Article Type. Research Article
Journal of Intelligent Systems: Theory and Applications 2(1) 2019: 7-12.

Eta Correlation Coefficient Based Feature Selection Algorithm for Machine Learning: E-Score Feature Selection Algorithm

Muhammed Kürşad UÇAR[1*]

[1]Sakarya University, Faculty of Engineering, Electrical-Electronics Engineering, 54187, Sakarya / Turkey

mucar@sakarya.edu.tr

Abstract
Feature selection algorithms are great importance in the field of machine learning. The primary function of feature selection algorithms is to select features in a meaningful way. Features Selection Algorithms methods are still being developed today. The reason for this is that data quantities are growing day by day. As the data increases, more advanced, better performance, feature selection algorithms are needed. In this study, Eta Correlation Coefficient based E-Score Feature selection algorithm was developed. Two versions were prepared for E-Score. We tested the performance of the E-Score method with three classifiers and compared with conventional F-Score Feature Selection Algorithm. According to the results, both versions of the E-Score feature selection algorithm have improved performance and is better than the F-Score. According to these results, it is thought that the E-Score Feature Selection Algorithm can be used in the field of machine learning.

Keywords: Eta Correlation Coefficient, E-Score Feature Selection Algorithm, Feature Selection Methods.

Makine Öğrenmesi için Eta Korelasyon Katsayısı Tabanlı Özellik Seçme Algoritması: E-Score Özellik Seçme Algoritması

Öz
Makine öğrenmesi alanında özellik seçme algoritmaları büyük öneme sahiptir. Çok büyük verilerin anlamlı bir şekilde azaltılması özellik seçme algoritmalarının temel işlevidir. Bu yöntemler günümüzde hala geliştirilmeye devam etmektedir. Bunun sebebi her geçen gün daha büyük verilerle çalışıyor olmasıdır. Veriler arttıkça daha gelişmiş, performansı daha iyi özellik seçme algoritmalarına ihtiyaç duyulacaktır. Bu çalışmada Eta Korelasyon Katsayısı tabanlı E-Score Özellik seçme algoritması geliştirilmiştir. Geliştirilen yöntem için iki farklı versiyon hazırlanmıştır. E-Score yönteminin performansı üç sınıflandırıcı ile test edilmiştir. Ayrıca literatürde bulunan F-Score Özellik Seçme Algoritması ile de kıyaslanmıştır. Elde edilen sonuçlara göre E-Score özellik seçme algoritmasının her iki versiyonu da performansı arttırmıştır. Ayrıca F-Score ile kıyaslandığında daha iyi başarı oranı elde etmiştir. Bu sonuçlara E-Score Özellik Seçme Algoritmasının makine öğrenmesi alanında kullanılabileceği düşünülmektedir.

Anahtar Kelimeler: Eta Korelasyon Katsayısı, E-Score Özellik Seçme Algoritması, Özellik Seçme Yöntemleri.

1. Introduction

In machine learning, datasets are the essential elements. Thanks to today's technology, the amount of collected data has reached enormous amounts. Massive data sometimes have a negative impact on the machine learning process (Guan *et al.*, 2014). Nowadays, one of the most significant problems in machine learning is that significant data lengthens the process and reduces performance. The reason for the decrease in performance is that the irrelevant data is in the cluster.

* Corresponding Author. Phone: +90 506 849 31 46
E-mail: mucar@sakarya.edu.tr

Recieved : Dec 18, 2018
Revision : Jan 9, 2019
Accepted : Jan 17, 2019

To solve this problem, Polat has developed algorithms to select the related properties from datasets (Polat and Güneş, 2009; Kavsaoğlu, Polat and Bozkurt, 2014). These algorithms are commonly called feature selection algorithms.

Feature selection algorithms aim to increase the performance of classification by selecting important features from datasets according to specific algorithms (Polat and Güneş, 2009; Guan *et al.*, 2014; Cai *et al.*, 2018). Training time, classification accuracy rate, data size, number of features selected affects performance. There are many different types of data in the datasets (Cai *et al.*, 2018). Therefore, a feature selection algorithm cannot be used in each dataset.

