Bulletin of Electrical Engineering and Informatics
Vol. 9, No. 4, August 2020, pp. 1569~1577
ISSN: 2302-9285, DOI: 10.11591/eei.v9i4.1959

New feature selection based on kernel

Zuherman Rustam, Sri Hartini
Department of Mathematics, University of Indonesia, Indonesia

Article Info

Article history:

Received Aug 10, 2019
Revised Oct 5, 2019
Accepted Dec 3, 2019

Keywords:

Feature selection
Kernel
Kernel matrix
K-means clustering
Polynomial kernel
Radial basis function kernel

ABSTRACT

Feature selection is an essential issue in machine learning. It discards the unnecessary or redundant features in the dataset. This paper introduced the new feature selection based on kernel function using 16 the real-world datasets from UCI data repository, and k-means clustering was utilized as the classifier using radial basis function (RBF) and polynomial kernel function. After sorting the features using the new feature selection, 75 percent of it was examined and evaluated using 10-fold cross-validation, then the accuracy, F1-Score, and running time were compared. From the experiments, it was concluded that the performance of the new feature selection based on RBF kernel function varied according to the value of the kernel parameter, opposite with the polynomial kernel function. Moreover, the new feature selection based on RBF has a faster running time compared to the polynomial kernel function. Besides, the proposed method has higher accuracy and F1-Score until 40 percent difference in several datasets compared to the commonly used feature selection techniques such as Fisher score, Chi-Square test, and Laplacian score. Therefore, this method can be considered to use for feature selection

Corresponding Author:

Zuherman Rustam,
Department of Mathematics,
University of Indonesia,
Depok 16424, Indonesia.
Email: rustam@ui.ac.id

1. INTRODUCTION

Feature selection is one of the essential methods in machine learning. The use of a dataset without adequate features makes prediction impossible. Conversely, using all features may also be impossible since the amount of available training data in accordance dimensionality is small [1]. Even though feature selection tends to cause biases when handling missing data [2], it can handle uncorrelated or redundant features, which improves prediction performance [3]. There are two types of feature selection, filter and wrapper technique [4, 5]. Depending on the characteristic of data, the filter technique evaluates features without using any classification algorithms [6] and is utilized for high dimensional data [7]. However, the wrapper technique utilizes a specific classifier to evaluate the quality of the selected feature and its subset effect on the algorithm performance [4, 8].

According to [9], the most standard filters are based on their predictive power, which is approached by several means such as Fisher score [10], Chi-Square test [11], Laplacian score [12], Pearson correlation [13], or mutual information [14]. Conversely, wrapper feature selection is one of the most common and practical techniques [15]. The ant colony algorithm with an artificial neural network [16], a genetic algorithm with k-nearest neighbors [17]. Binary PSO and mutation algorithm with decision tree [18] are the example of the wrapper method in feature selection. Feature selection reduces the dimension by eliminating inappropriate or redundant features. It contributes to making more improvements in the learning accuracy of computational intelligence [19]. Furthermore, it is relatively significant because, with the same training data, it tends to perform better with different subsets [20].

Many researchers have developed new feature selection methods. The large margin hybrid algorithm for feature selection (LMFS) proposed by Zhang et al. [21] successfully overcome the over-fitting between the optimal feature subset and a given classifier. Yuan et al. [22] proposed partial maximum correlation information (PMCI) as a new feature selection method that delivers relatively good performance with lower time complexity than others. LW-index with the Sequence forward search algorithm (SFS-LW), proposed by Liu et al. [23] obtained similar accuracy as the wrapper method.

Meanwhile, Chiew et al. [24] proposed the hybrid ensemble feature selection (HEFS) as the feature selection for machine learning-based phishing detection system that is highly desirable and practical. There was also a method known as the curious feature selection (CFS) which is motivated by artificial curiosity and positively impacts the accuracy of the learning model [25]. Moreover, the possibility to improve and developed a new feature selection is still an appealing issue. The kernel function is known as the function that commonly used in the machine learning method to separate the data linearly when the data cannot be linearly separable. In this paper, therefore, introduces a new algorithm for feature selection based on kernel. K-means clustering [26] was used to examine its performance by calculated accuracy and F1-Score.

2. PROPOSED METHOD

This research introduces a new feature selection algorithm based on kernel with three steps: we calculate the mean of features, apply the kernel function, and sort the feature importance. Let $X = \{C_1, C_2, \dots, C_k\}$ is a set of k classes that consists of n samples of the dataset with f features in which $x = (x_1, x_2, \dots, x_f) \in C_k$ and $|C_k| = n_k$. From the above-listed values, the mean of each feature in every class is computed. It provides the sense to understand and obtain its representative value. Consider the mean of f features in the k-th class as a vector $m_k = (\overline{x_1}, \overline{x_2}, \dots, \overline{x_f})^t$. These k vectors are then used to construct K by F matrix $M = [m_1 \quad m_2 \quad \cdots \quad m_k]^t$.

