MÜHENDİSLİK FAKÜLTESİ

Uluslararası Mühendislik Araştırma ve Geliştirme Dergisi

International Journal of Engineering Research and Development

Cilt/Volume:10 Sayı/Issue:3 Aralık/December 2018 **https://doi.org/10.29137/umagd.472881**

Seçilmiş Konferans Makalesi / Selected Conference Article

Diagnosis of Chronic Kidney Disease using Random Subspace Method with Particle Swarm Optimization

Parçacık Sürüsü Optimizasyonu ile Rastgele Alt Uzay Yöntemi Kullanılarak Kronik Böbrek Hastalığı Teşhisi

Kemal ADEM *[1]

[1]*Tokat Gaziosmanpasa University, Department of Informatics, 60150, Tokat*

Başvuru/Received: 21/10/2018 **Kabul/Accepted**: 12/12/2018 **Son Versiyon/Final Version**: 31/12/2018

Abstract

Late diagnosis of chronic kidney disease, a disease that has increased in recent years and threatens human life, may lead to dialysis or kidney failure. In this study, kNN, SVM, RBF and Random subspace data mining methods were applied on the data set consisting of 400 samples and 24 attributes taken from UCI for classification of chronic kidney disease with particle swarm optimization (PSO) based feature selection method. As a result of the study, the results of the application of each data mining method are compared with the resultant training and test results. As a result of the comparison, the best performance was found to be 99.75% accuracy with PSO and random subspace method. Moreover, as a method of data mining, it has been seen that the random subspace method has higher accuracy rates than the other methods.

Key Words

"Chronic kidney disease, Particle Swarm Optimization, Random subspace"

Öz

Son yıllarda artmış ve insan yaşamını tehdit eden bir hastalık olan kronik böbrek hastalığının geç teşhisi diyaliz veya böbrek yetmezliğine neden olmaktadır. Bu çalışmada kNN, DVM, RBF ve Rastgele alt uzay veri madenciliği yöntemleri, parçacık sürüsü optimizasyonu (PSO) temelli özellik seçim yöntemiyle kronik böbrek hastalığının sınıflandırılması için UCI'den alınan 400 örnek ve 24 öznitelikten oluşan veri setine uygulanmıştır. Çalışma sonucunda, her veri madenciliği yönteminin uygulamasının sonuçları, elde edilen eğitim ve test sonuçlarıyla karşılaştırılmıştır. Karşılaştırma sonucunda, en iyi performans PSO ve rastgele alt uzay yöntemi kullanılarak %99,75 doğruluk oranıyla bulunmuştur. Ayrıca, veri madenciliği yöntemi olarak, rastgele alt uzay yönteminin diğer yöntemlerden daha yüksek doğruluk oranlarına ulaştığı görülmektedir.

Anahtar Kelimeler

"Kronik Böbrek Hastalığı, Parçacık Sürüsü Optimizasyonu, Rastgele Alt Uzay"

**** kemal.adem@gop.edu.tr***

1. INTRODUCTION

Chronic kidney disease (CKD) refers to kidney failure, which typically affects leg swelling, vomiting, fatigue (Liao et al., 2012; Moyer, 2012). Blood and urine tests are performed to demonstrate that the kidney works functionally. For this reason, it is important to establish the screening mechanism to identify CKD symptoms so that important precautions can be taken to avoid any complications (Plantinga et al., 2010).

With the development of information technology, the amount of data that has emerged shows a rapid increase. It is estimated that data stored in the digital environment in the world have doubled every 20 months (Witten & Frank, 2005). It is difficult to process these data with increasing amount of data. Various data mining algorithms have been developed to solve this problem. In the literature, there are algorithms used in data mining and different studies to compare these algorithms. The health sector is also one of these areas. Data mining methods are used for diagnosis and treatment in this sector for obtaining more accurate results and preventing human errors (Karakoyun & Hacibeyoğlu, 2014). It helps the treatment process by identifying the current disease and anticipating future disease. In this study, accurate classification ratios and time performances will be compared by applying data mining methods for diagnosis and diagnosis of chronic kidney disease.

