Prediction for Pulmonary Disease Based on Diagnostic Reciepes and Classification

[1]Nidhi, [2]AP Nidhi

Department of Computer Science, Swami Vivekanand Institute of Engineering & Technology,
Punjab Technical University, Jalandhar, India

Abstract: ***In this research work we have developed a strategy in which the various parameters that influence the occurrence of pulmonary disease have been gathered from survey of doctors who specialize in diagnoses of pulmonary disease and diagnostic recipes involving if the else rules were built and given labels, which were used as target for machine learning algorithms [Logistic , SVM, RBF, Naïve Bayes] for identification of input dataset of symptoms of subjects . Multiple designs of these classifiers were implemented and best possible machine algorithm was identified for implementing the complete methodology. Results shows that there was no absolute answer for the design and selection of best possible machine algorithm as evident from the results based on multiple statistical tests, therefore , distance from ideal values of statistical test to find best classifier with most optimized parameters was calculated and the classifier which had closest to these ideal values was found and declared the best classifier for identification of pulmonary diseases presence or absence .as per results naïve bayes classifier is performing best which is evident from the statistical test scores .***

Keywords: SVM, RBF, Naïve Bayes, Logistic, pulmonary.

1. Introduction

Considerable population of the world does not have access to even bare minimum healthcare facilities. Majority of these are not able to avail these facilities because of lack of resources/funds, resultantly they suffer. Irony however is that most of the diseases are curable provided diagnosis and treatment is initiated at right time. Information Technology in general and predictive models in particular being used to facilitate the physicians in diagnosis of diseases. Bayesian networks have been extensively used in developing computer based systems for diagnosis of various diseases. The purpose of using these computer based systems in healthcare is to reduce the chances of faulty diagnosis and prevent cost and time over runs. However, the scope of the study in this research work will be limited as we will be confined to the information obtained. Doctors and dentists along with pilots, police, miners and social workers are considered to be members of high stress occupations [1]. This primarily is because of the reason that in these professions scope for error is minimal and in case a mistake happens there is a direct impact on the human life and psyche. It is therefore of utmost importance that people engaged in these professions exercise extreme precaution in decision making and ensure minimal error. In a country like India incorrect diagnosis may have severe implications. The current annual per capita public expenditure on health in the country is no more than Rupees Two Hundred, which is among the five lowest in the world. As per National Health Accounts [2], the total health expenditure in India from all the sources was Rupees 1, 33,776 crores, consisting of 4.25 percent of the GDP. Of the total health expenditure, the share of private sector was highest at 78 percent (private individual households 71 percent and private firms 7 percent); public sector at 20 percent (Central Government 7 percent, State Government 12 percent and Local Government 1 percent) and external flows contributed about 2 percent.

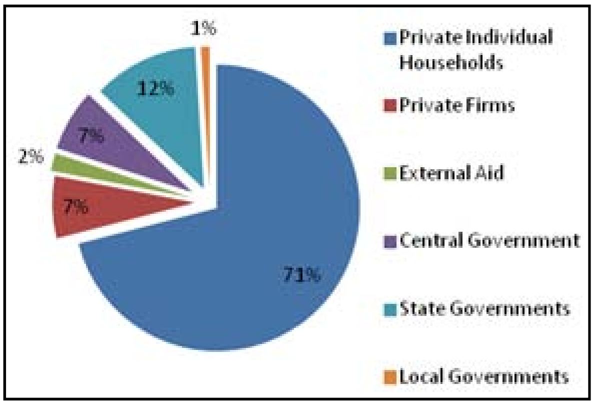

Figure 1: (Health Expenditure India – 2004) - Source: National health accounts 2004-2005

It has been estimated that less than 20 percent of the population, which seek OPD and less than 45 percent of that which seek indoor treatment avail of such services in public hospitals. These is despite the fact that most of these patients do not have the means to make out – of – pocket payment for private health services except at the cost of other essential expenditure for items such as basic nutrition. Stated plainly this means most of the people in India do not have the capacity to pay for the expensive tests. As such, it is very important for the doctors to make a correct diagnosis and if possible with minimum reliance on laboratory tests and more on their expertise based upon the presenting symptoms.

The present study is targeted to assist the doctors in making correct diagnosis by using the advantages offered by computer based systems. Computer based systems are expert systems designed to process knowledge and reach diagnostic decisions in the same way as human experts.

