IJMSIT

International Journal of Multidisciplinary Studies and Innovative Technologies

Volume : 3
Number: 2
Year: 2019
Pages: 135-138

Performance Analysis of Machine Learning Algorithms and Feature Selection Methods on Hepatitis Disease

Ebru Aydindag Bayrak[1*], Pinar Kirci[2] and Tolga Ensari[3]

[1, 2] *Engineering Sciences, Istanbul University-Cerrahpaşa, Turkey*
[3] *Computer Engineering, Istanbul University-Cerrahpaşa, Turkey*
[*] *ebruaydindag@gmail.com*

Abstract – In this study, some machine learning classification techniques are applied on Hepatitis data set acquired from UCI Machine Learning Repository. Naïve Bayes Classifier, Logistic Regression and J48 Decision Tree are used as classification algorithms and they have been compared according to filter-based feature selection methods. For filter-based feature selection, Cfs Subset Eval, Info Gain Attribute Eval and Principal Components have been used and the performance of them is evaluated in terms of precision, recall, F-Measure and ROC Area. Among the all used classification algorithms, Naïve Bayes Classifier has higher classification accuracy on the Hepatitis data set than the others with applied and non-applied filter-based feature selection. Moreover, we declare that the best filter-based feature selection is Principal Components because of the highest classification accuracy obtained with for hepatitis patients.

Keywords – Hepatitis, Machine Learning, Feature Selection, Classification, Diagnosis.

I. INTRODUCTION

Hepatitis is the term used to means inflammation of the liver. One of several viruses most often cause hepatitis and it is often called viral hepatitis. The most common types of viral hepatitis are Hepatitis A, Hepatitis B, and Hepatitis C [1]. The most known 5 main hepatitis viruses are called to as types hepatitis A, B, C, D and E. Because of the causing outbreak and occurring lots of deaths, these 5 types of hepatitis worry all people. Specially, types B and C cause to chronic disease in 100 of millions of people and also, they are the most extensive reason of liver cirrhosis and cancer [2].

In this paper, three of the most popular machine learning techniques are applied on Hepatitis data set and the result of applied machine learning techniques are compared using some performance metrics. The rest of the paper is organized as follows: The section II describes the researches about machine learning algorithms on health data set. The section III explains used material and the fundamental point of used machine learning methods. The experimental results are explained and illustrates in Section IV. The paper ends in Section V with the concluding remarks.

II. RELATED WORKS

Different machine learning techniques and algorithms have been used for the analysis of several disease like Cancer, Parkinson's, Alzheimer, Diabetes and Hepatitis. Especially, recent survey about hepatitis have been introduced as following:

In [3], Artificial Immune System (AIS) based classification and normal classification techniques have been applied many data sets of different diseases (Breast Cancer, Ecoli, Hepatitis, Heart (Statlog) and Pima Indians Diabetes Datasets from UCI). For AIS classification there are 8 different algorithms as CLONALG, CSCA, AIRS1, AIRS2, AIRS2 Parallel, Immunos1, Immunos2 and Immunos99. Therefore, ZeroR, J48, KStar and Simple Cart are used as normal classifiers in this study. The results of experiments show that AIS based classifiers have the best performance with compared to all the classifiers used for healthcare data set.

In [4], the performance of 15 classification algorithms on Hepatitis data set obtained from UCI Repository, have been compared in Tanagra tool. Also, Fisher filtering, Relief filtering and Step Disc have been used as feature selection method and classification algorithms are applied on data set using with and without feature selection to see impact of feature selection. The authors have mentioned that the classification algorithms have better performance without feature selection.

In [5], Naive Bayes (NB), Naïve Bayes Updatable, Bayes Net, Random Forest, J48 and Multi-Layer Perceptron (MLP) classification algorithms have been used for analyzing Hepatitis dataset. The results of classification algorithms are determined according to time and accuracy values. Naive Bayes has been performed better accuracy and less time on detection of hepatitis.

In [6], various classification algorithms have been applied on different UCI medical data sets for comparing their performance in terms of accuracy, specificity, precision and ROC Area. The authors used Naïve Bayes, Multi-Layer Perceptron, JRip, J48, IBK and Bagging are selected classifier algorithms.

In [7], data mining techniques used for diagnosing Hepatitis disease have been reviewed and shown that the accuracy of applied data mining methods can be used decision making in the early diagnosis of hepatitis disease.

In [8], the Decision Tree (DT) algorithm has been used a machine learning method as prediction of Hepatitis disease. This analysis is performed with different attribute numbers. It has suggested that the attributes Bilirubin and Varices plays an important role in detection abnormalities with 85.81% accuracy value is obtained from the overall study.

