IRSTI 20.23.19

E.N. Myrzagulova[1]
[1]Suleyman Demirel University
Kaskelen, Kazakhstan

COMPARATIVE ANALYSIS OF CIRRHOSIS PATIENTS' MEDICAL DATA BY USING SUPERVISED MACHINE LEARNING ALGORITHMS

Abstract. The number of patients who have liver disease (cirrhosis) have been continuously increasing in the last decades because of excessive consumption of alcohol, inhale of harmful gases, intake of contaminated food, pickles and drugs. By this time, computer intelligence allows physicians to significantly improve the accuracy of diagnoses, create super-effective drugs and significantly alleviate the course of various diseases of patients. Automatic classification tools may reduce burden on doctors. This paper evaluates the selected classification algorithms for the classification of some liver patient datasets. There were several algorithms that were used in this research work such as Logistic Regression, Naïve Bayes classifier, k-nearest neighbors, Neural Network algorithm and Random Forest. These algorithms are evaluated based on three criteria: Accuracy, Precision, and Recall.

Nowadays computer-aided diagnosis plays an important role in medicine and has no difference from the diagnosis of professional doctors. The aim of the research project is to work out an optimized algorithm for the detection Cirrhosis of the liver by using supervised machine learning algorithms, which can help to reduce costs and human resources. So the main purpose is to show that less complex algorithms can be used in order to establish a diagnosis of different cancer diseases and implement them.

Key words: Cirrhosis of the liver, Supervised Learning, Gaussian Naive Bayes, Correlation analysis, Performance Analysis, Neural networks, Random Forest, Logistic regression.

***

Аңдатпа. Соңғы онжылдықта ішімдіктің шамадан тыс қолдануынан, көмірқышқыл газының таралуынан, зиянды тағам мен есірткіні тұтыну нәтижесінде бауыр ауруы – циррозбен ауыратын науқастар саны өсуде. Қазіргі таңда жасанды интеллект науқастарға дәл диагноз қойып, тиімді дәрілерді құрастыра отырып медицина қызметкерлеріне үлкен септігін тигізуде. Жіктеу тәсілінің автоматтық құралдары дәрігерлерге жүгінетін жұмыстарды жеңілдетпек. Бұл зерттеу жұмысында цирроз ауруымен ауыратын науқастардың деректері

қолданыла отырып машиналық оқытудың алгоритмдерінің салыстырмалы анализі жасалынады. Осы ғылыми мақалада Logistic Regression, Naive Bayes classifier, K-nearest neighbors, Neural Network algorithm және Random Forest сынды алгоритмдер талданады. Бұл алгоритмдер дәлдік, тереңдік және толықтық сынды өлшемдер бойынша бағаланады.Осы зерттеу жұмысының негізгі мақсаты – бауыр циррозын анықтау мақсатында оңтайландырылған алгоритмді жасап шығару.

Түйінді сөздер: бауыр циррозы, қадағалауға үйрену, өнімділік талдау, клиникалық шешімдерді қолдау жүйелері, шешім ағаштар, логистикалық регрессия.

***

Аннотация. В последние десятилетия число пациентов с заболеванием печени (циррозом) постоянно растет из-за чрезмерного потребления алкоголя, вдыхания вредных газов, потребления зараженной пищи и наркотиков. В настоящее время искусственный интелект позволяет врачам значительно улучшить точность диагнозов, создать сверхэффективные лекарства и значительно облегчить ход различных заболеваний пациентов. Автоматические инструменты классификации могут снизить нагрузку на врачей. В данной исследовательской работе представлен сравнительный анализ медицинских данных пациентов с циррозом с использованием контролируемых алгоритмов машинного обучения. Существует несколько алгоритмов, которые использовались в этой исследовательской работе такие как логистическая регрессия, классификатор Наивного Байеса, алгоритм ближайшего соседа, алгоритм нейронной сети и случайный лес. Эти алгоритмы оцениваются по трем критериям: точность, глубина и полнота. Основная цель этой исследовательской работы – разработать оптимизированный алгоритм выявления цирроза печени.

Ключевые слова: цирроз печени, контролируемое обучение, анализ эффективности, системы поддержки клинических решений, деревья принятия решений, логистическая регрессия.

Introduction

It is not a secret that in this new era of science and technology, artificial intelligence is becoming an indispensable and essential element in the healthcare industry. Nowadays computer-aided diagnosis plays an important role in medicine and has no difference from the diagnosis of professional doctors. The aim of the research project is to provide an example of how humanity can bind the rapidly expanding knowledge base of medicine with the doctor treating each patient as an individual with faster diagnosis by using new algorithms as the task of ascertaining a correct diagnosis of a disease is one of the most essential problems in medicine. [1,page 21] The solution is machine

learning which one of the major branches of Computer Science is. It provides several fundamental tools for intelligent data analysis. This research work contains machine learning algorithms that were designed and used to analyze medical data sets that will help to classify diseased patients from normal [2,page 125].

