Karaelmas Science and Engineering Journal
Journal home page: http://fbd.beun.edu.tr
DOI: 10.7212%2Fzkufbd.v8i1.717

Research Article
Received / Geliş tarihi : 26.11.2016
Accepted / Kabul tarihi : 26.05.2017

Feature Selection Using Cubic Smoothing Spline and Robust Regression

Kübik Smoothing Spline ve Robust Regresyon Kullanılarak Özellik Seçimi

Övünç Polat*

Faculty of Engineering, Department of Electrical and Electronics Engineering, Akdeniz University, Antalya, Turkey

Abstract

An efficient feature selection approach based on the combination of cubic smoothing spline and robust regression is presented for classification applications in this study. Six different data sets are used to test the proposed feature selection algorithm. Cubic smoothing spline and robust regression terms are calculated for each attributes in related classification application with %50 of dataset. The success of proposed algorithm is evaluated by using K-Nearest Neighbor (KNN) algorithm and Discriminant analysis. Obtained simulation results show that proposed feature selection approach has high classification accuracy rate with fewer number of features. For example, in the Heart(Statlog) dataset classification problem, 66.6% classification accuracy is obtained for KNN (K = 5) and 13 feature values before using the feature selection algorithm, using the proposed feature selection approach, a classification success of 83.7% is achieved by using 6 features.

Keywords: Classification, Cubic smoothing spline, Feature selection, Robust regression

Öz

Bu çalışmada sınıflandırma uygulamaları için kübik smoothing spline ve robust regresyonun kombinasyonu temelli etkili bir özellik seçim yaklaşımı sunulmuştur. Altı farklı veri seti önerilen özellik seçim algoritmasını test etmek için kullanılmıştır. İlgili sınıflandırma uygulamasında verisetinin %50'si kullanılarak her bir özellik değeri için kübik smoothing spline ve robust regresyon terimleri hesaplandı. Önerilen algoritmanın başarısı K. En Yakın Komşu Algoritması ve Diskriminant analizi kullanılarak değerlendirilmiştir. Elde edilen benzetim sonuçları önerilen özellik seçim yaklaşımının daha az özellik sayısıyla yüksek sınıflandırma başarı oranına sahip olduğunu göstermektedir. Örneğin Kalp(Statlog) veri seti sınıflandırma probleminde önerilen özellik seçim algoritması kullanmadan önce KNN (K=5 için) ile 13 özellik değeri için %66.6 sınıflandırma doğruluğu elde edilirken, önerilen özellik seçim yaklaşımı kullanılarak 6 özellik ile %83.7 sınıflandırma başarısı elde edilmiştir.

Anahtar Kelimeler: Sınıflandırma, Kübik smoothing spline, Özellik seçimi, Robust regresyon

1. Introduction

Feature selection process aims to removal of irrelevant and redundant features in application. The purpose of feature selection is determined best features in order to increase the classification accuracy and decrease the computational cost (Liu et al. 2005, Shahzad et al. 2013).

This study presents an approach for feature selection the combination of cubic smoothing spline and robust regression. In the literature, cubic smoothing spline is used for different applications such as background correction (Kuligowski et al. 2010), local linear forecasts (Hyndman et al. 2005). Robust regression algorithm is a important tool for data analysis (Chen 2002). It can be used for outlier detection (Chen 2002 and Wang et al. 2014) and classification (Polat 2015).

Proposed approach is tested for 6 different dataset from UCI dataset archives (Machine Learning Repository 2016). Selected features are used in KNN (Dasarathy 1991) and Discriminant Analysis (Fukunaga 1990) classifier. Next section gives a feature selection procedure. Simulations and results are given in the last section.

2. Materials and Method

In proposed algorithm, cubic smoothing spline and robust regression terms are calculated for each attributes in classification application with %50 of dataset. Output

*Corresponding Author: ovuncpolat@akdeniz.edu.tr

Övünç Polat orcid.org/0000-0002-9581-2591

values are calculated for the other half of the dataset. Then, the mean absolute error values between the target value and obtained output value are calculated, and average error values are calculated for the each of smoothing spline and robust regression.

