Liver disease prediction using machine learning

Prof. Rageena M, Md Irfan Khan, Pooja Tandle, Pawar Arathi
Dept of Computer Science and Engineering. Guru Nanak Dev Engineering College, Karnataka, India

ABSTRACT

Data Mining technologies have been widely used in the process of medical diagnosis and prognosis, extensively. These data mining techniques have been used to analyze a colossal amount of medical data. Currently there is one of the prevalent diseases of 21st century is liver disorders annually killing so many people's round the worlds. A range of therapies have been provided by researcher to evaluate results. Early diagnosis is of considerable amount of significance in treating the disease. Diagnosis is of the physician skills conducting based on their knowledge's and experience yet an error might occurrence is here. Using various Artificial Intelligence methods for liver disorders diagnosis has recently become wide spreading data. These intelligent helps systems help physicians as diagnosis assistants. Now, various Artificial Neural Network system, Expert Systems, Fuzzy Neural Network, Classification, This paper provides a review of different Artificial System and expert system method in
diagnosis and detections of liver disease disorders acuteness is the key for results. Fuzzy Logic, and Swarm Intelligence are widely used.algorithm is considered to be a better performing algorithm when it comes to feature selection with an accuracy rate of 95.04%.

Keywords—*Artificial Intelligence; Artificial Neural Network; diseases dataset, Classification schemes, Training datasets, Machine learning, Classifiers, Algorithms, Classification models*

1. INTRODUCTION

The use of intelligence systems in medical diagnosis is increasing gradually. There is no doubt that evaluation of data taken from patients and the decisions of experts are the most important factors in diagnosis.

But, expert systems and different artificial intelligence techniques for classification also help experts in a great deal.

Classification systems, helping possible errors that can be done because of a fatigued or inexperienced expert to be minimized, provide medical data to be examined in a shorter time and more detailed. Problems with liver patients are not easily discovered in an early stage as it will be functioning normally even when it is partially damaged. Early diagnosis of liver problems will increase the patient's survival rate.

Liver disease can be diagnosed by analysing the levels of enzymes in the blood. Its weight comes around three pounds. The liver performs many essential functions related to digestion, metabolism, immunity and the storage of nutrients within the body. These functions make the liver as an important organ, without this, body tissues would quickly die from lack of energy and nutrients. Traditionally, liver disease can be diagnosed clinically by analyzing the levels of enzymes in the blood . In this research work, a combination of Naïve Bayes and Support Vector Machine (SVM) classifier algorithms are used for liver disease prediction.

2. LITERATURE SURVEY

Paul Mangiameli et al., [1] : Proposed model selection affects the decision support systems accurately. In their model selection, how to affects the accuracy of decision support system hydrides by single model and ensembles. They proposed single model is not more accurate than ensembles. Ahmed M. Hashem et al., [18] proposed to predict Liver Cirrhosis or fibrosis single stage classification model and multistage classification model. In their model based on Decision Tree, Neural Network, Nearest Neighborhood clustering and Logistic Regression.

Ziol.M et al.,[2]: proposed to evaluated liver fibrosis with chronic hepatitis C for patients using liver stiffness measurement (LSM).Z. Jiang.Z.,[4] proposed for discovering the corresponding degree of fibrosis by support vector machine (SVM).

Kemal Polat et al.,[3]: proposed resource allocation mechanism of AIRS was changed with a new one decided by Fuzzy-Logic. This approach called as Fuzzy- AIRS was used as a classifier in the diagnosis of Liver Disorders. In this Classification accuracies were evaluated by comparing them with reported classifier's accuracy, time and number of resources.

Piscaglia et al.,[4]: proposed to predict Liver cirrhosis and other liver-related diseases used by Artificial neural network. Dong-Hoi Kim et al.,[19] proposed machine learning technique and decision tree(C4.5).In this method is used for to predict the susceptibility to two liver diseases such as chronic hepatitis and cirrhosis from single nucleotide polymorphism(SNP) data . They also used to identify a set of SNPs relevant to those diseases.