Feature selection algorithms can be used wherever machine learning is available. For example, it is used in many areas such as image processing, signal processing, classification problems and data mining (Khotanzad and Hong, 1990; Goltsev and Gritsenko, 2012). As the problems develop, new solutions are developed. Recently, the Ensemble Feature Selection algorithms have been developed (Li, Gao and Chen, 2012; Elghazel and Aussem, 2015). This method combines performance with different feature selection algorithms to improve performance.

The performance of the feature selection algorithms developed in the literature is generally assessed by classification algorithms such as k-Nearest Neighborhood Algorithm (kNN), Support Vector Machines (SVMs), Radial Basis Function (RBF) (Huang, 1999; Cai *et al.*, 2018). A useful feature selection algorithm has a high accuracy rate and fast operation (Cai *et al.*, 2018).

Many feature selection algorithms have been developed in the literature. These can be developed based on statistical or different basic principles (Tsang-Hsiang Cheng, Chih-Ping Wei and Tseng, 2006; Khoshgoftaar *et al.*, 2012). In the literature, feature selection algorithms use three different methods according to the learning method (Cai *et al.*, 2018). These are Filter, Wrapper, and Embedded Model. In the filter model, the selection is made by considering the relationship between the features and the class label (Cai *et al.*, 2018). The calculation workload is less than the Wrapper model (Cai *et al.*, 2018). The filter model makes the selection of features according to a specific criterion (Cai *et al.*, 2018). The embedded method selects the features in the training process (Cai *et al.*, 2018). All these algorithms still need to be improved regarding performance.

In this article, we have developed an Eta correlation coefficient-based feature selection algorithm like the filter model. The features were selected according to the correlation value between the features and the class label and their performances were tested with kNN, Probabilistic Neural Networks (PNN) and SVMs.

In order to reach the highest level of quality, authors should comply with the rules set out in this template.

The template will be returned to the author for the reorganization of the articles not prepared by the template. Returned articles must be returned after they have been arranged by the rules.

2. Materials and Methods

Figure 1 shows the operation steps in this article. First, the feature selection algorithms select the features in the datasets. Then, various classifiers classify features. Finally, the performances of the classifiers are calculated.

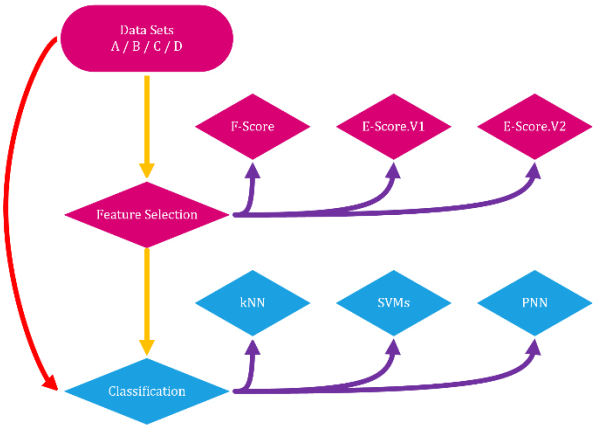

Figure 1. Flow diagram

2.1. Sample Datasets

Four datasets (A / B / C / D) were used to test the developed method (Table 1). These are downloaded from the UCI Machine Learning Repository (Andrzejak *et al.*, 2001; Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, 2001). The data includes the Electroencephalography (EEG) signal features. Each dataset has two labels (Epilepsy(1)/Non-Epilepsy(2)). Each dataset has 178 properties.