Today's hospitals are well equipped with data collection devices and data is gathered and shared in large information systems. Many artificial intelligence algorithms like genetic algorithms, fuzzy logic algorithms, Bayes-methods and many data mining tools like scikit-learn and neural network are used to diagnose the diseased data [3,page 11]. Some liver disease patients show symptoms of viral infection such as tiredness, abdomen pain, muscle pain, and vomiting, and loss of appetite. But few symptoms including swelling of the abdomen as well as limbs, jaundice, digestive bleedings etc., may be the advance cases of liver failure patients. The common attributes of liver data to find the liver status are SGOT, SGPT, ALP, Total Bilirubin, Direct Bilirubin, Total Proteins and Albumin. Algorithms and methods are applied on Medical Data in tests like LFT (Liver Function Test) to diagnose Hepatitis or liver disorder or jaundice. These patients need uninterrupted and well-coordinated medical treatment to diminish the death rate.

Materials

A. Data Collection

The data set was obtained from the UCI Repository of Machine Learning Databases [1]. This data set contains 416 liver patient records and 167 non liver patient records. The data set was collected from north east of Andhra Pradesh, India. Selector is a class label used to divide into groups(liver patient or not). This data set contains 441 male patient records and 142 female patient records. Any patient whose age exceeded 89 is listed as being of age «90».

1.Age of the patient
2.Gender of the patient
3.Total Bilirubin
4.Direct Bilirubin
5.Alkaline Phosphatase
6.Alanine Aminotransferase
7.Aspartate Aminotransferase
8.Total Proteins
9.Albumin
10.Albumin and Globulin Ratio
11.Selector: field used to split the data into two sets (patient with liver disease, or no disease)

B. Features Correlation

Methods

A. Logistic Regression

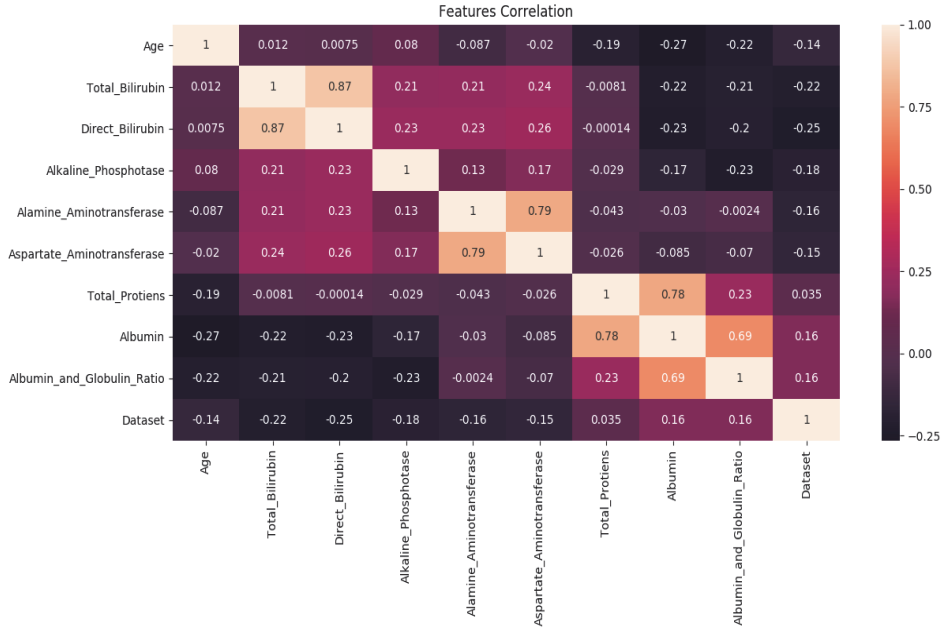

Fig. 1. Features Correlation

The logistic regression (LRM) has wide range of implications in medical research field. The LRM model is used for the classification of the attributes, which might help to classify the outcome. The distinctive feature of the model is that the outcome variable is dichotomous. The result is not bounded to a linear form. As a result, the created model can be used to classify a new provided data via placing them in a model for the probability P, the detailed information is provided in [4,page 66].

B. Multi-Layer Perceptron Neural Networks

Multi-layer perceptron neural networks (MLP) are processing devices, which closely resemble a model of the neuronal structure of the mammalian cerebral cortex. Large MLPs might have hundreds or thousands of processor units, whereas a mammalian brain has billions of neurons with a corresponding increase in magnitude of their overall interaction and emergent behavior. Generally the neural network has three components or layers. The first layer is called input layer, through which it gets data inside network on our case disease related attributes. The second layer is called hidden where all operations performed. The last layer called out where network make final decision regarding patient's condition.

C. k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression.In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression

D. Bayesian Classifiers are statistical classifiers based on Bayes theorem. Bayesian classification is very simple and it shows high accuracy and speed when applied to large data bases. It works on one assumption that is the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class conditional independence [6,page 231].

E. Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Evaluation of the classifier to measure the quality is commonly evaluated based on the data in the confusion matrix. Several standard measures have been defined for correct and incorrect classification results of the matrix. The most common practical measure to evaluate the performance is accuracy, which is defined as the proportion of the total number of instances that were classified correctly.