N. Ramkumar, S. Prakash, S. Ashok Kumar, K Sangeetha, "Prediction of liver cancer using Conditional probability Bayes theorem" [5]: Malignant growth is the one of the unsafe infection on the planet. Malignant growth spreads in lungs, liver, bosom, bones and so forth. Liver malignancy is the most hazardous and it will proceed with long-lasting. The side effects of a liver malignant growth are Jaundice, loss of weight, yellow shaded pee, spewing, torment in the upper right stomach area, sweats, fever and amplified liver. The liver malignancy which starts in the liver separated from moving from another piece of the body is called as essential liver disease. A disease which spreads all other pieces of the body lastly it achieves liver is called an auxiliary liver malignant growth. The liver is one of the critical pieces of the human. WHO reviews state out of 100,000 individuals, around 30 individuals have experienced liver malignant growth and generally it influences the African and Asian nations prior. These days it turned into a well-known ailment. The most widely recognized sort of a liver malignant growth is called hepatocellular carcinoma, this specific influences male as opposed to female. The liver malignant growth happens for the most part because of the more liquor utilization. Numerous information mining calculations, artificial insight ideas are utilized to anticipate liver disease. The likelihood of anticipating the liver malignant growth is performed utilizing the Bayes hypothesis with the WEKA tool.

Mafazalyaqeen Hassoon, Mikhak Samadi Kouhi, Mariam Zomorodi Moghadam, Moloud Abdar, "Rule Optimization of Boosted C5.0 Classification Using Genetic Algorithm for Liver disease Prediction" [5]: One of the fascinating and vital subjects among scientists in the field of therapeutic and software engineering is diagnosing disease by considering the highlights that have the most effect on acknowledgements. The subject talks about another idea which is called Medical Data Mining (MDM). Undoubtedly, information mining techniques utilize diverse ways, for example, characterization and grouping to arrange maladies and their indications which are useful for diagnosing. This paper presents another technique for liver illness analysis to help specialists and their patients in finding the sickness side effects and decrease quite a while of diagnosing and counteract passings. The proposed strategy will streamline the tenets discharged from Boosted C5.0 grouping technique with the Genetic Algorithm (GA), to expand the determination time and exactness. So as opposed to utilizing a transformative calculation for creating rules, the hereditary calculation is utilized for improving and diminishing tenets of another calculation. We demonstrate that our proposed methodology has better execution and throughput in correlation with other work in the field. The precision is improved from 81% to 93% in our work

2.1 Findings

The mechanisms that are currently used in the prediction of liver disease are prone to have different levels of accuracy and effectiveness. The sense of importance, though, is determined by the need of the hour. Different diseases demand accuracy of a different set of parameters and might not demand the same set of inferences, throughout more than a single case. In the near future, the study reflects that there was a decent amount of accuracy that was achieved. However, the agenda of this paper is to improvise on those lines and come up with better accuracy standards. The slack in the accuracy in the recent cases had been tackled by designating different combinations to be considered, while the case study is being considered. The existing models are also reflective of certain issues that pertain to the handling of the training dataset and data elements.

3. ANALYSIS OF FACTORS AFFECTING ACCURACY

When it pertains to machine learning, the inferential information sets are a product of the commonly observed observations that often reflect this sense of pattern in the collection of data or the respective data set. There is a process of characterization, which is reaffirmed by the study and the accounting of a voluminous amount of data that relates to the context, and the study of the same. On the same lines, when the data is of limited volume and is curtly presented for consideration in the machine learning algorithm, it is difficult to come up with accurate and seemingly glaring inferential patterns and thereby the result of predictive analysis of a set of data. There are a set of issues that continue to challenge the accuracy of the machine learning algorithms that are used for predictive analysis.

(a) The Quantity of data that is involved: The concise, yet precise nature of this argument being - the more the data, the more accurate the result of the predictive analysis. With lesser data, the accuracy and effectiveness of the predictive process decline.

(b) Scope of the issue: With machine learning paradigms demanding a huge collection of data for analysis, it is important to give due importance to the selectiveness of the features, that would pivotally define the boundaries of context, in any given problem. It is important to maintain relevance and sync with the issue/problem statement.

(c) Parameters that are involved as a part of the method: The study and analysis of the algorithm and the larger system, as a whole should also be feasible to be executed by nontechnicians and absolute rookies with the basic understanding of the functioning of the system. In modern machine learning algorithms, the sense of innovation is reassured with the involvement of more than a single parameter that is involved in the analysis of the scope. These multiple parameter settings are induced and boundary by only the user's understanding, experience and the knack of being about to intuit the necessary parameters that need involvement/tweaking.