Table 1. Sample datasets

Information	**Datasets**			
	A	**B**	**C**	**D**
Epilepsy	1150	1150	1150	1150
Non-Epilepsy	1150	1150	1150	1150
Total	2300	2300	2300	2300
Number of Features	178	178	178	178

2.2. Eta correlation coefficient

In the literature, there are many correlation calculation methods. However, each data group needs the appropriate unique correlation formula (Alpar, 2010). There are various types of data in the field of machine learning. Class labels are often Unordered Qualitative variables. Eta Correlation Coefficient (r_{pb}) is used when calculating the correlation coefficient

between qualitative and continuous numerical variables (Equation 1) (Alpar, 2010). The method changes when the data type changes (Alpar, 2010).

$$r_{pb} = \frac{\overline{Y_1} - \overline{Y_0}}{s_y}\sqrt{p_0 p_1} \quad (1)$$

In the equation, $\overline{Y_0}$ and $\overline{Y_1}$ are the average of the data in class 0 and 1 respectively. s_y is the standard deviation of all data in both classes (Equation 2).

$$s_y = \sqrt{\frac{\sum Y^2 - \frac{(\sum Y)^2}{n}}{n}} \quad (2)$$

N, N_0 and N_1 is the number of elements of the total, Class 0 and Class 1 respectively. Equation 3 shows p_0 and p_1.

$$p_0 = \frac{N_0}{N}, p_1 = \frac{N_1}{N} \quad (3)$$

2.3. Feature selection based on Eta correlation coefficient: Eta-Score

In this study, we have developed the Eta correlation coefficient-based feature selection algorithm. The algorithm has two versions (E-Score.V1 - E-Score.V2).

2.3.1. Selection Criteria 1 - E-Score.V1

Figure 2 shows the E-Score.V1 process steps. First, the Eta correlation coefficient (*Eta* or r_{pb}, Equation 1) for each feature is calculated. Second, the Eta threshold is determined (*Eta* or r_{pb}, Equation 1). If $Eta > Eta_{mean}$, that feature is selected.

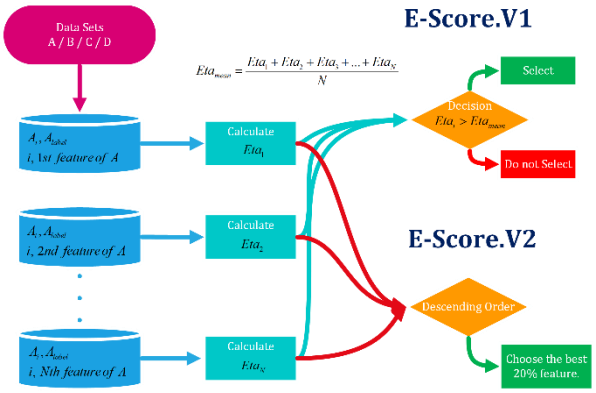

Figure 2. Flow diagram for E-Score

$$Eta_{mean} = \frac{Eta_1 + Eta_2 + Eta_3 + ... + Eta_N}{N} \quad (4)$$

2.3.2. Selection Criteria 2 - E-Score.V2

Figure 2 shows the E-Score.V2 process steps. First, the *Eta* value for the features is sorted in descending order. The first 20% of the features are selected. Eighty percent of the value of a book is hidden in 20 percent of the pages (Koch, 2014). The purpose of this study is to reduce the data size by 80% and to improve system performance. In addition to this aim, performance evaluations were made by selecting the features from 1% to 100% (Figure 3).

2.3.3. Performance Evaluation

kNN, PNN, and SVMs were used to test the proposed algorithms. The performances of these classifiers were measured according to the following criteria. Accuracy rate, sensitivity, specificity, and the working time are performance evaluation criteria. Also, the working time of the algorithm was evaluated.

The datasets for the classification process are divided into two sets: Training (50%) and Test (50%) (Table 2). Besides, different Training / Test rates for E-Score.V1 have been tried, and the accuracy rate is shown graphically for each classifier and each data group (Figure 4).