In [9], the survey of machine learning algorithms for disease diagnostic has been provided a comparative analysis of different machine learning algorithms applied on diabetes, dengue, hepatitis, heart and liver diseases. It has been asserted that the feed forward neural network classifies hepatitis disease with 98 % accuracy value. Also, the positive and negative sides of these algorithms have been highlighted.

In [10], As to be Naïve Bayes, Naïve Bayes Updatable, Neural Network, J48, K-star, FT Tree and LMT different classification algorithms are used for analyzing Hepatitis prognostic data set which is obtained from UCI. Also, Neural Connection used for comparing the performance of data mining algorithms to diagnose hepatitis disease. It is shown that based on experimental results, Naïve Bayes algorithm has higher accuracy than others.

In [11], biomedical data sets obtained from UCI Machine Learning repository were classified with C4.5 algorithm. Heart, hepatitis, promoters and splice data sets were subjected to preprocessing and applied in [11]. The authors state that it is very important which preprocessing of the data sets used when comparing different data mining applications.

In [12], the performance of 6 different classification algorithms on the 10 clinical data sets acquired UCI database (breast cancer, hepatitis, thyroid and others) were analyzed. The authors observed that there was not a single classifier method that showed the highest performance in all data sets, but Multi-Layer Perceptron and Naive Bayes methods were relatively successful.

For further studies, the authors state a new idea on machine learning task [17]. They declare “Digital Data Forgetting” and “Big Cleaning” concepts to get rid of unimportant data parts of the data set. Because, we are in digital life and computers store more data (images, cameras, phones, user experiences, medical data and others) day by day in our life. So, it will be hard to store all of the data forever. Moreover, it will be hard to process and get results from these massive data sets. Therefore, big data may be cleaned not only for bio-informatics applications but also for other applications like computer vision, robotics, financial in the future studies.

III. Materials and Method

A. Data Set

In this study, Hepatitis data set is analyzed and acquired from University of Irvine California (UCI) Machine Learning Repository [13]. This data set consists of 155 samples (number of instances) are classified live or die. Also, Hepatitis data set has 19 attributes and their characteristics are categorical, integer and real values.part1 [13].

In preprocessing phase, data cleaning, data aggregation, data transformation, data reduction data are made available for analysis. These operations may affect the performance of the model. Transactions depend on the practitioner's perspective. Some different interference on the data set can have different results in different algorithms [14].

B. Feature Selection

For feature selection part, we use Cfs Subset Eval, Info Gain Attribute Eval and Principal Components. CfsSubsetEval evaluates the value of a subset of attributes, according to individual predictive ability and the degree of redundancy. Also, it uses BestFirst as search method. InfoGainAttributeEval is other attribute evaluator and it evaluates the value of attributes by calculating information gain according to the class, and it uses Ranker as search method [15]. Principal Components performs a principal components analysis and transformation of the data. It is used with Ranker search and it reduce dimension of data by choosing enough eigenvectors to provide for some 0.95 (default value) percentage of the variance in the original data [16].

Table 1. The attributes of Hepatitis data set.

Number of attributes	Attribute	Values
1	Class	Die, Live
2	Age	10,20,30,40,50,60,70,80
3	Sex	Male, Female
4	Steroid	No, Yes
5	Antiviral	No, Yes
6	Fatigue	No, Yes
7	Malaise	No, Yes
8	Anorexia	No, Yes
9	Liver Big	No, Yes
10	Liver Firm	No, Yes
11	Spleen Palpable	No, Yes
12	Spiders	No, Yes
13	Ascites	No, Yes
14	Bilirubin	0.39,0.80,1.20,2.00,3.00,4.00
15	Alk Phosphate	33,80,120,160,200,250
16	Sgot	13,100,200,300,400,500
17	Albumin	2.1,3.0,3.8,4.5,5.0,6.0
18	Protime	10,20,30,40,50,60,70,80,90
19	Histology	No, Yes

C. Classification Methods

In this paper, we applied Naïve Bayes, Logistic Regression and J48 Decision Tree machine learning classification algorithms to classification of Hepatitis data set to learn performance of machine learning methods (Table 2). We also applied k=10 cross validation for the testing parts. For k=10 cross validation, the data set is divided into 10 parts and each part is used once as a testing set and the remaining 9 parts as a training set.

Naïve Bayes Classifier

Naive Bayes Classifier is one of the most popular machine learning classification technique that based on the Bayesian theorem. It is simple method, such that in small quantities training data can be classify the given examples. The Naive Bayes Classifier algorithm which is used to calculate present and past frequency occurrences, can explain as Equation (1) [18]:

$$P(A \setminus B) = P(B \setminus A)*P(A)/ P(B) \quad (1)$$

P(A): Prior probability of A. It counts only the occurrences of A.
P(A \ B): Posterior probability of A, given B.
P(B \ A): Posterior probability of B, given A.
P(B): Prior probability of B.