In robust regression process, the ordinary least squares analysis used for all datasets. When the calculated error value for each attributes greater than the average error value, related feature is removed from the feature set for each of the smoothing spline and robust regression. Then, remaining the same features for both methods are used to classification. If same features are not remain for cubic smoothing spline and robust regression, the combination of the remaining features for each of these methods are used to classification. Figure 1 shows the feature selection procedure.

3. Results

The success of proposed feature selection algorithm is examined by the Pima Indians Diabetes, iris, Heart (Statlog), balance scale, glass and seeds dataset from UCI dataset archives (Machine Learning Repository 2016). The datasets are classified using KNN and Linear Discriminant Analysis. 50% of the dataset is used as reference set for KNN. For K=1 and K=5, classification accuracy rates are determined in this work. For balance scale dataset, 312 samples are used as reference data. The remaining 313 samples are used for test. Properties of used all datasets is given in table 1.

The numbers of selected features by using cubic smoothing spline, robust regression and the combination of these techniques for tested datasets is given in table 2. For balance scale dataset, number of features is the same before and after the selection process using combined approach. It found that all of the features to be effective using proposed method.

Table 1. The selected datasets.

Datasets	Attributes	Instances	Classes
Diabetes	8	768	2
Iris	4	150	3
Heart Statlog	13	270	2
Balance Scale	4	625	3
Glass	9	214	6
Seeds	7	210	3

The classification accuracy rates using KNN given in table 3. As can be seen from table 3, the accuracy rates are about the near or higher after feature selection by using the combination of cubic smoothing spline and robust regression for KNN classifier. The obtained accuracy rates using Linear Discriminant Analysis given in table 4. As can be seen from table 4, the classification accuracy rates for diabetes, iris and Heart Statlog datasets are about the same or near after feature selection by using the combination of cubic smoothing spline and robust regression for Linear Discriminant Analysis.

4. Conclusion

In this work, a feature selection method is designed based on the combination of cubic smoothing spline and robust regression. The proposed approach is carried out for six different dataset. The success of proposed selection method is evaluated by using KNN classifier and Linear Discriminant Analysis. Simulation results show that the classification accuracy rate for KNN classifier is about the near or higher after feature selection process for all dataset by using the combination of cubic smoothing spline and robust regression. The proposed feature selection approach can be tested using different classifiers.

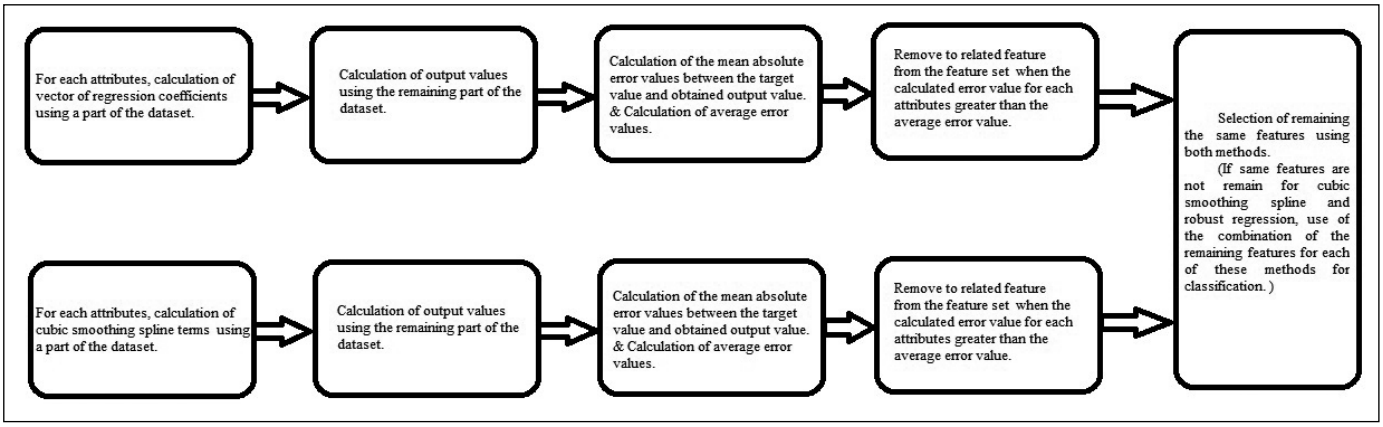

Figure 1. The procedure of the selection of features.

