DESIGN AND IMPLEMENTATION OF EFFICIENT DISEASE PREDICTION SYSTEM USING REGULAR EXPRESSION

Rupeshsing Ramesh Patil	Mr. Dinesh D Puri	Dr. Manoj E. Patil
M.E Student	Assistant Professor	Associate Professor
Dept. of Computer Engg.	Dept. of Computer Engg.	Dept. of Computer Engg.
SSBT's COET, Bambhori	SSBT's COET, Bambhori	SSBT's COET, Bambhori
rupeshsingpatil@gmail.com	ddpuri@gmail.com	mepatil@gmail.com

Abstract

Healthcare is a sector involving decisions which have very high cost and risk associated with them. One bad choice can cost a person his/her 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. Using Regular Expression, prediction of diseases with the help of symptoms proves to be of great help in treating patients. An expression which is used for suggesting a set of strings needed for a specific purpose is called a Regular Expression. Listing members or elements is a very generic way for pointing out a set of strings which is finite. Many researchers have worked in disease prediction on the basis of symptoms using classification method of sequential search but use of regular expressions makes it more efficient and effective. It saves a lot of time and space.

Keywords: *Regular Expression, Machine Learning, Natural Language Processing.*

1. Introduction

Regular expressions (REs) provide an expressive and powerful formalism to capture the structure of events, messages and documents. Because of their expressive power, Regular Expressions (REs) are rapidly becoming an integral part of language specifications for many important application scenarios. An expression which is used for suggesting a set of strings needed for a specific purpose is called a Regular Expression. Listing members or elements is a very generic way for pointing out a set of strings which is finite. Prediction, Decision making and recommendation are some areas where the computer system has proved its worth. Its use in such fields is trending now a days. For more than a decade, the research is being done in such fields. The progress made recently in medical science field can be ascribed to the advancements made in the computer science field. Nevertheless, finding about medical behavior is a herculean task and has to be done by getting help from medical professionals. On the basis of symptoms, occurrence of all diseases shows a pattern. The system tries to find out probable disease by basing its decision on symptoms. Our system tries to quantify this.

Natural language processing (NLP) is a subpart of computer science, information engineering, and artificial intelligence pertained with the interactions between computers and human languages, in particular about programming computers to process, analyze large amounts of natural language data. The symptoms provided by the patient are in common tongue of humans but to make understandable for machines, NLP comes into picture.

Artificial Intelligence has an application and that is Machine Learning. It renders the system to learning new things and improvising from the past mistakes and experiences. This is done automatically without any involvement of explicit command or program. Developing a computer programs which can access data and later can make use of it for them for learning is one of the attempts of machine learning mechanism. It saves a great deal of time as it processes the data run time. Code learns from past cases and rectifies mistakes, resulting in better and fast output. Training data sets are used primarily for this purpose to train the algorithm.

Healthcare is a very high risk sector. Prediction of disease, by analyzing patients' symptoms, model can be very helpful in better care of patient and in addressing treatment. Regular Expression enables us to a very efficient and effective method of classifying and searching patterns from database.

2. Related work

- Mani Shankar et. al. in [1] presented the new method which recognizes disease and also predicts cure time. The whole prediction is based on symptoms of a patient. They have used sequential search method. The prime objective of this study is to recognize the disease from provided symptoms. In this paper, the author presented a reinforcement learning based approach given by Barto[2]. In this method, high rewards are given to the desired outcomes. Also the low rewards are allotted to the opposite outcomes that are undesired. This technique ensures one thing that desired outcomes gets recognized most of the time.

- Wasan et. al. [3] has worked on identification of patterns in medical data. They have designed an application which uses data mining techniques. Application can be used as diagnostic tool. Hospital management system is a potential field. Knowledge discovery in that promises great deal and it can be benefitted from such developed techniques.

- Scales et. al. [4] has used data mining technique. When we have large amounts of data, data mining promises a great help in finding patterns from that. For example Medical diagnostics is field where large amounts of data are generated. Identifying patterns in that may prove of immense value.

- Comparison between data mining algorithms and tools has seen surge recently. Durairaj and Ranjani [5] performed same comparative study on several diseases. Over Existing datasets, they have also examined a success rate of techniques related to medical field. In conclusion, combination of multiple data mining methods may yield better results in medical field.