After that, the kernel transformation is performed on every pair of mean vectors m_i, m_j where $i \neq j$ by projecting them into high dimensional feature space using the function as follows:

$$k(m_i, m_j) : XxX \to F \tag{1}$$

This research utilizes two kernel functions, namely Gaussian radial basis function (RBF) and polynomial kernel functions with several kernel parameters. The formulas are shown in (2)-(3).

$$\text{RBF kernel function: } k(m_i, m_j) = exp\left(-\frac{\|m_i - m_j\|^2}{2\sigma^2}\right) \tag{2}$$

$$\text{Polynomial kernel function: } k(m_i, m_j) = \left(m_i \cdot m_j + 1\right)^h \tag{3}$$

The result of this transformation is then stored as kernel matrix as given in (4):

$$K = \left[k(m_i, m_j)\right] = k_{ij} \tag{4}$$

In addition, the feature importance depends on this kernel matrix. Finally, the total entries of every row or the total number of kernel representation of the mean are computed. It is calculated using (5):

$$S_i = \sum_{j=1}^{k} k_{ij} \;, \quad i = 1, 2, \dots, f \tag{5}$$

Its value is then decreasingly sorted, which shows the order of features used represents the feature importance of the dataset. After that, the order of these features is considered in performing feature selection.

3. RESEARCH METHOD

3.1. Dataset

In these experiments, 16 real-world datasets from UCI data repository [27] are utilized to examine the performance of the proposed method with details summarized in Table 1.

Table 1. The real-world dataset characteristic

Dataset	Number of samples	Number of features
Iris	150	4
Thyroid disease	215	5
Credit score	100	6
Breast cancer Wisconsin (BCW) (Diagnostic)	569	30
Glass identification	214	9
Letter recognition	20000	16
Statlog (Landsat satellite)	6435	36
Wine	178	13
Statlog (Vehicle silhouettes)	946	18
Housing	506	13
Machine	209	6
Mammographic mass	961	5
Seismic-bumps	2584	18
Cardiotocography	2126	21
Forest type mapping	326	27
Image segmentation	2310	19

3.2. Algorithm

The new feature selection based on kernel consists of three steps: we calculate the mean of features, apply the kernel function, and sort the feature importance. The new feature selection algorithm based on kernel is given in Figure 1. This paper utilized only 75 percent of the first features after sorting the features which are used in the evaluation. K-means clustering, using 10-fold cross-validation is further used to examine the model by utilizing reduced features in the new feature selection algorithm. The k-means clustering algorithm is shown in Figure 2.

Input: $X = \{C_1, C_2, \dots, C_k\}$ where $x = (x_1, x_2, \dots, x_f) \in C_k$ and $|C_k| = n_k$
Output: sorted features

1. Calculate the mean of each class: $m_k = (\overline{x_1}, \overline{x_2}, \dots, \overline{x_f})^t$
2. Construct the matrix $M = [m_1 \quad m_2 \quad \cdots \quad m_k]^t$
3. Compute kernel matrix $K = [k(m_i, m_j)] = k_{ij}$ where $i \neq j$ and $k(m_i, m_j)$ is calculated based on the kernel type that was used.
4. Find the value $S_i = \sum_{j=1}^{k} k_{ij}$ with $i = 1, 2, \dots, f$, and sort this value decreasingly. The index of the sorted S_i is the index of features that will be first used.

End

Figure 1. Our new feature selection based on kernel algorithm

Input: $X = \{x_1, x_2, \dots, x_n\}, c, m_i, m_f, \varepsilon$, T (the maximum number of iterations allowed).
Output: $V = \{v_1, v_2, \dots v_c\}$, $R = [r_{ik}], 1 \leq i \leq n, 1 \leq k \leq c$.