(d) Features in the data: It is imperative for any machine learning algorithm or the developer/data analyst to be able to sparsely collate the raw data and project the potentiality in the rich feature space. This is expected to accelerate the learning process of a machine learning system.

(e) Quality of Data: Any data that is to serve as a template for critical studies, fabrication, analysis and research of any subject - needs to be thoroughly checked on qualitative grounds. This is because even the slightest sense of lethargy can vandalize the integrity of the process and compromise on the potential and the expectancy, to be able to deliver.

4. SYSTEM ANALYSIS

4.1 Objective

The objective of this paper is to be able to predict the occurrence of liver disease in a sample dataset/ training data set, in order to be able to calculate the predictability to the greater magnitude of accuracy using the appropriate machine learning algorithm.

4.2 Present System

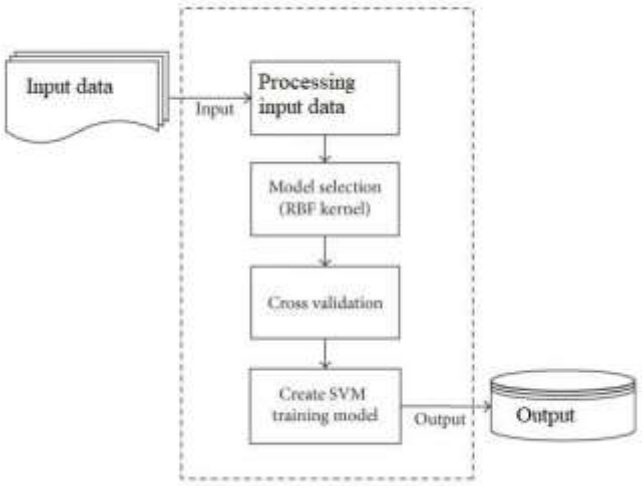
The present system shares the same objective but encompass different methodologies to arrive at a relatively less accurate conclusion. The qualitative superiority that these methods have over one another is dependent on the accuracy of the results produced. There are different aspects of the data that are used in order to parametrically come to a definite conclusion over the prediction of liver disease. Fuzzy logic has been developed for the classification of patients with liver cirrhosis. In gastroenterology, the Child-Pugh score is used to assess the prognosis of chronic liver disease, mainly cirrhosis. It was originally made to predict the mortality during surgery.

It is now used to determine the prognosis, the required strength of treatment and the necessity of liver transplantation.

The following score is instrumental in utilizing the five clinical measures of liver disease and each of these measures are scored between 1 and 3, with 3 indicating a serious condition of organ deterioration.

4.3 Proposed System

Machine learning is understandably one of the most extensively utilized paradigms of big data management where a significantly high set of distinct raw data can be collated effectively to make appropriate inferences and eventually to come up with a usual collection of contextually useful collection of integrative information. With the onset of the exponential technological explosion in the field of medicine, there is a felt need to handle a colossal set of data, thereby managing and utilizing the same to make effective and informative inferences for the doctors and patients.



The proposed system is build expert system which employ Fuzzy C-Means for the diagnosis of LDs is developed in an environment characterized by Microsoft Window XP professional Operating System, Microsoft Access Database Management system, Visual Basic Application Language and Microsoft Excel. Neuroph and Crystal reports were used for neural network analysis and graphical representation. An approach for analyzing clusters to identify meaningful pattern for determining whether a patient suffers from LD or not is presented. The system provides a guide for diagnosis of LDs within the decision making framework. The process for the medical diagnosis of LD starts when an individual consults a physician (doctor) and presents a set of complaints (symptoms). The physician then requests further information that will further aid in the proper diagnosis of the disease.

4.4 Advantages of the Proposed System

Considering the certain differences that have been adopted in the current system the following are the distinct advantages that are observed:

- **The performance classification of liver-based diseases is further improved:** with the far deepened understanding of the different kinds of ailments in the field of medicine, the different set of parameters to distinctly determine the kind of liver disease and its occurrence has become a far less complicated task. With advancements in data mining paradigms and software architectures like Hive, R, easing up the data collection process, the preprocessing and evaluation stages are given more attention to.
- **Time complexity and accuracy can be measured by various machine learning models, so that we can measures different parameters, owing to the needs of the user:** Every prediction system is based on the kind of parameters that it is expected to accept, compare and then finally come to a predictive conclusion. Accordingly, there are different algorithms that are used to model the predictive system to suit the context. The different machine learning algorithms judge the kind of disease and the testing parameters.
- **Different machine learning having high accuracy of the result:** In comparison to other methodologies considered, the right machine learning algorithm can aptly increase the efficiency of the results that are expected out of the predictive system.
- **Risk factors can be predicted early by machine learning models:** The machine learning algorithms predict the risk factors through simple methodologies of analysis the inconsistencies in the collective training data set and their respective parameters.

