms and tools on various diseases. They have also analyzed the success rate of medical techniques over existing datasets. They have concluded that combination of multiple data mining methods may yield better results in medical domain. A method for predicting disease risk through feature selection has been proposed by Yang et. al. [8]. They have applied random forest and SVM techniques for this purpose on multiple UCI datasets.

Meisamshabanpoor and Mehregan Mahdavi [9] proposed a methodology for prediction of diseases and their cure time based on symptoms. Their method classifies diseases into groups based on age and Body Mass Index. They have developed a collaborative filtering method which considers neighborhood selection for prediction. Their method does not take into account different coefficients/weights for different symptoms of a disease. S Sudha and S Vijiyarani [10] suggested to use data mining techniques to predict diseases mainly of three kinds. They focused more on heart diseases, diabetes and breast cancer. They used different algorithms for predicting different diseases.

Most of the literature on the subject of disease prediction

978-1-4799-1734-1/15 $31.00 © 2015 IEEE
DOI 10.1109/ICACCE.2015.66

IEEE computer society

using data mining techniques is centred around any one disease. For example, in [11], the authors have used a hybrid technique to predict Asthama Disease. They have used a combination of Naive Bayes and Neural Network. In [12] Krishnaiah et. al. have suggested the use of Fuzzy K-NN approach for heart disease prediction. In [13], the authors have used WEKA tool with 10 fold cross validation to predict Dengue Disease. Their dataset consists of features derived from symptoms of disease as well as the current condition of the patient. Although the experiments performed in above-mentioned papers are very useful and yield good results, they are limited to only one disease. In the proposed work, we try to overcome this barrier by introducing a generalized approach of disease prediction.

The proposed approach is based on reinforcement learning [14] by Barto. In this method, desired outcomes are rewarded highly while undesired outcomes are given low rewards. This ensures that desired outcome is identified correctly most of the times when encountered.

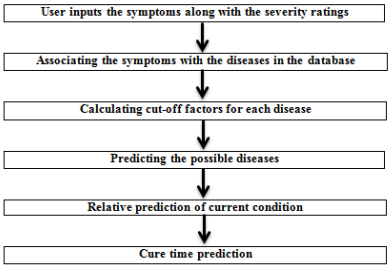
User inputs the symptoms along with the severity ratings
Associating the symptoms with the diseases in the database
Calculating cut-off factors for each disease
Predicting the possible diseases
Relative prediction of current condition
Cure time prediction

Fig. 1: Flow Chart of the Proposed System

A. Reading Input

In this step the user enters the symptoms he/she is facing with the severity ratings.

B. Associating symptoms with diseases

In this step the symptoms are then matched with the database entries, and the common symptoms and the possible diseases are selected for further processing.

C. Calculating cut-off factors

In this step we calculate the cut-off factor based on the symptoms given. The process of calculating the cut-off factor is as follows : Suppose the user enters n symptoms S_1, S_2, ..., S_n with severity ratings R_1, R_2,..., R_n.

Suppose m distinct diseases D_1, D_2, ..., D_m in our database have some symptoms in common with the symptoms specified by the user. For each such disease, we calculate three different parameters 1,2 and 3:

$$CS_i = (t_i/s_i) * 100 \tag{1}$$

$$CRS_i = (\sum_{i=1}^{t_i}(b_i)/\sum_{i=1}^{s_i}(a_i)) * 100 \tag{2}$$

$$SR_i = ((\sum_{i=1}^{t_i}(b_i * k_i))/(\sum_{i=1}^{t_i}(b_i)) * 5) * 100 \tag{3}$$

where, CS = Percentage of Common Symptoms

CRS = Percentage of Comparison Rating of Symptoms

SR = Percentage of Severity Rating

s_i = number of symptoms for D_i

t_i = number of common symptoms between user input and D_i

r_i = [a_1, a_2,...., a_{s_i}], comparison rating for each symptom of D_i

w_i = [b_1, b_2,...., b_{t_i}], comparison rating for common symptoms between user input and D_i

x_i = [k_1, k_2,...., k_{t_i}], severity rating for common symptoms between user input and D_i

The above-mentioned three factors can be combined to make a common cut-off factor as given by 1:

$$CF_i = ((CS_i^2 + CRS_i^2 + SR_i^2)/3)^{0.5} \tag{4}$$

where, CF = Cut-off Factor

The normalization of Cut-off Factor requires a division by $3^{0.5}$.