Table 2. Datasets distribution for the test and training process

Class	For A dataset			For B dataset		
	Training (%50)	**Test (%50)**	**Total**	**Training (%50)**	**Test (%50)**	**Total**
Epilepsy	1150	1150	2300	1150	1150	2300
Non-Epilepsy	1150	1150	2300	1150	1150	2300
Total	**2300**	**2300**	**4600**	**2300**	**2300**	**4600**
Class	**For C dataset**			**For D dataset**		
	Training (%50)	**Test (%50)**	**Total**	**Training (%50)**	**Test (%50)**	**Total**
Epilepsy	1150	1150	2300	1150	1150	2300
Non-Epilepsy	1150	1150	2300	1150	1150	2300
Total	**2300**	**2300**	**4600**	**2300**	**2300**	**4600**

3. Results

This study aims to develop a new feature selection algorithm in the field of machine learning. For this, we established the Eta correlation coefficient-based E-Score Feature Selection Algorithm with two different versions (Section 2.3.). The improved method has been tested in different classifiers according to some performance criteria (Section 2.3.3.). The E-Score was also compared with the F-Score Feature Selection algorithm available in the literature (Polat and Güneş, 2009).

The working time of the E-Score algorithm was measured for four different datasets (Table 3). Besides, the working time performance of the algorithm was compared with the F-Score feature selection algorithm (Table 3).

Table 3. Results of E-Score working time evaluation

Datasets	All Features	F-Score		Eta-Boost.V1		Eta-Boost.V2	
		Number	Time (sec)	Number	Time (sec)	Number	Time (sec)
A	178	60	0.019	79	0.171	36	0.170
B	178	68	0.020	85	0.166	36	0.165
C	178	62	0.019	77	0.166	36	0.165
D	178	54	0.019	73	0.166	36	0.164

sec: Second

kNN, PNN and SVMs classifiers evaluated the performance of the E-Score algorithm. According to the performance results, kNN classifier and for each dataset (A/B/C/D), E-Score.V2 is the best-performing feature selection algorithm among other algorithms (Table 4). Besides, E-Score.V1 has similar performance with F-Score (Table 4).

In the PNN classifier, the performance of feature selection algorithms depends on the datasets (A/B/C/D) (Table 4). E-Score.V1 is the best feature selection algorithm for SVMs (Table 4). When the feature selection algorithms examined the effects of the classifiers operating time, E-Score.V2 most successful feature selection algorithm (Table 4).

Table 4. Evaluation of the performance of the E-Score feature selection algorithm

A

Classifier	kNN				PNN				SVMs			
Performance	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**
All Features	86.78	0.74	1.00	0.33	93.17	0.87	1.00	2.56	99.61	0.99	1.00	0.12
F-Score	88.91	0.78	1.00	0.11	92.17	0.94	0.90	0.79	99.13	0.98	1.00	0.08
Eta-Boost.V1	87.35	0.75	1.00	0.13	91.91	0.90	0.94	1.04	99.26	0.99	1.00	0.08
Eta-Boost.V2	90.39	0.81	1.00	0.06	63.26	0.98	0.29	0.52	98.96	0.98	1.00	0.07

B

Classifier	kNN				PNN				SVMs			
Performance	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**
All Features	85.17	0.70	1.00	0.32	91.09	0.84	0.98	2.55	98.13	0.96	1.00	0.14
F-Score	87.30	0.75	1.00	0.11	78.57	0.87	0.70	0.96	96.74	0.95	0.98	0.28
Eta-Boost.V1	86.35	0.73	1.00	0.15	88.43	0.88	0.89	1.13	97.57	0.96	0.99	0.11
Eta-Boost.V2	88.43	0.77	1.00	0.07	73.61	0.93	0.54	0.52	95.74	0.94	0.98	0.24

C

Classifier	kNN				PNN				SVMs			
Performance	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**
All Features	83.09	0.66	1.00	0.32	89.13	0.80	0.98	2.65	97.91	0.97	0.99	0.14
F-Score	85.96	0.72	1.00	0.11	81.00	0.91	0.71	0.80	96.48	0.95	0.98	0.11
Eta-Boost.V1	85.09	0.70	1.00	0.12	75.83	0.88	0.63	1.02	96.78	0.96	0.98	0.10
Eta-Boost.V2	87.65	0.75	1.00	0.07	59.70	0.96	0.23	0.53	95.96	0.95	0.97	0.09