4.5 Advantages of Machine Learning Algorithms

Machine learning is a functionality of a system to be able to learn through the extensive usage of examples that pose a set of conditions that can be incorporated as a part of the self improvement process without being coded by a programmer. The result, thus obtained is then used by the corporate, in order to make actionable inferences for decision making. It has its roots related to data mining and close association with Bayesian predictive modelling. The data is taken as an input by the machine and the result is formulated as the output. Typical machine learning algorithms are utilized in trying to improve the user experience by providing recommendations using historical data. This would be an opportunistic approach to utilise this unsupervised learning to do the same.

4.6 Machine Learning vs. Traditional Programming

In traditional programming paradigm, the programmer is required to analyse, study and code all the rule subordinations in accordance with the experts and their recommendations on an advisory capacity. These rules act as the logical foundation for the machine. When the system grows, there is a rising need in the complexity of these systems and the need to incorporate more and more rules. This can get too haphazard to maintain. Machine learning replaces the conventional paradigms under these circumstances. The learning systems in the machine learning paradigm are centric on enabling the system to derive theses functional rules in order to inferentially use a set of example patterns to derive those rules and build a solid logical foundation in a system.

4.8 Inferences

The functionality of the proposed system has to be tested for the kind of limitations that could put up constraints on the operations of the system. The powerfulness of the system is tested by exploring the limits using data that the system has never been acclimated to or the kind of data that is unexplored at every level. The new data that is incorporated into the system, is incorporated and transformed into a features vector, go through the model and then conclusively come up with a prediction.

The life of Machine Learning programs is straightforward and can be summarized in the following points:

- Define a question
- Collect data
- Visualize data
- Train algorithm
- Test the Algorithm
- Collect feedback
- Refine the algorithm
- Loop 4-7 until the results are satisfying
- Use the model to make a prediction

4.9 Supervised learning

A calculation utilizes preparing information and criticism from people to get familiar with the relationship of offered contributions to a given yield. For example, an expert can utilize showcasing cost and climate gauge as information to foresee the offers of jars. You can utilize administered realizing when the

yield information is known. The calculation will foresee new information. There are two classifications of regulated learning:

- **Classification task:** Imagine you need to foresee the sexual orientation of a client for a business. You will begin gathering information on the tallness, weight, work, pay, obtaining crate, and so forth from your client database. You know the sexual orientation of every one of your client; it must be male or female. The target of the classifier will be to dole out a likelihood of being a male or a female (i.e., the name) in light of the data (i.e., highlights you have gathered). At the point when the model figured out how to perceive male or female, you can utilize new information to make an expectation. For example, you just got new data from an obscure client, and you need to know whether it is a male or female. On the off chance that the classifier predicts male = 70%, it implies the calculation is certain at 70% that this client is a male, and 30% it is a female. The mark can be of at least two classes. The above precedent has just two classes, yet on the off chance that a classifier needs to anticipate object, it has many classes (e.g., glass, table, shoes, and so forth each article speaks to a class)
- **Regression task:** When the yield is persistent esteem, the assignment is a relapse. For example, a money-related investigator may need to gauge the estimation of a stock dependent on a scope of highlights like value, past stock exhibitions, and macroeconomics record. The framework will be prepared to gauge the cost of the stocks with the least conceivable blunder.

4.10 Python IDLE Follow these instructions to write and run a simple Python program using the IDLE editor:

1. Start IDLE (see screen above). You will then see a window entitled "Python Shell"

Fig: Python Shell window

2. From the Python Shell window, select New Window from the File menu.

3. You will see a window entitled "Untitled"

Fig: Untitled Window

4. From the File menu, select Save As, and select a folder to save your Python program file.

8. The following program statement will run under Python 2.x or Python 3.0

Type in the following text into this window (make sure the word print is all in lower case):

Print ("Hello World")

If you're running Python 2.x, the text will automatically change colour to look like this

Print ("Hello World")

If you're running Python 3.0.x, the text will automatically change colour to look like this:

Print ("Hello World")

9. To run this program, select Run Module from the Run menu. You should see a reminder to save the Source (your program).