Table 2. The Numbers of Selected Features.

Datasets	Before feature selection	After feature selection for cubic smoothing spline	After feature selection for robust regression	After feature selection for the combination of cubic smoothing spline and robust regression
	Number of Features	Number of Selected Features	Number of Selected Features	Number of Selected Features
Diabetes	8	7	3	3
Iris	4	2	2	2
Heart Statlog	13	10	7	6
Balance Scale	4	2	2	4
Glass	9	8	4	4
Seeds	7	2	1	3

Table 3. The average classification accuracy rates by using KNN for K=1 and K=5.

Datasets	Before feature selection		After feature selection for cubic smoothing spline		After feature selection for robust regression		After feature selection for the combination of cubic smoothing spline and robust regression	
	Accuracy for K=1	Accuracy for K=5	Accuracy for K=1	Accuracy for K=5	Accuracy for K=1	Accuracy for K=5	Accuracy for K=1	Accuracy for K=5
Diabetes	67.1%	73.4%	68.2%	74.4%	67.7%	73.1%	67.7%	73.1%
Iris	94.6%	92.0%	93.3%	97.3%	93.3%	97.3%	93.3%	97.3%
Heart Statlog	55.5%	66.6%	75.5%	77.0%	78.5%	72.5%	81.4%	83.7%
Balance Scale	79.4%	84.2%	48.7%	65.3%	48.7%	65.3%	79.4%	84.2%
Glass	45.7%	58.8%	45.7%	58.8%	43.9%	58.8%	43.9%	58.8%
Seeds	84.7%	87.6%	79.0%	82.8%	45.7%	50.4%	84.7%	86.6%

Table 4. The average classification accuracy rates by using Linear Discriminant Analysis.

Datasets	Before feature selection	After feature selection for cubic smoothing spline	After feature selection for robust regression	After feature selection for the combination of cubic smoothing spline and robust regression
	Accuracy	Accuracy	Accuracy	Accuracy
Diabetes	78.9%	79.1	76.0%	76.0%
Iris	96.0%	96.0	96.0%	96.0%
Heart Statlog	87.4%	84.4	84.4%	82.9%
Balance Scale	50.0%	63.4	63.4%	50.0%
Glass	58.8%	57.9	42.9%	42.9%
Seeds	96.1%	81.9	56.1%	84.7%

5. Acknowledgment

The research has been supported by the Research Project Department of Akdeniz University, Antalya, Turkey.

6. References

Chen, C. 2002. Robust Regression and Outlier Detection with the ROBUSTREG Procedure. *Proceedings of the 27th SAS Users Group Int. Conference*, Cary NC: SAS Institute, Inc.

Dasarathy, BV. 1991. Nearest Neighbor (NN) Norms: NN pattern classification techniques. *IEEE Computer Society Press*, Los Alamitos, CA.

Fukunaga, K. 1990. Introduction to Statistical Pattern Recognition, 2nd ed., San Diego, CA: Academic.

Hyndman, RJ., King, ML., Pitrun, I., Billah, B. 2005. Local Linear Forecasts Using Cubic Smoothing Splines. *Aust. N. Z. J. Stat.*, 47: 87–99.

Kuligowski, J., Carrión D., Quintás, G., Garrigues, S., Guardia, M. 2010. Cubic smoothing splines background correction in on-line liquid chromatography–Fourier transform infrared spectrometry. *J Chr. A*, 1217: 6733–6741.

Liu, H., Lei, Y. 2005. Toward Integrating Feature Selection Algorithms for Classification and Clustering. *IEEE Trans. K. D. E.,* 17 (4): 491-502.

Machine Learning Repository, 2016. Center for Machine Learning and Intelligent Systems. Retrieved from: http://archive.ics.uci.edu/ml/

Polat, Ö. 2015. A robust regression based classifier with determination of optimal feature set. *J. App. R. T.*, 13:443-6.

Shahzad, W., Asad, S., Khan, MA. 2013. Feature subset selection using association rule mining and JRip classifier. *Int. J. P. S.* 8(18):885–896.

Wang, J., Xiong, S. 2014. A hybrid forecasting model based on outlier detection and fuzzy time series – A case study on Hainan wind farm of China. *Energy*, 76: 526–541.