Comparative Study of Heart Disease Classification

Simge EKİZ
Düzce University, Engineering Faculty,
Computer Engineering Department
Düzce, Turkey
simgeekiz@duzce.edu.tr

Pakize ERDOĞMUŞ
Düzce University, Engineering Faculty,
Computer Engineering Department,
Düzce, Turkey
pakizeerdogmus@duzce.edu.tr

Abstract— The aim of this paper is to compare two important machine learning platform results for the same dataset. With this aim, we conducted an experiment to classify heart disease both in Matlab© environment and WEKA©, by using six different algorithms. Linear SVM, Quadratic SVM, Cubic SVM, Medium Gaussian SVM, Decision Tree and Ensemble Subspace Discriminant machine learning approaches are used for classifying the heart disease.

Keywords—heart disease; UCI Machine Learning; Classification; Support Vector Machines; Ensemble Learning; Decision Tree

I. INTRODUCTION

Heart disease, also known as coronary artery disease, indicates any pernicious conditions that affect the heart[1]. Heart disease has the largest proportion of deaths in the world. In 2012, approximately 17.5 million people died from heart disease, meaning that it consists of the 31% of all global deaths[2]. Moreover, heart disease death toll rises every year. It is expected to grow more than 23.6 million by 2030[3].

Many lives could be saved by diagnosing on time. Thus, diagnosing the heart disease is very important and it should be performed precisely [4]. However, it is a highly complicated task. Automating this process would help to overcome the issues with the diagnosis [5]. To achieve this, researchers proposed many tools and methodologies.

This paper is organized as follows: In section 2, we present a brief overview of related work. In section 3, we give information about the dataset and we explain the theoretic background of the data mining algorithms. Then our methods are described in section 4. Experimental results are given in section 5. Finally, we conclude our research in the last section.

II. RELATED WORKS

There is a plenty of prior work on classifying heart diseases. One of the early studies in the field, 77.0% accuracy was obtained by using logistic regression algorithm [6]. In the same year, the accuracy score was reported as 78.9% in another study where the author made use of CLASSIT conceptual clustering system [7]. It was observed that, accuracy scores of 81.11% and 81.48% could be reached, by using C4.5 and Naïve Bayes algorithms, respectively [8]. Likewise, a combined system of Generalized Discriminant Analysis and Least Square Support Vector Machine obtained 82.05% accuracy score [9]. More recently, Das et al. used an ensemble of neural networks and managed to get as high as 89.01% accuracy [1]. All the above mentioned studies were performed on heart disease dataset which was created and maintained by University of California Irvine (UCI) [10].

III. STUDY ENVIRONMENT AND BACKGROUND THEORIES

A. Dataset Description

We use the Hungarian heart disease dataset from UCI machine learning repository [10]. The dataset contains 294 samples. Although dataset has 76 raw attributes, it is decided that only 14 of them are actually useful. The fourteenth field indicates the heart disease in the patient. 0 represents absence of disease and 1 to 4 numbers represents the presence of heart disease. The complete list of the features is listed in Table 1.

Table 1: Description of attributes

1. age	
2. sex	
3. cp	chest pain type (four values)
4. trestbps	resting blood pressure
5. chol	serum cholesterol in mg/dl
6. fbs	fasting blood sugar > 120 mg/dl
7. restecg	resting electrocardiographic results (values 0, 1 and 2)
8. thalach	maximum heart rate achieved
9. exang	exercise induced angina (1 = yes; 0 = no)
10. oldpeak	ST depression induced by exercise relative to rest
11. slope	the slope of the peak exercise ST segment
12. ca	number of major vessels (0-3) colored by flourosopy
13. thal	3 = normal; 6 = fixed defect; 7 = reversable defect
14. num (the predicted attribute)	0 = absence; 1-4 = presence of heart disease

978-1-5386-0440-3/17/$31.00 ©2017 IEEE

B. Theoretical Background

1) Decision tree

Decision tree follows a tree metaphor. Thus, it has root and leaves. Every path from the root to leaf will form a classification rule. Basing in that rule, data choose which path to go and the last leaf of a tree shows the class of that data. For building a decision tree, training data are needed [11]. After the tree is built, test data that we want to handle can be classified.