Click on OK to save.

Then you will see your program running in a Python Shell window

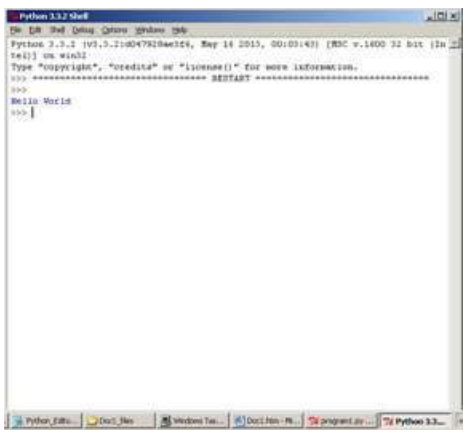


Fig: Python Shell window

Close all Python windows to quit Python.

5. RESULTS

5.1 Front-End

6. CONCLUSION

In this paper the problem of summarizing the different algorithm of data mining is used in the field of medical prediction are discussed. The main focus is on using different algorithm and combination of several targets attributes for different types of disease prediction using data mining. A foremost class of problems in medical science absorbs the diagnosis of disease, based upon an assortment of tests carried out upon the patient. When several tests are involved, the ultimate diagnosis may be difficult to obtain, yet for a medical expert. This has given rise, over the past few decades, to automated problem-solving tools, intended to assist the physician in making sense out of the welter of data. In healthcare, data mining is becoming increasingly more essential.

The selection of data mining approaches depends on the nature of the dataset if the dataset consist of the labelled features then the classification techniques can be suggested for best prediction. If the dataset is with unlabelled features then the clustering techniques are best suited for pattern recognition. If the optimization of the results needs to be improvised means then bio inspirational based techniques are best suited. Keeping in consideration with these existing problems this paper aims to survey the existing approaches in the field of medical sciences and the importance of data mining techniques used by various authors. The study reveals the importance of life threading disease should be diagnosed.

Firstly, a naive predictor and a benchmark model ('Logistic Regression') were run on the dataset to determine the benchmark value of accuracy. The greatest difficulty in the execution of this project was faced in two areas- determining the algorithms for training and choosing proper parameters for fine-tuning. Initially, I found it very vexing to decide upon 3 or 4 techniques out of the numerous options available in sklearn.

7. REFERENCES

[1] G. E. Sakr, I. H. Elhajj and H. A. Huijer, "Support vector machines to define and detect agitation transition," IEEE Transactions On Affective Computing, vol. 1, pp. 98-108, December 2010.

[2] M. Haitham, A. Angari and A. V. Sahakian, "Automated recognition of obstructive sleep apnea syndromeusing support vector machine classiifier," IEEE Transactions On Information Technology In Biomedicine,vol. 16, pp. 463-468, May 2012.

[3]D. Y. Tsai and S. Watanabe, "Method for optimization of fuzzy reasoning by genetic algorithms and is application to discrimination of myocardial heart disease," IEEE Nuclear Science Symposium and Medico1 Imaging Conference, pp. 2239-2246, December 1966.

[4] A. M. Anbarasi and N. Iyengar, "Enhanced prediction of heart disease with feature subset selection using genetic algorithm," International Journal of Engineering Science and Technology, vol. 2, pp. 5370-5376, November 2010.

[5] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," IEEE Intelligent Systems, pp. 44-49, March 1998.

[6] C. L. Huang and C. Jawing, "A ga-based feature selection and parameters optimization for support vector machines," Expert Systems with applications, vol. 31, pp. 231-240, October 2006.

[7] J. Z. H. Yan and C. Xiao, "Selecting critical clinical features for heart diseases diagnosis with a real coded genetic algorithm," Applied Soft Computing, vol. 8, pp. 1105-1111, March 2008.

[8] A. Rajkumar and G. S. Reena, "Diagnosis of heart disease using datamining algorithm," Global Journal of Computer Science and Technology, vol. 10, pp. 38-43, December 2010.

[9]S. Palaniappan and R. Awang, "Intelligent heart disease prediction system using data mining techniques,"International Journal of Computer Science and Network Security, pp. 343-350, January 2008.

[10]. W. G. Baxt, "Application of artificial neural networks to clinical medicine,"Lancet, vol.346, pp. 1135-1138, October 1995