D

Classifier	kNN				PNN				SVMs			
Performance	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**	**Acc**	**Sen**	**Spe**	**T**
All Features	81.43	0.64	0.99	0.32	48.78	0.90	0.08	2.62	94.30	0.96	0.93	0.34
F-Score	83.78	0.69	0.98	0.08	48.91	0.96	0.02	0.75	93.39	0.93	0.93	0.25
Eta-Boost.V1	82.74	0.67	0.99	0.12	49.48	0.96	0.03	0.97	93.78	0.95	0.93	0.27
Eta-Boost.V2	85.26	0.73	0.97	0.07	49.43	0.98	0.01	0.52	92.70	0.93	0.92	0.24

Acc Accuracy Rate (%) , Sen Sensitivity, Spe Specificity, T Time (second)

E-Score.V2 selects only the first 20\% of all features. The percentage change can increase performance (Figure 3). In kNN and SVMs, small performance changes were observed due to the number of features (Figure 3). However, PNN performance is highly variable depending on the number of features (Figure 3).

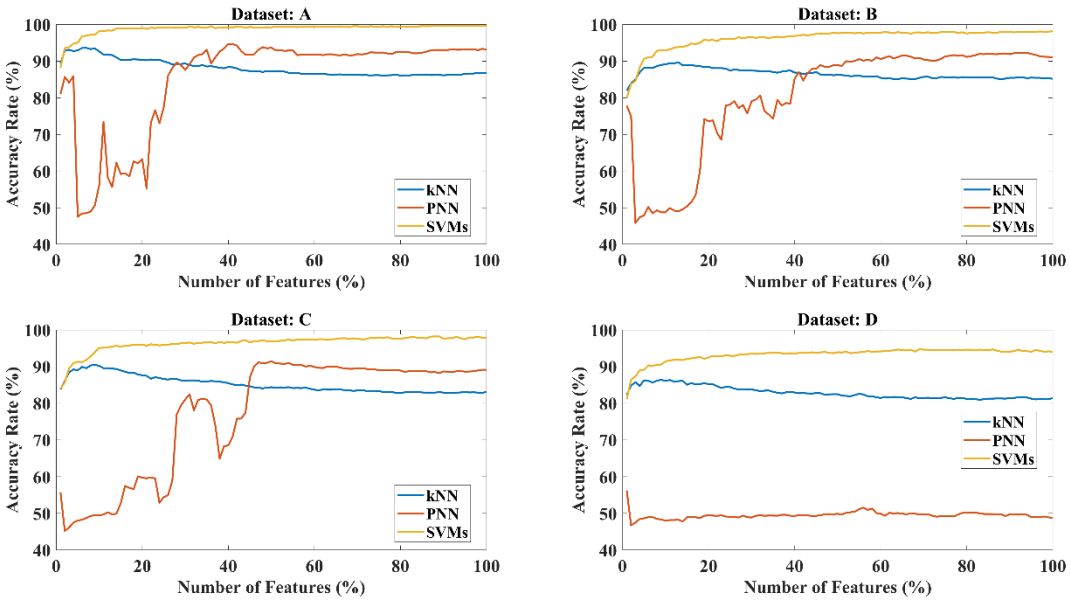

Figure 3. For E-Score.V2, Accuracy rates for selected properties in different percentages

Training and Test rates are 50% and 50% for classification. For the E-Score.V1 algorithm, the change of the test data was monitored from 5% to 95% (Figure 4). If the test data exceeds 65-70%, system performance decreases (Figure 4). When the test dataset is 50%, the system performance is maximum (Figure 4).