2. Proposed Work

Since there is an urgent need to automate the process of analysis of the datasets of symptoms in terms of taking advantage of machine algorithms, so that availability of large datasets of symptoms of patients suffering from diseases can come aid to the current medical practitioner to diagnose diseased that can be easily be identified by using diagnostic recipes based on the experience of the doctors. Therefore, we suggest method that can utilize in both the alternative and mainstream medical field for easy and accurate diagnostics of patients having pulmonary disease or not.

3. Implementation Steps/Results

1) Literature survey and interaction with specialists for collecting information pertaining to pulmonary diseases.
2) Collection of data pertaining to symptom matrix.
3) Data preprocessing to remove missing and superfluous values.
4) Framing “If-then rules” to labeling and identification of PD
5) Generating labeled dataset on the basis of these rules.
6) Running classifiers and selecting the most accurate classifier.

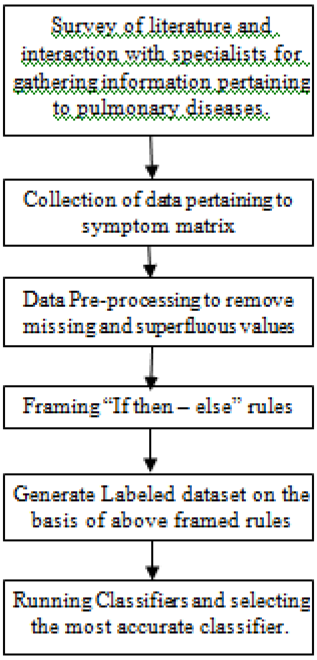
Survey of literature and interaction with specialists for gathering information pertaining to pulmonary diseases.
Collection of data pertaining to symptom matrix
Data Pre-processing to remove missing and superfluous values
Framing “If then – else” rules
Generate Labeled dataset on the basis of above framed rules
Running Classifiers and selecting the most accurate classifier.

Figure 3.1: Flow chart of the entire methodology

The data pertaining to parameters under study were collected from the leading Bollywood websites.

- Data Normalization must be used to reduce the number of samples and the complexity of the neural network and the computation time of the neural network.
- For the classification schemes, it was found that training the model with a large number of test data and with fast training algorithm would greatly enhance the accuracy and hence the reliability of the system.
- The design of our classifier was done by running the neural network with different number of hidden layers and it was apparent from the graphs that it affected the accuracy [3]
- It was found that as we increase the number of hidden layers there was also an increase in computation time but high order of accuracy is also achieved until we have reached the maximum of hidden layers, therefore, we need an optimal combination of parameters to achieve 93.3% accuracy Figure 3.1 Methodology Flow Chart

Step 1: Identification of the symptoms

Preliminary data bank of symptoms - For this study initial data bank of symptoms was collected on the basis of a survey questionnaire to be administered to doctors seeking their opinion and rating of symptoms based upon the strength to correctly predict pulmonary disease. An initial data bank was thus developed for further classification and study. This was followed by prioritizing the symptoms for accurate disease prediction and diagnosis.

Symptom Selection - Symptom selection was regarded as the problem of feature selection. Symptoms are essential to diagnose a disease. Therefore, a strong predicting model of syndrome is based on key symptoms. In this phase, we investigated which symptoms influence the predicted syndromes most. [5]

Step 2 and 3 – Creation of symptom matrix

Based upon the above process a symptom matrix representing the symptoms associated with pulmonary diseases was created Table 4.1.

Table 4.1: Symptom Matrix

1	Symptoms	10	Swollen Ankles
2	Fever	11	Shivering
3	Wheezing	12	Blood in the cough
4	Shortness of breath	13	Morning Headaches
5	Cough Dry	14	Dizzy spells
6	Productive Cough	15	Restlessness
7	Difficulty in blowing out air	16	Slurring of speech
8	Weight Loss		
9	Tiredness and Fatigue		

Step 4 and 5 – Framing if then rules and Developing Labeled Dataset

On the basis of the symptoms If-then-rules were framed in consultation with the doctors. The labeled dataset so obtained is represented in Table 4.2

Table 4.2: Labeled dataset so obtained

Symptoms Present	*Predicted Diagnosis*
High Fever, Shivering, Fatigue, Blood in sputum	Pulmonary disease (Pneumonia)
Wheezing, Dry Cough,	Pulmonary Disease (Asthma)
Productive Cough, Fever, Morning Headache,	Pulmonary Disease (COPD)
Difficulty in blowing out air, Tiredness, Weight loss	Pulmonary Emphysema
Swelling of ankles, Dizzy spells, Restlessness, Slurring of speech	Advanced Pulmonary Disease (Advanced COPD)

Summarization of the results of the study is shown in Table 4.3.