- Yang et. al. [6] did some research in the field of predicting disease risk. A method for same uses a feature selection. Random SVM and forest methodologies have employed by them. The multiple UCI datasets have been used.

- Meisamshabanpoor and Mehregan Mahdavi [7] did research on prediction of diseases. The method used for predicting is based on symptoms. The cure time required for curing a disease is also predicted. BMI (Body Mass Index) and the age of a person are taken into consideration for classification of diseases in to separate groups. Collaborative filtering method is new addition in the system. It takes into consideration the neighborhood selection for predicting disease. The only drawback of the system is that of weights or coefficients for symptoms. That is not taken into consideration for the prediction of diseases.

- S Sudha and S Vijiyarani [8] used data mining method in their study. They have worked with focus on mainly three types of diseases. They also predict disease using above mentioned method. Separate algorithms have been used for separate disease. The main three types are breast cancer, diseases related to heart and diabetes.

3. Problem Definition

In today's scenario, the efficient prediction of diseases takes up great deal of importance due to ever increasing proneness of humans to illness. For algorithm to work, we have created a datasets of diseases and its symptoms. In this abstract, the main challenge is constructing the new data structure. Regular expression syntax is used on large number of patterns. That provided, it needs to be compiles into new data structure. It needs to be done to effectively increase the efficiency in the runtime operations. An input string has to be matched with the pattern from large number of patterns runtime with maximum efficiency. Every input string is matched with some pattern. That is the property of that set consisting of patterns. Sometimes there may be several patterns getting matched with the string. This is the main obstacle for classifying efficiently. Regular expression must be member of a class. Only in that case, it will be matched with the class from huge chunk of classes. Also each class has its own regular expression. Syntax in regular expression is very expressive. It can be used in many numerous applications. Any patterns which are expressed in any syntax can be converted in to regular expression syntax. It is very handy when it comes to string matching, sorting etc.

A. The objectives of the proposed system
  - To resolve errors in efficiently classifying symptom strings into set of patterns with an overall accuracy.
  - To reduce time and space requirements required using regular expressions
  - To improve the disease prediction using regular expression for better treatment.
  - To utilize machine learning to better prediction of results as cases come by.

3. Proposed Work

The system is trained to find a match in the dataset. The existing significant work done on disease prediction for many different diseases using different methods. But the work carried out on using Regular Expression is limited. In the proposed work, first the input string will be converted into regular expression using code, which then will be used to match with the preexisting database of diseases and its symptoms. The dataset will be in regular expression format.

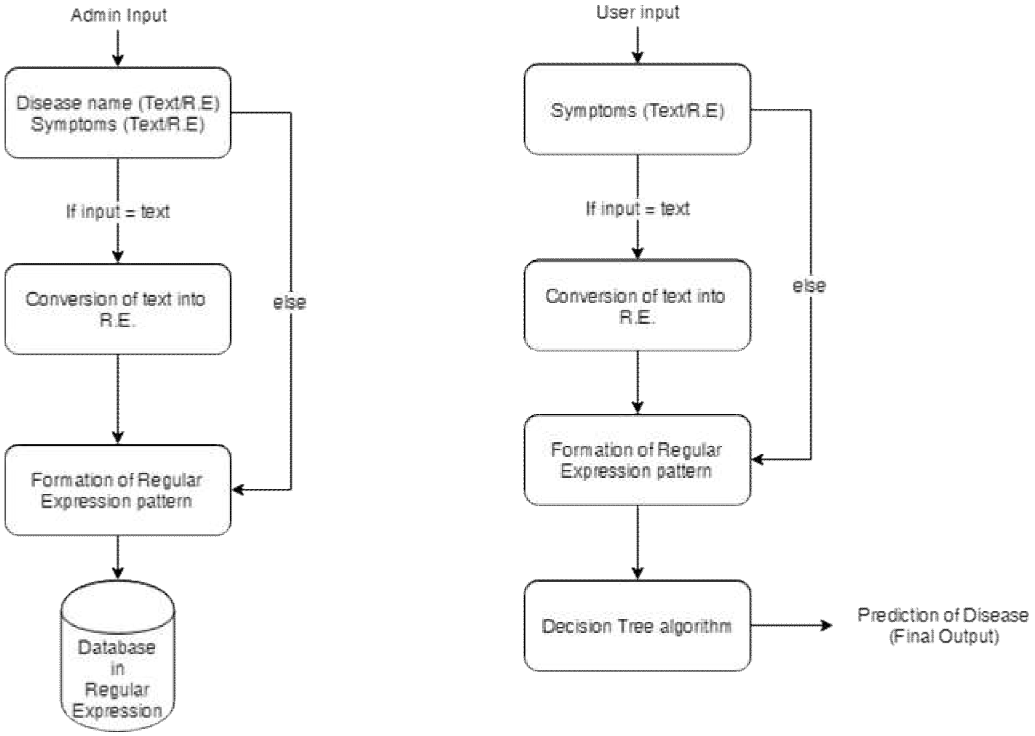
Admin Input
Disease name (Text/R.E)
Symptoms (Text/R.E)
If input = text
Conversion of text into R.E.
else
Formation of Regular Expression pattern
Database in Regular Expression
User input
Symptoms (Text/R.E)
If input = text
Conversion of text into R.E.
else
Formation of Regular Expression pattern
Decision Tree algorithm
Prediction of Disease (Final Output)