In the literature, studies using data mining methods related to chronic kidney disease have been screened. As a result of the scan, it was seen that Artificial Neural Network (ANN), Radial Basis Function (RBF), Naive Bayes, Logistic Regression, Support Vector Machine (SVM), Decision Trees (C4.5, Random Tree, Random Forest) and kNN were used as data mining methods. In studies using the data set with 400 records and 24 features of 'UC Irvine Machine Learning Repository' (UCI) on Chronic Kidney Disease, the use of the ANN method resulted in a 72% (Sunil & Sowmya, 2017), the use of RBF, ANN and LR algorithms with a 10-fold cross-validation method resulted in the best performance using the 99.66% (Rubini et al., 2015), The use of SVM, NB, RBF, ANN and RF data mining methods results in best results with RF 95% (Kumar, 2016), SVM method was used together with ClassifierSubsetEval, WrapperSubsetEval, CfsSubsetEval, and FilterSubsetEval feature selection algorithms results in best results with FilterSubsetEval and SVM 98.5% (Polat et al., 2017), RepTree, BFTree and J48Tree methods were applied with Adaboost algorithm and in best results BFTree model with Adaboost 99.25% (Başar et al., 2016), SVM with radial basis kernel function 93.75% (Ravindra et al., 2018), Relief and KNN methods were used together to achieve 99% success rates (Kayaalp et al., 2018). Table 1 summarizes the classification results of chronic kidney disease by data mining methods.

Table 1. Classification results of chronic kidney disease with data mining methods.

Authors	Year	Number of data	Method	Accuracy (%)
Rubini et al.	2015		ANN	99.66
Kumar et al.	2016		Random Forest	95
Başar et al.	2016		Adaboost-BFTree	99.25
Polat et al.	2017	400	FilterSubsetEval-SVM	98.5
Sunil et al.	2017		Naive Bayes	72
Ravindra et al.	2018		SVM with radial basis kernel	93.75
Kayaalp et al.	2018		Relief, kNN	99

The most important contribution of this study to the literature is to select the most significant features with PSO for the first time in the data set and to use a Random Subspace method to achieve a very high success rate. Data set identification, attribute selection and classification algorithm used in the study are examined in the second section. In the third section, the parameters used in the methods and experimental studies are given. In the last section of the study, the results are evaluated.

2. MATERIAL AND METHOD

The data set to be used as an example for diagnosing and predicting chronic kidney disease with data mining techniques is derived from the 'UC Irvine Machine Learning Repository' database with 400 samples and 24 (11 numeric, 13 nominal) features (Dua & Karra, 2017). In the dataset, there are 400 samples, 250 of which belong to diseased persons and the other 150 data consist of data belonging to healthy persons.

2.1. Particle Swarm Optimization (PSO)

PSO is a calculation based on intelligence techniques inspired a lot of social behavior (Kennedy & Eberhart, 1995). This method simulates the behavior of flying birds and the exchange of information they have to solve their problems (Kennedy & Eberhart, 2001). The PSO method is used for optimization in many areas (Adem et al., 2018; Delice et al., 2017; Collotta et al., 2017). The steps for selecting an attribute with PSO are as follows:

Step 1. Generate the starting range by starting positions and velocities of the randomly generated particles according to the parameters entered at the beginning.
Step 2. Calculate the fitness values of all the particles in the swarm.
Step 3. Find the local best estimate (*pbest*) for each particle in the current iteration. Find the global best (*gbest*) in local bests in current iteration.
Step 4. Change the pbest if the fitness value of the relevant particle is greater than pbest.
Step 5. If the fitness value of the relevant part is also bigger than *gbest*, change *gbest* and update the part related to the fitness value.
Step 6. Is the number of iterations finished?
Step 7. If the number of iterations has not finished go to Step 2.
Step 8. If the number of iterations is finished, the solution is to select the best attributes with the highest fitness value according to the classification results in the final positions of the particles.

According to the working steps in the algorithm, each particle represents a binary representation of *N*, the total number of features in the data set. Each bit in this display represents an attribute, and '1' indicates that the attribute is selected, and '0' indicates that the attribute is not selected. Each particle updates its position according to its fitness values, first by local, then by global best. In this way, selection of the attributes that achieve the global best is performed.

2.2. Random Subspace Method

The Random Subspace method is a kind of community classification algorithm consisting of various classifiers at the subspace of the attributes in the dataset (Ho, 1998). The classification results are based on the outputs of the individual classifiers selected by the majority vote. This method can be used with many classifiers such as linear classifier, kNN, support vector machine, decision trees (Marina, 2002; Tremblay, 2004; Dacheng et al., 2006). This method is quite advantageous because smaller parts can be better trained because it uses subspaces of the actual data size. This algorithm is very successful in data sets with a large number of attributes, especially because it works with subspaces of the data set. This method is of interest to researchers because it reduces over-learning, introduces a general model, shorter training time, and an easy-to-understand and simple structure than other classical models (Ho, 1998; Kayal & Kannan, 2017). In this study, Random Forest method was used because of its high accuracy rate as a basic classifier in random subspace method.