Logistic Regression

Logistic Regression provides the relationship between the independent variables and the dependent variable if the dependent variable is a categorical. According to characteristic of dependent variable; it can be named binary, ordinal and multi nominal logistic regression [19]. Logistic regression resembles in many respects to linear regression however, they diverge in critical respect. In Logistic regression, outcome can be explained through a special mathematical transformation and weighted sum that is called logit. This transformation allows all weighted sum to be correspond a value in between 0 and 1 [20].

J48 Decision Tree

Classification with decision trees is basically based on the creation of a decision tree, the application of each instance in the database to that tree and the classification of the instance according to the result. Compared with other classification methods, decision trees are easier to generate and understand. There are many algorithms developed based on decision trees and these algorithms differ in terms of the selection of root, node and branching criteria [21]. J48 algorithm has been used as decision tree method in this study.

Table 2. Comparative analysis of machine learning methods on Hepatitis data set.

Method	Used in our study
Statistical Classifier	Naïve Bayes Classifier
Regression	Logistic Regression
Decision Tree	J48

IV. EXPERIMENTAL RESULTS AND DISCUSSION

In this study, we used Naïve Bayes, Logistic Regression and J48 Decision Tree classification algorithms to classification of Hepatitis data set to measure the performance of these algorithms. The used feature selection methods are Cfs Subset Eval and Info Gain Attribute Eval and Principal Components.

Firstly, Naïve Bayes, Logistic Regression and J48 Decision Tree algorithm were applied Hepatitis data set without feature selection. Secondly, same experiments were repeated for three feature selection methods. 11, 17 and 20 features were selected respectively Cfs Subset Eval, Principal Components and Info Gain Attribute Eval feature selection methods.

The performance of applied machine learning classification algorithms was demonstrated in Table 3. The classification accuracy is the performance metric with k=10 cross validation for compared algorithms. The classification algorithms are compared by applied and non-applied feature selection methods. According to Table 3, the highest accuracy was obtained for Hepatitis data set with Naïve Bayes classification algorithm with Principal Components feature selection approach. Also, evaluation of feature selection methods was shown Table 4 in terms of precision, recall, F-Measure and ROC Area metrics.

Table 3. Comparison of classification accuracy for machine learning classification algorithms.

	Without Feature Selection	**With Feature Selection**		
Classification Algorithms		Cfs Subset Eval	Info Gain Attribute Eval	Principal Component
Naïve Bayes	84.51 %	87.74 %	84.51 %	88.38 %
Logistic Regression	82,58 %	86.45 %	82.58 %	83.87 %
J48 Decision Tree	83.87 %	81.29 %	83.87 %	81.94 %

Table 4. Evaluation of applied feature selection methods for Hepatitis data set.

Classification Algorithm	**Feature Selection Method**	**Precision**	**Recall**	**F-Measure**	**ROC Area**
Naive Bayes	Cfs Subset Eval	0.87	0.87	0.87	0.89
	Principal Components	0.88	0.88	0.88	0.88
	Info Gain Attribute Eval	0.85	0.84	0.84	0.86
Logistic Regression	Cfs Subset Eval	0.86	0.86	0.86	0.89
	Principal Components	0.82	0.83	0.83	0.82
	Info Gain Attribute Eval	0.81	0.82	0.81	0.80
J48 Decision Tree	Cfs Subset Eval	0.79	0.81	0.79	0.69
	Principal Components	0.81	0.81	0.81	0.69
	Info Gain Attribute Eval	0.82	0.83	0.82	0.70

V. CONCLUSION

Hepatitis is the most common disease for all people in the world. Therefore, any improvement for early diagnosis and prediction liver disease is more important for a healthy life. In this paper, we discussed most popular machine learning classification methods for Hepatitis disease classification. Naïve Bayes Classifier, Logistic Regression and J48 Decision Tree classification methods were used as machine learning

classifier. A comparative analysis was performed on the basis of filter-based feature selection algorithms to classify hepatitis disease. The evaluation of applied feature selection methods was compared based on performance metrics such as precision, recall, F-Measure and ROC Area.

When we do not apply feature selection method, Naïve Bayes Classifier shows the highest accuracy value % 84.51 compared other classification algorithms. After the application of feature selection methods, in terms of classification accuracy Naive Bayes Classifier with Principal Components feature selection method shows best performance than the others. Logistic Regression shows the best performance with Cfs Subset Eval feature selection method. J48 Decision Tree shows the best performance with Info Gain Attribute Eval and non-applying feature selection methods.

The results of this study will make contributions in the diagnosis of hepatitis disease with several machine learning classification algorithms and feature selection methods.