2) Support Vector Machine (SVM)

SVM is developed to classify the data set which contains separable classes. It performs classification by finding the optimum hyper plane that maximizes the margin between classes [12].

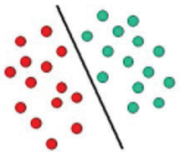

Fig. 1. Linearly Separable Dataset

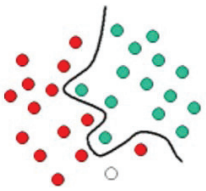

Fig. 2. Linearly Not Separable Dataset

Figure 1 shows classification of the Linear SVM which separates data into their respective groups with a line. Originally SVMs classify the data with linear separator but in most cases like in Figure 2, data cannot be separated with linearly. So in order to make optimal separation, SVM needs more complex structure. To overcome this problem kernel functions are developed. Kernel functions are used to map nonlinearly separable classes of data into a higher dimension so the data can be linearly separable in this high dimensional space [13].

Equation (1) defines the kernel function:

$$K(X_i, X_j) = \phi(X_i).\phi(X_j) \quad (1)$$

There are four types of kernels that can be used in Support Vector Machines models. These include linear, polynomial, radial basis function (RBF) and sigmoid [9]. And the equations have been indicated respectively.

$$K(X_i, X_j) = X_i, X_j \qquad Linear \quad (2)$$

$$K(X_i, X_j) = (\gamma X_i.X_j + C)^d \qquad Polynomial \quad (3)$$

$$K(X_i, X_j) = \exp(-\gamma |X_i - X_j|^2) \qquad RBF \quad (4)$$

$$K(X_i, X_j) = \tanh(\gamma X_i.X_j + C) \qquad Sigmoid \quad (5)$$

Each object in the input space which is in a low dimension transform in a higher dimensional space with the help of kernel functions [14]. Linear SVM uses the linear formula, Quadratic SVM and Cubic SVM uses the Polynomial formula respectively with degree two and degree three. Medium Gaussian SVM uses the RBF formula.

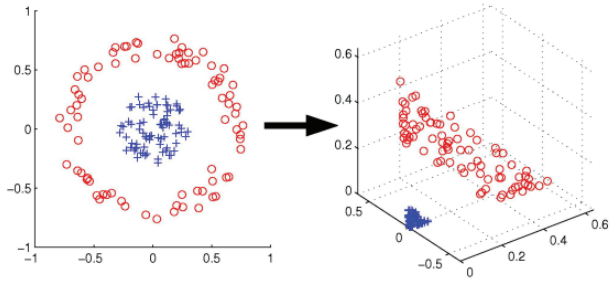

Fig. 3. Transforming the data can make it linearly separable

Figure 3 shows that SVM rearrange the objects and does some complex data transformations with the kernel functions [15].

3) Ensemble Learning

Ensemble Learning is a combination technique that combines multiple machine learning algorithms to obtain better prediction results. Ensemble Subspace Discriminant creates an ensemble of discriminant classifiers using the Random Subspace method. Moreover, random subspace method ensembles the classifiers by randomly distributing the features to its sub-classifiers. [16]. Thus, this method helps to avoid the problems of high dimensionality.

IV. METHOD

To identify the heart disease we use data mining techniques. We train various supervised machine learning algorithms on the UCI heart disease dataset. The simulations are conducted using Hungarian reprocessed dataset consisting of 294 instances and 14 features. To implement the algorithms we make use of WEKA, Matlab Classification Learner, and Neural Network Pattern Recognition Matlab application. In our experiment we use 10 fold cross-validation that divides the whole dataset into 10 groups. One group is used for testing and the rest of them are used for training and this process repeats 10 times.

First of all, we convert the data into a 294x14 dimensional matrix. We apply Linear SVM, Quadratic SVM, Cubic SVM, Medium Gaussian SVM, Complex Decision Tree algorithms. Additionally, we use an ensemble method called Subspace Discriminant Learner. The algorithms are implemented by using Matlab Classification Learner with 10 fold cross-validation.