1. Initialization: $V^0 = \{v_1, v_2, \dots v_c\}$
2. Compute the value of $\|x_i - v_j\|$
3. Update membership of the data point x_i in k^{th}-cluster according to: $r_{ik} = \begin{cases} 1 & , if\ k = \arg\min \|x_i - v_j\|^2 \\ 0 & , \text{otherwise} \end{cases}$
4. Update cluster center V^t using the equation below. $v_j^{(t)} = \frac{\sum_{i=1}^{n} r_{ij} x_i}{\sum_{i=1}^{n} r_{ij}}$
5. If $\|V^{(t-1)} - V^{(t)}\| < \varepsilon$ or $T = t$, then the iteration stops. Otherwise, $t = t + 1$ and go back to step 2;

End

Figure 2. K-means clustering algorithm

3.3. Performance metrics

In evaluating the performance of our new feature selection based on kernel, we utilize confusion matrix respect to the result of k-means clustering. The confusion matrix consists of four possible outcomes: true positives (TP), false negatives (FN), true negative (TN), and false positive (FP) [28]. If the positive instance is correctly predicted, it is counted as a true positive. If not, it is called a false negative. Then if the negative instance is correctly predicted, it is counted as true negative. If not, it is called a false positive [29].

In this paper, the confusion matrix is used to compute the performance metrics such as accuracy and F1-Score, where their formulas are as shown in (6)-(7):

$$\text{Accuracy} = \frac{\text{TP+TN}}{\text{TP+FN+TN+FP}} \quad (6)$$

$$\text{F1} - \text{Score} = \frac{2 * \text{sensitivity} * \text{precision}}{\text{sensitivity} + \text{precision}} \quad (7)$$

with sensitivity and precision is defined as given in (8)-(9):

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP+FN}} \quad (8)$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}} \quad (9)$$

4. RESULT AND DISCUSSION

4.1. The performance of our new feature selection based on RFB kernel function

In this section, the performance of k-means clustering was examined using the new feature selection based on RBF kernel function. Several kernel parameter σ were utilized with the analysis of the result based on each performance measurement, as shown in Table 2. This table shows that the method used has excellent performance almost in all real-world datasets, with the majority obtained when σ=1000 is used. In addition, the Machine dataset had the highest accuracy when σ=0.0001. The accuracy is constant for every value of the kernel parameter for several datasets. Moreover, F1-score performance is shown in Table 3.

Table 2. The accuracy performance of our method on the real-world datasets using RBF kernel

Dataset	Kernel parameter of RBF kernel function									
	0.0001	0.001	0.05	0.1	1	5	10	50	100	1000
Iris	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**
Thyroid disease	98.14	98.32	98.38	98.42	98.43	98.45	98.45	98.46	**98.47**	**98.47**
Credit score	94.44	96.11	96.67	96.94	97.11	97.22	97.30	97.36	97.41	**97.44**
Breast cancer Wisconsin (Diagnostic)	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**
Glass identification	94.29	94.76	94.92	95.00	95.05	95.08	95.10	95.12	95.13	**95.14**
Letter recognition	97.19	98.32	98.76	98.99	99.13	99.22	99.28	99.33	99.37	**99.40**
Statlog (Landsat satellite)	91.62	92.64	93.04	93.25	93.37	93.46	93.52	93.56	93.60	**93.63**
Wine	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**
Statlog (Vehicle silhouettes)	87.22	87.33	87.37	87.39	87.40	87.41	87.41	**87.42**	**87.42**	**87.42**
Housing	85.42	85.52	85.55	85.57	85.58	85.59	85.59	85.59	**85.60**	**85.60**
Machine	**85.46**	85.30	85.25	85.22	85.21	85.19	85.19	85.18	85.18	85.17
Mammographic mass	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**
Seismic-bumps	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**
Cardiotocography	88.80	89.74	90.05	90.21	90.30	90.37	90.41	90.44	90.47	**90.49**
Forest type mapping	90.07	92.80	93.79	94.30	94.60	94.80	94.95	95.06	95.14	**95.21**
Image segmentation	96.42	97.29	97.64	97.81	97.92	97.99	98.04	98.08	98.11	**98.14**

Table 3. The F1-Score performance of our method on the real-world datasets using RBF kernel

Dataset	Kernel parameter of RBF kernel function									
	0.0001	0.001	0.05	0.1	1	5	10	50	100	1000
Iris	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**
Thyroid disease	98.89	99.00	99.04	99.05	99.06	99.07	99.08	99.08	99.08	**99.09**
Credit score	88.37	91.57	92.68	93.25	93.60	93.83	93.99	94.12	94.21	**94.29**
Breast cancer Wisconsin (Diagnostic)	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**
Glass identification	96.25	96.55	96.65	96.70	96.73	96.75	96.77	96.78	**96.79**	**96.79**
Letter recognition	97.94	98.57	98.89	99.07	99.18	99.26	99.32	99.36	99.39	**99.42**
Statlog (Landsat satellite)	92.26	92.93	93.21	93.36	93.45	93.52	93.56	93.60	93.62	**93.65**
Wine	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**
Statlog (Vehicle silhouettes)	84.96	85.14	85.20	85.23	85.25	85.26	85.27	85.28	85.28	**85.29**
Housing	84.68	84.92	85.01	85.05	85.07	85.09	85.10	85.11	**85.12**	**85.12**
Machine	**83.78**	83.60	83.54	83.50	83.49	83.47	83.46	83.46	83.45	83.45
Mammographic mass	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**
Seismic-bumps	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**
Cardiotocography	90.74	91.28	91.47	91.57	91.62	91.66	91.69	91.71	91.73	**91.74**
Forest type mapping	92.08	93.53	94.14	94.46	94.67	94.80	94.90	94.98	95.04	**95.09**
Image segmentation	96.78	97.45	97.72	97.87	97.96	98.02	98.06	98.09	98.12	**98.14**