Table 4.3: Summary of the results

Algorithm	*Logistic Regression*	*RBF*	*SVM*	*Naïve*
MAE	0.1	0.2	0.006	0.0003
AUC	1	1	0.994	1
KS	0.9878	1	0.988	1
RMSE	0.0781	0.1	0.078	0.002
FP	0.1	0.1	0.006	0.2
TP	1	1	0.993	1
Precision	1	1	0.994	1
Recall	1	1	0.993	1

4. Conclusion

The study was intended to develop a framework for pulmonary disease diagnosis using Bayesian Belief networks, to evaluate its effectiveness with respect to actual diagnosis made by the physicians and compare the results obtained with other prediction models.

As a pre-requisite to the study a dictionary of symptoms was created, these symptoms. A focus group discussion with doctors was carried out to group these symptoms in a manner that would assist in making correct diagnosis.

The results of the study indicate that naïve bayes classifier are better at predicting pulmonary diseases least MAE indicates that there is minimal deviation in predicted diagnosis and actual diagnosis being carried out by the doctors. AUC, Kappa value, TP, Precision and Recall all these parameters have values equal to 1. This is highly significant as already stated AUC describes effectiveness in predicting true positive rates averaged over all false positive rates. A value of 1 effectively means successful prediction of true positive rates. Unity value of Kappa statistics, True Positive (TP) and Precision (accuracy) further reinforces the effectiveness of Naïve Bayes Algorithm. Recall value is an indicator of the ability of the algorithm to reproduce the same results over a period of time. Recall value of 1 for naïve bayes classifier indicate that the algorithm when subjected to same conditions over different period of times gives the same results, indicating high reliability of the algorithm.

5. Future Scope

In a developing country like India having considerable population living below the poverty line, there is an urgent need to use information technology in medicine. One such area where IT can be used is, in assisting the physicians to diagnose a disease without much reliance on expensive laboratory tests. Further, in-order to make health care accessible and affordable to entire population of the country, this (using IT) may be the only alternative.

The focus of the present study has been to develop a Naïve Bayes for diagnosing pulmonary diseases. As the results indicate the suggested algorithm performs better on most of the parameters under study, however, further improvements are possible. Table 5.1 indicates that probably more accurate results can be obtained by combining Naïve Bayes and RBF algorithms, since on some parameters (RMSE and FP) RBF performs better than Naïve Bayes [7].

The present study only focuses on absence or presence of pulmonary disease. It does not further go into pin pointing the type of pulmonary disease (whether it is pneumonia, asthma, COPD etc.). A more in depth study may be required to go into the detailed analysis.

6. Acknowledgement

I am thankful to A P Nidhi, Assistant Professor, Swami Vivekanand Institute of Engineering & Technology, Banur, for providing constant guidance and encouragement for this research work.

References

[1] L.Zhang, J.Luo, S.Yang. “Forecasting Box Office Revenues of Movies with BP Neural Networks”. Expert Systems with Applications 2009, vol. 36 (3) part 2, page 6580-6587.

[2] K.J.Lee, W. Chang. “Bayesian Belief Network for Box Office Performance: A Case Study of Korean Movies”. Expert Systems with Applications, 2009, vol. 36 (1), page 280-291.

[3] Nidhi and AP Gurpinder Kaur, Pulmonary Disease Diagnosis Using Bayesian Belief Networks: a Conceptual Framework, International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[4] T.Charitos, L.C.V.Gaag, S Visscher. “A Dynamic Bayesian Network for Diagnosing Ventilator Associated Pneumonia in ICU Patients”. Expert Systems with Applications 2009, vol. 36 (2), page 1249-1258.

[5] J. C. Burges, “A tutorial on support vector machines for pattern recognition,” Data Mining and Knowledge Discovery, vol. 2, no. 2, pp. 121–167, 1998.

[6] Y. W. Chen and C. J. Lin, “Combining SVMs with various feature selection strategies,” Taiwan University, vol. 207, pp. 315–324, 2006.

[7] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” Machine Learning, vol. 29, no. 2-3, pp.131–163, 1997.