Next, we convert our data to arff format by adding the data formats of the features to be able to implement the same algorithms in WEKA environment. In the WEKA environment, the names of the approaches may differ than the

ones in the other platforms. With the 10 fold cross-validation rate, we choose the J48 algorithm for implementing decision tree approach. We use LibSVM to apply Support Vector Machines approaches. We train SVM algorithms with the 2nd degree polynomial kernel, 3rd-degree polynomial kernel, radial basis function kernel, and linear kernel to be able to make comparisons with Quadratic SVM, Cubic SVM, Medium Gaussian SVM, and Linear SVM, correspondingly. In WEKA, we use RandomSubspace method and choose the LDA algorithm to ensemble that uses discriminant analysis algorithm so we implement the same algorithms we use in Matlab.

V. Experimental Results

Table 2 demonstrates the accuracy scores of six algorithms based on the environments.

Table 2: Accuracy Scores of Machine Learning Algorithms

Algorithm\Accuracy	MATLAB	WEKA
Decision Tree	60.9%	67.7 %
Linear SVM	65.3%	67.3%
Quadratic SVM	65.0%	58.5%
Cubic SVM	61.9%	52.3 %
Medium Gaussian SVM	67.0%	63.9 %
Ensemble Subspace Discriminant	67.7%	67.0%

It is evident from the Table 2 that J48 in WEKA and Subspace Discriminant in Matlab has the highest classification accuracy with the rate of 67.7%. Support Vector Machine Algorithms has proportional rate. In both environments Linear SVM has more correctly classify instances among SVM results. Moreover, cubic SVM in WEKA environment has the lowest classification accuracy that is 52.3% among all the other algorithms. So J48 and ensemble learner which ensembles subspace algorithm and discriminant analysis outperforms SVMs in terms of classification accuracy.

Most of the Matlab SVM algorithms outperform the WEKA SVM algorithms as it can be seen in Figure 4. However, there is a tie between these platforms on performing subspace discriminant ensemble learning. The decision tree algorithm shows a better result when we use WEKA to apply it.

VI. Conclusions

In this study, we compared two different platform performances for heart diseases classification. We used Matlab© and WEKA© for the classification of presence of heart disease. We used SVM, Subspace Discriminant and Decision Tree algorithms for the classification. J48 and Subspace Discriminant methods performance are better than the others. Among SVM, linear SVM outperforms the cubic, quadratic and Radial Basis Kernel in both platforms.

If we compare both platforms in term of speed, there is no noticeable difference. However, SVM algorithms take longer to process than the other algorithms in both platforms. On the one hand, the decision tree method gives more reliable results on WEKA. On the other hand, most of the SVM algorithms in Matlab are performing better than the WEKA SVM algorithms. Additionally, Matlab is more flexible in terms of ease of implementation.

As a result, even there are more successful classification results taken from other classifiers in the literature, our classification results are not very high, since we used present functions in both platform. In the future work, we need to perform more experiments to find the optimal parameter values. So if the different parameters are given for the functions and tested for different values, the performance of the classification results will be better. If feature selection algorithms and some data pre-processing techniques are used, the results will be better. Also, comparing different algorithms by using different data mining techniques can improve the classification.