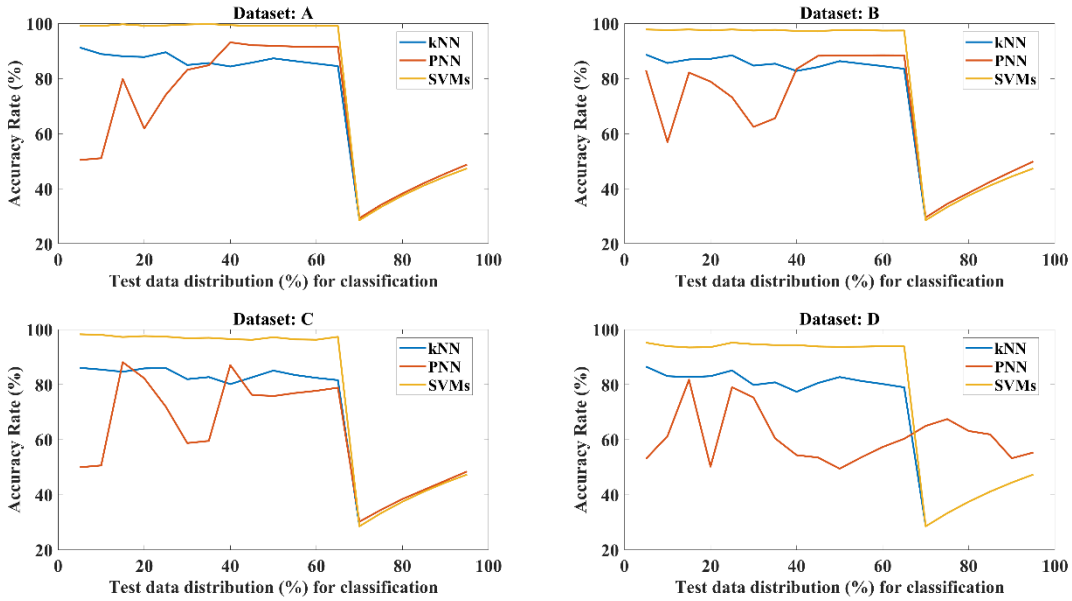

Figure 4. E-Score.V1 feature selection algorithm performance for different Training / Test distributions

4. Discussion and Conclusions

A new feature selection algorithm has been developed with this study. Feature selection algorithms are an essential part of machine learning. These algorithms are required to shorten the duration of learning and to minimize the number of features (Polat and Güneş, 2009; Guan *et al.*, 2014; Kavsaoğlu, Polat and Bozkurt, 2014; Cai *et al.*, 2018). The number of features selected by the E-Score method is between 20-40\% compared to the total number of features. This reduces the workload considerably. Besides, E-Score increases the classification performance of the system. E-Score performance is quite good compared to the F-Score feature selection algorithm in the literature (Polat and Güneş, 2009). E-Score has reduced the workload and improved the performance of the system, such as feature selection algorithms in the literature (Polat and Güneş, 2009; Guan *et al.*, 2014; Kavsaoğlu, Polat and Bozkurt, 2014; Cai *et al.*, 2018).

E-Score is a correlation-based feature selection algorithm. As the E-Score is statistical-based, the correlation between features and intergroup correlation can be accurately estimated (Alpar, 2010). However, the method can only be applied between qualitative and continuous numerical variables. For other data types, similar process with E-Score is recommended, but it is recommended to use the correlation formulas according to the data type.

According to the results obtained in the study, each feature selection algorithm does not adapt to each dataset. Performance has improved. However, there is no corresponding improvement in each dataset.

As a result, when the E-Score feature selection algorithm is examined regarding performance, it is considered to be a quality method that can be used in the field of machine learning.

Acknowledgment

Matlab-based codes for the E-Score Feature Selection Algorithm are available from GitHub.

Access connection: https://github.com/MKUCAR/E_Score_Feature_Selection.git

Please refer to each publication you use the method or code.