As the measurement that concerns equally in sensitivity and precision, the F1-Score performance of our method was also excellent. The best performance was obtained when kernel parameter σ=1000 used. In addition, to the performance metrics above, the running time also was evaluated, and its result is summarized in Table 4. The result of the running time, which is calculated in second, varies regarding the value of kernel parameter. Except for the Letter Recognition dataset, the algorithm performs fast for almost all of the datasets.

Table 4. The running time performance of our method on the real-world datasets using RBF kernel

Dataset	Kernel parameter of RBF kernel function									
	0.0001	0.001	0.05	0.1	1	5	10	50	100	1000
Iris	0.13	0.14	**0.11**	**0.11**	0.13	**0.11**	0.16	0.13	**0.11**	**0.11**
Thyroid disease	0.25	0.23	**0.22**	**0.22**	0.25	0.23	0.33	**0.22**	**0.22**	**0.22**
Credit score	0.05	0.06	0.05	0.05	**0.03**	0.05	0.06	0.06	0.05	0.06
Breast cancer Wisconsin (Diagnostic)	**1.30**	1.31	**1.30**	1.33	1.34	1.31	1.33	1.36	1.50	1.73
Glass identification	0.23	0.25	0.27	**0.22**	**0.22**	**0.22**	0.27	0.23	**0.22**	0.27
Letter recognition	297.31	317.42	344.06	310.23	296.70	317.45	**279.84**	280.25	280.98	280.41
Statlog (Landsat satellite)	11.05	10.95	11.02	11.00	11.05	11.13	11.39	11.34	11.03	**10.94**
Wine	**0.13**	**0.13**	0.14	**0.13**	0.17	**0.13**	0.14	**0.13**	**0.13**	**0.13**
Statlog (Vehicle silhouettes)	3.22	3.22	3.19	3.23	3.22	3.22	3.36	3.44	3.25	**3.16**
Housing	1.13	1.09	1.20	1.11	1.09	**1.08**	1.09	1.11	1.20	1.13
Machine	0.19	**0.17**	**0.17**	**0.17**	**0.17**	**0.17**	0.20	**0.17**	0.19	**0.17**
Mammographic mass	1.83	1.84	**1.81**	**1.81**	1.84	1.86	**1.81**	1.83	**1.81**	1.88
Seismic-bumps	1.16	1.13	1.13	1.14	1.14	1.13	1.13	1.13	1.16	**1.11**
Cardiotocography	2.92	2.92	2.91	2.98	**2.84**	2.91	2.95	2.91	2.89	**2.84**
Forest type mapping	1.27	1.28	1.30	1.25	**1.23**	1.31	1.30	1.27	1.28	1.28
Image segmentation	22.13	22.14	22.13	22.42	22.31	22.14	22.17	**22.09**	22.19	22.39

4.2. The performance of our new feature selection based on polynomial kernel function

After evaluating the new feature selection performance using RBF kernel function, the new feature selection based on the polynomial kernel function in this section all evaluates the accuracy, F1-Score, and running time. The accuracy performance is shown in Table 5. Opposite with the RBF kernel function, the accuracy performance of the new feature selection based on polynomial kernel is not affected by the polynomial degree.

Table 5. The accuracy performance of our method on the real-world datasets using polynomial kernel