Recall is the mean proportion of actual positives which are correctly identified.

Precision is the mean proportion of positives which are relevant. F-measure is a harmonic mean of recall and precision.

IV. Simulation results

In this study, the dataset of patients with liver diseases, which containing 10 original features by using five machine learning methods, Gaussian Naive Bayes, K-Nearest Neighbors, Logistic Regression, Multi-Layer Perceptron and Random Forest used for classification. The performance metrics like accuracy, recall, precision and f-measure along with error metrics for all features has been performed using 10-fold cross-validation. The simulations were performed by using WEKA3.8 machine learning tool.

Table 1

Algorithms performance

algorithm	accuracy	precision	recall
GNB	0.57484611	0.94253663	0.432853717
KNN	0.65871308	0.7501016282	0.7813749001
LG	0.71699216	0.7532971101	0.9064053105
MLP	0.69293654	0.7682966792	0.9976019185
RF	0.67933218	0.7509854276	0.8726757724

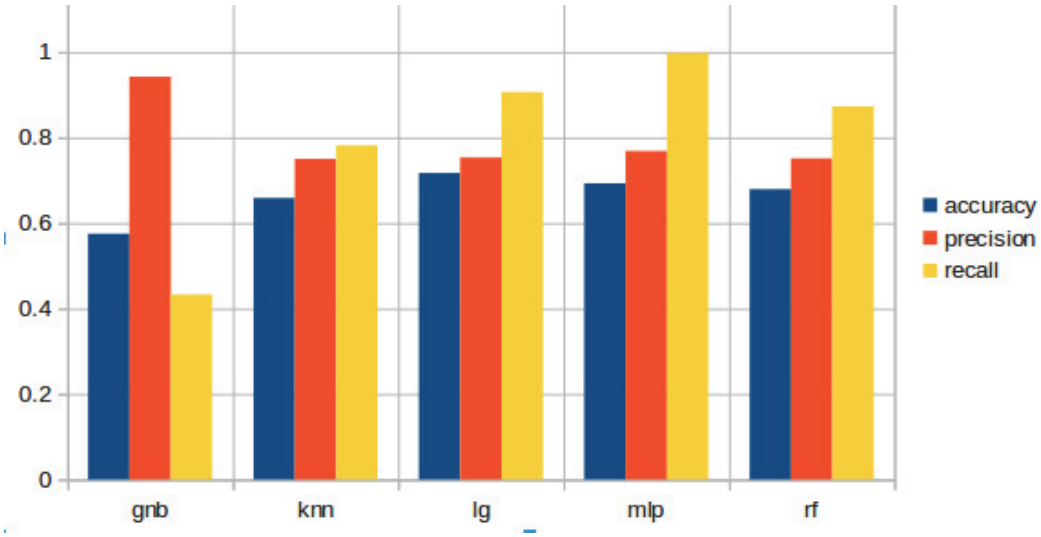

Fig 2. Results

V. Conclusion

In this study, popular Classification Algorithms were considered for evaluating their classification performance in terms of Accuracy, Precision, and Recall in classifying cirrhosis patients' dataset. The performance of Bayesian Classifier was the best for all performance and error metrics. As a result shows Gaussian Naive Bayes methods can be a good and practical choice to classify a medical data.

Physician's responsibility is to diagnose disorders of patients and treat them at the earliest. Any wrong diagnosis can put in danger a patient's life and may even cause his/her death. In this regard, the use of different methods of artificial intelligence and expert system is a promising strategy for the future. In this paper, an algorithm called separation of points by planes has been used to classify diseased liver patients from the healthier and helped in diagnosing the dangerous hepatitis or liver disorder. As a result shows Gaussian Naive Bayes methods can be a good and practical choice to classify a medical data.

References:

1 UCI Repository of Machine Learning Databases, University of California at Irvine, Department of Computer Science [Electron. resource]. - (date visited: 12.02.2018), 2013. – URL: https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset)

2 Ramana, B.V., Babu, M.S., Venkateswarlu, N.B. A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis //International Journal of Computer Science Issues – 2012. – Vol 9. № 2. – P. 506-516.

3 Ramana, B.V., Babu, M.S., Venkateswarlu, N.B. A Critical Study of Selected Classification Algorithms for Liver Disease Diagnosis, International Journal of Database Management Systems (IJDMS) – 2011. – Vol.3. – № 2. – P. 101-114

4 Norvig, P., Russell, S. Artificial Intelligence: A Modern Approach. – New Jersey: Prentice Hall, 2002. – 1152 p.

5 Dunham, M.H. Data mining: Introductory and advanced topics // M.H. Dunham. – London: Pearson Education, 2006. – 315 p.

6 Babu, M.S., Ramana, B.V., Kumar, B.R. New Automatic Diagnosis of Liver Status Using Bayesian Classification // International Conference on Intelligent Network and Computing (ICINC 2010), Kuala Lumpur 2010. – P. 3085-3088.