REFERENCES

[1] U.S. Food and Drug Administration Homepage, [Online]. Available: https://www.fda.gov/patients/get-illnesscondition-information/hepatitis-b-c

[2] World Health Organization Homepage, [Online]. Available: https://www.who.int/features/qa/76/en/

[3] R. K. Das, M. Panda, N. Mahapatra, and S. S. Dash, "Application of Artificial Immune System Algorithms on Healthcare Data", in 2017 International Conference on Computational Intelligence and Networks, 2017, pp. 110-114.

[4] P. Nancy, V. Sudha, and R. Akiladevi, "Analysis of feature Selection and Classification algorithms on Hepatitis Data", *International Journal of Advanced Research in Computer Engineering & Technology*, Volume 6, Issue 1, 2017.

[5] T. Karthikeyan, and P. Thangaraju, "Analysis of Classification Algorithms Applied to Hepatitis Patients", *International Journal of Computer Applications*, 62(15), 2013.

[6] B. V. Ramana, and R. S. K Boddu, "Performance Comparison of Classification Algorithms on Medical Datasets", In 2019 IEEE 9th Annual Computing and Communication Workshop and Conference, 2019, pp. 140-145.

[7] S. O. Hussien, S. S. Elkhatem, N. Osman, and A. O. Ibrahim, "A Review of Data Mining Techniques for Diagnosing Hepatitis", in 2017 Sudan Conference on Computer Science and Information Technology, 2017, pp. 1-6.

[8] V. Shankar sowmien, V. Sugumaran, C. P. Kartikeyan, and T. R. Vijayaram, "Diagnosis of Hepatitis Using Decision Tree Algorithm", *International Journal of Engineering and Technology*, Vol 8, pp. 1411-1419, 2016.

[9] M. Fatima, and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic", *Journal of Intelligent Learning Systems and Applications*, 9(01), 1, 2017.

[10] F. M. Ba-Alwi, and H. M. Hintaya, Comparative Study for Analysis the Prognostic in Hepatitis Data: Data Mining Approach, *International Journal of Scientific & Engineering Research*, Vol 4, Issue 8, August-2013.

[11] Ö. Yıldız, T. Dayanan, and İ. Düzdar Arfun, "Comparison of Accuracy Values of Biomedical Data with Different Applications Decision Tree Method", in 2018 Electric Electronics, Computer Science, Biomedical Engineering's Meeting, 2018, pp. 1-4.

[12] E. Seğmen, and A. Uyar, "Performance Analysis of Classification Models for Medical Diagnostic Decision Support Systems", Signal Processing and Communications Applications Conference, 2013, pp. 1-4.

[13] UCI Homepage, [Online]. Available: https://archive.ics.uci.edu/ml/datasets/hepatitis

[14] C. Coşkun, and A. Baykal, Veri Madenciliğinde Sınıflandırma Algoritmalarının bir Ornek Uzerinde Karşılaştırılması, *Akademik Bilişim*, 2011, 1-8.

[15] C. Luan, and G. Dong, "Experimental Identification of Hard Data Sets for Classification and Feature Selection Methods with Insights on Method Selection", *Data and Knowledge Engineering*, Vol 118, 41-51, 2018.

[16] S. Priya, and R. Manavalan "Optimum Parameters Selection Using ACOR Algorithm to Improve the Classification Performance of Weighted Extreme Learning Machine for Hepatitis Disease Data Set", IEEE International Conference on Inventive Research in Computing Applications, 2018, pp. 986-991.

[17] M. Gunay, E. Yildiz, Y. Nalcakan, B. Asiroglu, A. Zencirli, and T. Ensari, "Digital Data Forgetting: A Machine Learning Approach", IEEE International Symposium on Multidisciplinary Studies and Innovative Technologies, 2018, pp. 1-4.

[18] E. Aydindag Bayrak, and P. Kirci, *Intelligent Big Data Analytics in Health. In Early Detection of Neurological Disorders Using Machine Learning Systems*, pp. 252-291, IGI Global, 2019.

[19] E. Karabulut, and R. Alpar, *Lojistik Regresyon, Uygulamalı Çok Değişkenli İstatistiksel Yöntemler*, Detay Yayıncılık Ankara, ISBN: 978-605-5437-42-8, 2011.

[20] C. Yoo, L. Ramirez, and J. Liuzzi, Big Data Analysis Using Modern Statistical and Machine Learning Methods in Medicine, *International Neurourology Journal*, 18 (2), 50, 2014.

[21] P. Tapkan, L. Özbakır, and A. Baykasoğlu, "Weka ile Veri Madenciliği Süreci ve Örnek Uygulama", Endüstri Mühendisliği Yazılımları ve Uygulamaları Kongresi, 2011.