The cut-off factor for each disease will be calculated in the manner described above.

D. Disease Prediction

In this step, the minimum score for each disease, min (CF_i) will be stored in the database and updated periodically. At the time of disease prediction, the cut-off factor for each disease Dl is compared with the minimum cut-off factor for that disease. If for a disease D, CF_i min (CF_i); then that disease is displayed to the user as a part of prediction.

E. Relative Condition Prediction

In this step, the Severity Rating for the predicted disease will be used to give the user an approximate idea of his condition on a quantitative scale.

Suppose the Severity Rating of a predicted disease is SR_p . For the same disease, find the number of users from the database who have similar or higher severity rating. Let this number be N_p. Also, find out the total number of users who have suffered from this disease in the past. Let this number be N_t. Therefore, the percentage of users with same or higher severity rating is given by 5:

$$N\% = (N_p/N_t) * 100 \tag{5}$$

The reason for including the higher severity-ratings in the above calculation is to give a relative measure to the user to approximate his/her condition with respect to others. This cannot be done if only similar severity-ratings are considered.

F. Cure Time Prediction

In this step, suppose for a predicted disease D_i severity rating is SR_i. Find all the severity ratings of this disease from the database table. Assume the ratings are [DSR_1, DSR_2,., DSR_k]. Moreover, find the corresponding cure time in each case. Assume the cure time be denoted as [DC_1, DC_2, , DC_k].

For each DSRj in the list, the number similarity[15] is computed as 6:

$$NS_j = 1 - |SR_i - DSR_j|/(|SR_i| + |DSR_j|) \tag{6}$$

The cure time can be predicted as 7:

$$C_i = (\sum_{i=1}^{k} NS_i * DC_k)/(\sum_{i=1}^{k} NS_i) \tag{7}$$

These predictions are based on the datasets which we used. These datasets were made by us and contained the medical history of some users. The dataset contained the diseases which we had already suffered, the duration in which it was cured and the symptoms which we faced with the severity rating. The cut-off factors were first calculated and then the minimum cut-off factors were defined.

III. Results and Analysis

The main purpose of our approach is to predict the disease based on symptoms. For implementing this methodology, we made a web interface which had two main functions. The first one was to take the medical history of the patient and the second one was for taking the present symptoms of the patient with the severity rating.

For implementing this methodology, we needed a big dataset. We made this dataset using the information we got from the students in our campus. The dataset consisted of the weight, height, age and the medical history of the student. In the medical history, the student had to tell about all the diseases that he/she suffered from, the time it took for it to cure and the symptoms that he/she suffered with the severity rating of 1-5 as shown in Table 1.

TABLE I: Disease and Symptoms

Serial Number	Disease	Symptom	Rating
1	Hepatitis-A	Fever	2
1	Hepatitis-A	Fatigue	2
1	Hepatitis-A	Jaundice	3
1	Hepatitis-A	Dark Urine	2

The previous paper[9] proposes a methodology to predict diseases and the cure time of diseases based on symptoms. Their proposed methodology assumes equal importance to all the symptoms. In our paper, we give a severity rating to all the symptoms, because of this the prediction is better than their prediction. S. Vijiyarani and S.Sudha[10] proposed data mining methods for prediction of selected diseases where as our approach is based on all types of diseases.To the best of our knowledge our implementation will provide better results when compared to existing works.

In our implementation, the output matched quite fairly with the expected results. With the dataset we used, the accuracy that we got was quite high.

We were able to predict several types of diseases and the cure time with a very good accuracy. This is quite a difficult task because predicting a disease and its cure time is based on different criterias like immune system of the patient etc.