3. EXPERIMENTAL STUDIES

The selection of attributes used in experimental studies and the evaluation of data mining methods were performed with MATLAB R2017b and WEKA software. The hardware was a computer with 16 GB DDR3 memory with an Intel Core i7 6700 HQ processor. Parameters used in the data mining methods used in the study were determined as experimental results and k = 3 and Euclidean distance criterion in kNN method, kernel function in SVM method, Gauss function as activation function in RBF neural network and random forest tree method in random subspace method. The most significant 17 attributes of the data set were identified by the PSO method applied with 500 particles and 1000 iteration results. 17 attributes selected using PSO method; blood pressure, blood pressure, creatinine, sodium, potassium, hemoglobin, erythrocyte bulk, white blood cell count, hypertension, diabetes, appetite, foot ailment and anemia are specific gravity, albumin, red blood cells, blood sugar. The dimension of the new attribute data set obtained as a result of the attribute selection was (400 x 17). A 5-fold cross validation method was applied to the data set used in the study during the training and test phases. 70% of the data set was used in training and the rest 30% were used as testing. Models created with data mining methods use the accuracy and Kappa values as evaluation criteria. The equation used to calculate the accuracy is given in Equation 1.

$$\boldsymbol{Accuracy} = \frac{\boldsymbol{TP+TN}}{\boldsymbol{TP+TN+FP+FN}} \quad (1)$$

In Equation 1, TP is true positive, TN is true negative, FP is false positive and FN is false negative (Metz, 1978). The equations used in the Kappa analysis, another evaluation criterion used in the study, are given in Equations 2 and 3.

$$P_G = \frac{\sum_{i=1}^{n} G_i}{n} \qquad P_B = \frac{\sum_{i=1}^{n} R_i C_i}{n^2} \quad (2)$$

$$Kappa = \frac{P_G - P_B}{1 - P_B} \quad (3)$$

The observed and expected probabilities of P_G and P_B given in Equations 2 and 3 respectively, G_i: the observed frequency i. in row and column, R_i: total frequency i. in line, C_i: total frequency i. in column and n: total number of observations (Gujarati, 1999; Rey et al., 2012). The accuracy and Kappa analysis values obtained by using kNN, SVM and RBF data mining models are given in Table 2.

Table 2. Accuracy and kappa analysis values of the experiments performed.

Feature Selection	Classification Method	Number of Feature	Accuracy (%)	Kappa Value
None	kNN	24	95.75	0.914
	SVM		97.75	0.9526
	RBF		98.5	0.9683
	Random Subspace		99.5	0.9837
PSO	kNN	17	95.75	0.914
	SVM		98.25	0.963
	RBF		98.75	0.9735
	Random Subspace		**99.75**	**0.9947**

In this study, we compared the results of using kNN, SVM and RBF data mining methods together with the PSO feature selection algorithm on the dataset related to chronic kidney disease from UCI. The results in Table 2, it is observed to increase the classification accuracy of the feature selection process with PSO. The best performance is seen that a result of comparison with 99.75% accuracy and 0.9947 Kappa value of the random subspace method with PSO.

4. CONCLUSION

Chronic kidney disease has been increasing and a disease that threatens human life in recent years. In the literature, lots of data mining studies related to the diagnosis and classification of the disease were observed. In this study, for the classification of chronic kidney disease, kNN, SVM, RBF and Random subspace data mining methods were applied with PSO feature selection method on UCI data set. PSO attribute selection method has been shown to positively influence the classification success. As a data mining method, random subspace method has been found to have higher accuracy rates than other methods. In future studies, it is considered to increase the success rate by using different feature selection algorithms and different data mining methods.

REFERENCES

Adem, K., Hekim, M., & Demir, S. (2019). Detection of hemorrhage in retinal images using linear classifiers and iterative thresholding approaches based on firefly and particle swarm optimization algorithms. Turkish Journal of Electrical Engineering & Computer Sciences. Doi: 10.3906/elk-1804-147.

Başar, M. D., Sarı, P., Kılıç, N., & Akan, A. (2016). Detection of chronic kidney disease by using Adaboost ensemble learning approach, In IEEE Signal Processing and Communication Application Conference (SIU), 2016 24th . 773-776.

Collotta, M., Pau, G., & Maniscalco, V. (2017). A fuzzy logic approach by using particle swarm optimization for effective energy management in IWSNs. IEEE Transactions on Industrial Electronics, 64(12), 9496-9506.