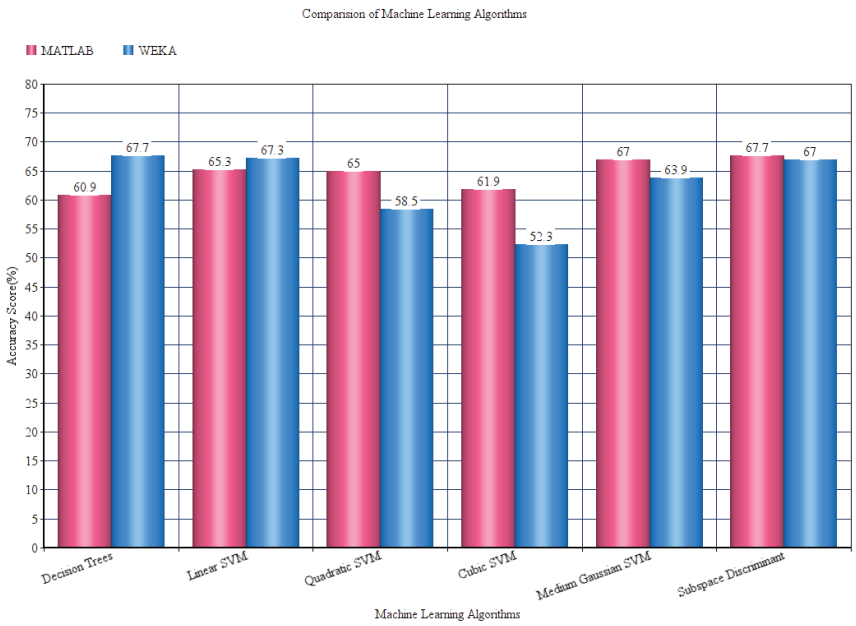

Figure 4. Accuracy Scores of Machine Learning Algorithms

REFERENCES

[1] R. Das, I. Turkoglu, and A. Sengur, "Effective diagnosis of heart disease through neural networks ensembles," *Expert Syst. Appl.*, vol. 36, no. 4, pp. 7675–7680, 2009.

[2] World Health Organization, "Cardiovascular diseases (CVDs)." [Online]. Available: http://www.who.int/mediacentre/factsheets/fs317/en/. [Accessed: 20-Mar-2017].

[3] D. Mozaffarian *et al.*, "AHA statistical Update," *Hear. Dis. Stroke. Circ.*, 2015.

[4] M. Ramaraj and T. Antony Selvadoss, "A Comparative Study of CN2 Rule and SVM Algorithm and Prediction of Heart Disease Datasets Using Clustering Algorithms," in *Network and Complex Systems*, 2013, vol. 3, no. 10, pp. 1–6.

[5] L. Parthiban and R. Subramanian, "Intelligent heart disease prediction system using CANFIS and genetic algorithm," *Int. J. Biol. Biomed. Med. Sci.*, vol. 3, no. 3, 2008.

[6] R. Detrano *et al.*, "International application of a new probability algorithm for the diagnosis of coronary artery disease," *Am. J. Cardiol.*, vol. 64, no. 5, pp. 304–310, 1989.

[7] J. H. Gennari, P. Langley, and D. Fisher, "Models of incremental concept formation," *Artif. Intell.*, vol. 40, no. 1–3, pp. 11–61, 1989.

[8] N. Cheung, "Machine learning techniques for medical analysis. School of Information Technology and Electrical Engineering," B. Sc. Thesis, University of Queenland, 2001.

[9] K. Polat, S. Güne\cs, and A. Arslan, "A cascade learning system for classification of diabetes disease: Generalized discriminant analysis and least square support vector machine," *Expert Syst. Appl.*, vol. 34, no. 1, pp. 482–487, 2008.

[10] K. Bache and M. Lichman, "UCI Machine Learning Repository," *University of California Irvine School of Information*, vol. 2008, no. 14/8. p. 0, 2013.

[11] M. C. Tu, D. Shin, and D. Shin, "A comparative study of medical data classification methods based on decision tree and bagging algorithms," in *Dependable, Autonomic and Secure Computing, 2009. DASC'09. Eighth IEEE International Conference on*, 2009, pp. 183–187.

[12] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.

[13] Anonim, "Support Vector Machines (SVM)," 2017. [Online]. Available: http://www.statsoft.com/Textbook/Support-Vector-Machines. [Accessed: 13-Apr-2017].

[14] M. Hoffman, "Support Vector Machines — Kernels and the Kernel Trick." pp. 10–13, 2006.

[15] M. I. Jordan, "The Kernel Trick." [Online]. Available: https://people.eecs.berkeley.edu/~jordan/courses/281B-spring04/lectures/lec3.pdf. [Accessed: 13-Apr-2017].

[16] X. Li and H. Zhao, "Weighted random subspace method for high dimensional data classification," *Stat. Interface*, vol. 2, no. 2, p. 153, 2009.