Figure 1 Proposed System Architecture - a. Training algorithm b. Testing

In this approach, initial tests are used to filter out some of the unwanted strings. It saves a lot of time as it reduces total number of computations required to be done. Initial test are to be chosen very carefully as quality of tests will matter a lot. The length of a string is taken into consideration for finalizing initial tests. The first character of a string or any character which might be present in the string anywhere are used for formulating tests. These tests provide us with the information which can be used for get rid of unwanted patterns. For example, let’s assume that string contains character ‘z’. If test fails, then we can conclude that string cannot match with other patterns as well. The next step is checking for remaining patterns. In the proposed work, Regular expression approach is used for efficient matching of strings in dataset.

The main introduction is about initial tests. The training data is used and the decision tree is compiled. The order to be followed is also decided.

The framework of proposed approach is described in Figure 1. The proposed framework shows the working of efficient disease prediction system using Regular Expression in which user gives symptoms in either textual form or regular expression form.

The steps involved in flow of the proposed system are given below:

1. Firstly, admin provides training data to the system
2. Once the system is set up, user enters symptoms in either textual or regular expression form
3. Algorithm searches match for input regular expression pattern
4. At last, Algorithm predicts a disease with the approximation in percentage.

5. Conclusion

The Regular Expression approach is well advanced which executes and gives output instantly. The method here used is way faster than any contemporary string classification methods like sequential matching. The applications are innumerable with this new way of classification due to less time taken as well as space for execution. This technique provides new paradigm in classification arena.

6. Acknowledgement

First and foremost, I would like to thank my guides and advisors Mr. Dinesh D Puri and Dr. Manoj E Patil, for being a wonderful instructor and always being in bright spirits. I am incredibly grateful to Principal, Head of the department Dr. Patnaik sir and the Professors in the M.E. Computer Engineering Program at SSBT's COET, Jalgaon. Without their time and support, I would not have been able to complete this work. Last, but not least, I would like to thank my family and friends who also supported me throughout my studies.

7. References

[1] Mani Shankar, Mayank Pahadia et al, “A Novel Method for Disease Recognition and Cure Time Prediction Based On Symptoms", 2015 Second International Conference on Advances in Computing and Communication Engineering.

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

[3] Siri Krishan Wasan, Vasudha Bhatnagar, and Harleen Kaur. “The impact of data mining techniques on medical diagnostics”, Data Science Journal, 5(19):119126, 2006.

[4] Roshawnna Scales and Mark Embrechts, “Computational intelligence techniques for medical diagnostics”, in Proceedings of Walter Lincoln Hawkins, Graduate Research Conference.

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

[6] 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.

[7] Meisamshabanpoor and Mehregan Mahdavi, “Implementation of a recommender system on medical recognition and treatment”, IJEEEE, 2(4):315318, 2012.

[8] S Sudha “Disease prediction in data mining technique a survey”, IJCAIT, 2(1):1721, 2013.

[9] Miranda Mowbray, William Horne, Prasad Rao, "Efficient classification of strings using regular expressions", Hewlett Packard Labs HPE-2017-03

[10] Rajiv Subrahmanham, Wei Huang, Dil Ibrahim and Debabrata Dash, “Scalable Security Information and Event Management”, HP Tech Con 2011