References

Alpar, R. (2010) *Applied Statistic and Validation - Reliability*. Detay Publishing. Available at: https://books.google.com.tr/books/about/Uygulamalı_istatistik_ve_geçerlik_güv.html?id=ITk1MwEACAAJ&pgis=1 (Accessed: 11 January 2016).

Andrzejak, R. G. *et al.* (2001) 'Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state', *Physical Review E - Statistical Physics, Plasmas, Fluids, and Related Interdisciplinary Topics*, 64(6), p. 8. doi: 10.1103/PhysRevE.64.061907.

Andrzejak RG, Lehnertz K, Rieke C, Mormann F, David P, E. C. (2001) *UCI Machine Learning Repository: Epileptic Seizure Recognition Data Set, UCI*. Available at: https://archive.ics.uci.edu/ml/datasets/Epileptic+Seizure+Recognition (Accessed: 14 August 2018).

Cai, J. *et al.* (2018) 'Feature selection in machine learning: A new perspective', *Neurocomputing*. Elsevier, 300, pp. 70–79. doi: 10.1016/J.NEUCOM.2017.11.077.

Elghazel, H. and Aussem, A. (2015) 'Unsupervised feature selection with ensemble learning', *Machine Learning*. Springer US, 98(1–2), pp. 157–180. doi: 10.1007/s10994-013-5337-8.

Goltsev, A. and Gritsenko, V. (2012) 'Investigation of efficient features for image recognition by neural networks', *Neural Networks*. Pergamon, 28, pp. 15–23. doi: 10.1016/J.NEUNET.2011.12.002.

Guan, D. *et al.* (2014) 'A Review of Ensemble Learning Based Feature Selection', *IETE Technical Review*, 31(3), pp. 190–198. doi: 10.1080/02564602.2014.906859.

Huang, D.-S. (1999) 'Radial Basis Probabilistic Neural Networks: Model and Application', *International Journal of Pattern Recognition and Artificial Intelligence*. World Scientific Publishing Company, 13(07), pp. 1083–1101. doi: 10.1142/S0218001499000604.

Kavsaoğlu, A. R., Polat, K. and Bozkurt, M. R. (2014) 'A novel feature ranking algorithm for biometric recognition with PPG signals.', *Computers in biology and medicine*, 49, pp. 1–14. doi: 10.1016/j.compbiomed.2014.03.005.

Khoshgoftaar, T. *et al.* (2012) 'First Order Statistics Based Feature Selection: A Diverse and Powerful Family of Feature Seleciton Techniques', in *2012 11th International Conference on Machine Learning and Applications*. IEEE, pp. 151–157. doi: 10.1109/ICMLA.2012.192.

Khotanzad, A. and Hong, Y. H. (1990) 'Rotation invariant image recognition using features selected via a systematic method', *Pattern Recognition*. Pergamon, 23(10), pp. 1089–1101. doi: 10.1016/0031-3203(90)90005-6.

Koch, R. (2014) *The 80/20 Principle and 92 Other Powerful Laws of Nature: The Science of Success*. Available at: https://www.amazon.com/80-20-Principle-Secret-Achieving/dp/1486213421 (Accessed: 23 September 2018).

Li, Y., Gao, S.-Y. and Chen, S. (2012) 'Ensemble Feature Weighting Based on Local Learning and Diversity', *AAAI*. Available at: https://www.semanticscholar.org/paper/Ensemble-Feature-Weighting-Based-on-Local-Learning-Li-Gao/733e4973aec6d9de139781a76ca2f6b3f05b293b (Accessed: 24 September 2018).

Polat, K. and Güneş, S. (2009) 'A new feature selection method on classification of medical datasets: Kernel F-score feature selection', *Expert Systems with Applications*, 36(7), pp. 10367–10373. doi: 10.1016/j.eswa.2009.01.041.

Tsang-Hsiang Cheng, Chih-Ping Wei and Tseng, V. S. (2006) 'Feature Selection for Medical Data Mining: Comparisons of Expert Judgment and Automatic Approaches', in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*. IEEE, pp. 165–170. doi: 10.1109/CBMS.2006.87.