To the best of our knowledge, we can say that our method is superior to all the other existing works because we predict the diseases and the cure time based on the symptoms that the patient mentions, which is similar to the real life doctor-patient interaction.

Our method is a preliminary step towards disease prediction and cure time prediction. But these predictions are not a replacement for the existing medical tests and procedures, they are just a basic step towards finding out the problems faced by the patient.

IV. Conclusion and Future Work

As we have already mentioned, we address the problem of predicting diseases and their respective cure time based on the symptoms. The main focus was on the classification of symptoms based on their severity and importance and using this knowledge to calculate a numerical value to identify diseases. Although the method was tested in a limited environment with high accuracy, it can be extended to larger settings. Apart from this, we also estimated the cure time of a disease based on the experiences of other patients. We also provide a severity rating for the current condition, relative to the other users with similar symptoms. The future work can focus on using the medical history of the user with current symptoms in prediction of diseases. The test results for various medical conditions can be used to further improve the reliability of the system. Since the results are dependent on the experience of previous users, it is important to isolate genuine experiences from fake ones.

V. Acknowledgement

We would like to whole heartedly thank Mr. Shridhar Domanal and Mr. Subhayan Mukherjee for their continuous support and guidance through our project. We would also like to thank the fellow students of NITK for sharing their medical history.

References

[1] Gediminas Adomavicius and YoungOk Kwon. New recommendation techniques for multicriteria rating systems. *Intelligent Systems, IEEE*, 22(3):48–55, 2007.

[2] Paul Resnick and Hal R Varian. Recommender systems. *Communications of the ACM*, 40(3):56–58, 1997.

[3] Wikipedia. List of medical symptoms, 2015. [Online; accessed 22-January-2015].

[4] WebMD. Disease symptoms, 2015. [Online; accessed 22-January-2015].

[5] Siri Krishan Wasan, Vasudha Bhatnagar, and Harleen Kaur. The impact of data mining techniques on medical diagnostics. *Data Science Journal*, 5(19):119–126, 2006.

[6] Roshawnna Scales and Mark Embrechts. Computational intelligence techniques for medical diagnostics. In *Proceedings of Walter Lincoln Hawkins, Graduate Research Conference from the World Wide Web: http://www. cs. rpi. edu/˜ bivenj/MRC/proceedings/papers/researchpaper. pdf*, 2002.

[7] M Durairaj and V Ranjani. Data mining applications in healthcare sector a study. *Int. J. Sci. Technol. Res. IJSTR*, 2(10), 2013.

[8] Jing Yang, Dengju Yao, Xiaojuan Zhan, and Xiaorong Zhan. Predicting disease risks using feature selection based on random forest and support vector machine. In *Bioinformatics Research and Applications*, pages 1–11. Springer, 2014.

[9] Meisamshabanpoor and Mehregan Mahdavi. Implementation of a recommender system on medical recognition and treatment. *IJEEEE*, 2(4):315–318, 2012.

[10] S Sudha. Disease prediction in data mining technique–a survey. *IJCAIT*, 2(1):17–21, 2013.

[11] Saloni Aneja and Sangeeta Lal. Effective asthma disease prediction using naive bayesneural network fusion technique. In *Parallel, Distributed and Grid Computing (PDGC), 2014 International Conference on*, pages 137–140. IEEE, 2014.

[12] V Krishnaiah, G Narsimha, and N Subhash Chandra. Heart disease prediction system using data mining technique by fuzzy k-nn approach. In *Emerging ICT for Bridging the Future-Proceedings of the 49th Annual Convention of the Computer Society of India (CSI) Volume 1*, pages 371–384. Springer, 2015.

[13] Kashish Ara Shakil, Shadma Anis, and Mansaf Alam. Dengue disease prediction using weka data mining tool. *arXiv preprint arXiv:1502.05167*, 2015.

[14] Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 1998.

[15] Geoffrey I Webb Claude Sammut. Encyclopedia of machine learning. *Springer Science+Business Media*, 2011.