Dacheng, T, Xiaoou, T, Xuelong, L, & Xindong W. (2006). Asymmetric bagging and random subspace for support vector machinesbased relevance feedback in image retrieval. IEEE Transactions on Pattern Analysis and Machine Intelligence. 28(7), 1088-99.

Delice, Y., Aydoğan, E. K., Özcan, U., & İlkay, M. S. (2017). A modified particle swarm optimization algorithm to mixed-model two-sided assembly line balancing. Journal of Intelligent Manufacturing, 28(1), 23-36.

Dua, D., & Karra Taniskidou, E. (2017). UCI Machine Learning Repository [http://archive.ics.uci. edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

Eberhart, R., & Kennedy, J. (1995). A new optimizer using particle swarm theory. In Micro Machine and Human Science, 1995. MHS'95., Proceedings of the Sixth International Symposium on (pp. 39-43). IEEE.

Gujarati, N. D. (1999). Temel Ekonometri. Çev. Ümit Şenesen ve Gülay G. Şenesen. 4. Baskı, 401-674, Literatür Yayınları, İstanbul.

Ho T.K. (1998). The Random Subspace Method for Constructing Decision. IEEE Transactions on Pattern Analysis and Machine Intelligence, Lucent Tech no l., AT&T Bell Labs., Murray Hill, 20(8): 832 – 844.

Karakoyun, M., & Hacıbeyoğlu M. (2014). Biyomedikal veri kümeleri kullanarak makine öğrenmesi sınıflandırma algoritmalarının karşılaştırılması. 2014 October 9-10 [Akıllı Sistemlerde Yenilikler ve Uygulamaları (ASYU) Sempozyumu. İzmir/Turkey].

Kayaalp, F., Basarslan, M. S., & Polat, K. (2018). A hybrid classification example in describing chronic kidney disease. In 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT) (pp. 1-4). IEEE.

Kayal, P., & Kannan, S. (2017). An Ensemble Classifier Adopting Random Subspace Method based on Fuzzy Partial Mining. Indian Journal of Science and Technology, 10(12), 1-8.

Kennedy, J., & Eberhart, R. C. (1999). The particle swarm: social adaptation in information-processing systems. In New ideas in optimization (pp. 379-388). McGraw-Hill Ltd., UK.

Kumar, M. (2016). Prediction of Chronic Kidney Disease Using Random Forest Machine Learning Algorithm. International Journal of Computer Science and Mobile Computing, 5(2), 24-33.

Liao, M., Sung, C., Hung, K., Wu, C., Lo, L., & Lu, K. (2012). Insulin Resistance in Patients with Chronic Kidney Disease. Journal of Biomedicine and Biotechnology.

Marina, S. (2002). Bagging, boosting and the random subspace method for linear classifiers. Pattern Analysis and Applications. 5(2), 121–135.

Metz, C. E. (1978). Basic principles of ROC analysis. In Seminars in nuclear medicine (Vol. 8, No. 4, pp. 283-298). WB Saunders.

Moyer, V. A. (2012). Screening for chronic kidney disease: Us preventive services task force recommendation statement. Annals of internal medicine, vol. 157, no. 8, pp. 567–570.

Plantinga, L. C., Tuot, D. S., & Powe, N. R. (2010). Awareness of chronic kidney disease among patients and providers. Advances in chronic kidney disease, vol. 17, no. 3, pp. 225–236.

Polat, H., Mehr, H. D., & Cetin, A. (2017). Diagnosis of Chronic Kidney Disease Based on Support Vector Machine by Feature Selection Methods, Journal of medical systems, 41(4), 55.

Ravindra, B. V., Sriraam, N., & Geetha, M. (2018). Classification of non-chronic and chronic kidney disease using SVM neural networks. International Journal of Engineering & Technology, 7(1.3), 191-194.

Rey, T, Kordon, A, & Wells, C. (2012). Applied Data Mining for Forecasting Using SAS. SAS Institute Inc, USA, 2012.

Rubini, L. J., & Eswaran, P. (2015). Generating comparative analysis of early stage prediction of Chronic Kidney Disease. Journal Of Modern Engineering Research, 5(7), 49-55.

Sunil, D., & Sowmya, B. P. (2017). Chronic Kidney Disease Analysis using Data Mining.

Tremblay, G. (2004). Optimizing nearest neighbour in random subspaces using a multi-objective genetic algorithm. 17th International Conference on Pattern Recognition. p. 208–11. Crossref

Witten, I. H., & Frank, E. (2005). Data Mining Practical Machine Learning Tools and Techniques. 2nd ed., San Francisco/ABD.