Dataset	Kernel parameter of polynomial kernel function									
	1	2	3	4	5	6	7	8	9	10
Iris	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**	**98.00**
Thyroid disease	**98.51**	**98.51**	**98.51**	**98.51**	**98.51**	**98.51**	**98.51**	**98.51**	**98.51**	**98.51**
Credit score	**97.78**	**97.78**	**97.78**	**97.78**	**97.78**	**97.78**	**97.78**	**97.78**	**97.78**	**97.78**
Breast cancer Wisconsin (Diagnostic)	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**	**90.00**
Glass identification	**95.24**	**95.24**	**95.24**	**95.24**	**95.24**	**95.24**	**95.24**	**95.24**	**95.24**	**95.24**
Letter recognition	**99.69**	**99.69**	**99.69**	**99.69**	**99.69**	**99.69**	**99.69**	**99.69**	**99.69**	**99.69**
Statlog (Landsat satellite)	**93.89**	**93.89**	**93.89**	**93.89**	**93.89**	**93.89**	**93.89**	**93.89**	**93.89**	**93.89**
Wine	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**	**91.12**
Statlog (Vehicle silhouettes)	**87.45**	**87.45**	**87.45**	**87.45**	**87.45**	**87.45**	**87.45**	**87.45**	**87.45**	**87.45**
Housing	**85.62**	**85.62**	**85.62**	**85.62**	**85.62**	**85.62**	**85.62**	**85.62**	**85.62**	**85.62**
Machine	**85.14**	**85.14**	**85.14**	**85.14**	**85.14**	**85.14**	**85.14**	**85.14**	**85.14**	**85.14**
Mammographic mass	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**	**75.33**
Seismic-bumps	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**	**80.85**
Cardiotocography	**90.68**	**90.68**	**90.68**	**90.68**	**90.68**	**90.68**	**90.68**	**90.68**	**90.68**	**90.68**
Forest type mapping	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**	**95.83**
Image segmentation	**98.35**	**98.35**	**98.35**	**98.35**	**98.35**	**98.35**	**98.35**	**98.35**	**98.35**	**98.35**

Meanwhile, F1-Score considers both sensitivity and precision are also similar for every polynomial degree, as shown in Table 6. Table 7 demonstrates the running time performance the method utilized. In addition, it still needs a long time for letter recognition dataset but performs well for other datasets. The performance also varied according to the polynomial degree used.

Table 6. The F1-Score performance of our method on the real-world datasets using polynomial kernel

Dataset	Kernel parameter of polynomial kernel function									
	1	2	3	4	5	6	7	8	9	10
Iris	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**	**98.04**
Thyroid disease	**99.11**	**99.11**	**99.11**	**99.11**	**99.11**	**99.11**	**99.11**	**99.11**	**99.11**	**99.11**
Credit score	**95.00**	**95.00**	**95.00**	**95.00**	**95.00**	**95.00**	**95.00**	**95.00**	**95.00**	**95.00**
Breast cancer Wisconsin (Diagnostic)	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**	**87.71**
Glass identification	**96.86**	**96.86**	**96.86**	**96.86**	**96.86**	**96.86**	**96.86**	**96.86**	**96.86**	**96.86**
Letter recognition	**99.68**	**99.68**	**99.68**	**99.68**	**99.68**	**99.68**	**99.68**	**99.68**	**99.68**	**99.68**
Statlog (Landsat satellite)	**93.85**	**93.85**	**93.85**	**93.85**	**93.85**	**93.85**	**93.85**	**93.85**	**93.85**	**93.85**
Wine	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**	**91.66**
Statlog (Vehicle silhouettes)	**85.32**	**85.32**	**85.32**	**85.32**	**85.32**	**85.32**	**85.32**	**85.32**	**85.32**	**85.32**
Housing	**85.18**	**85.18**	**85.18**	**85.18**	**85.18**	**85.18**	**85.18**	**85.18**	**85.18**	**85.18**
Machine	**83.41**	**83.41**	**83.41**	**83.41**	**83.41**	**83.41**	**83.41**	**83.41**	**83.41**	**83.41**
Mammographic mass	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**	**75.74**
Seismic-bumps	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**	**73.05**
Cardiotocography	**91.86**	**91.86**	**91.86**	**91.86**	**91.86**	**91.86**	**91.86**	**91.86**	**91.86**	**91.86**
Forest type mapping	**95.54**	**95.54**	**95.54**	**95.54**	**95.54**	**95.54**	**95.54**	**95.54**	**95.54**	**95.54**
Image segmentation	**98.33**	**98.33**	**98.33**	**98.33**	**98.33**	**98.33**	**98.33**	**98.33**	**98.33**	**98.33**

Table 7. The running time performance of our method on the real-world datasets using polynomial kernel

Dataset	Kernel parameter of polynomial kernel function									
	1	2	3	4	5	6	7	8	9	10
Iris	0.14	0.16	**0.11**	0.13	0.13	**0.11**	0.16	0.14	**0.11**	0.16
Thyroid disease	0.23	0.23	**0.22**	0.22	0.30	0.27	0.23	0.25	0.25	**0.22**
Credit score	0.06	0.06	0.08	**0.05**	**0.05**	**0.05**	0.06	**0.05**	**0.05**	0.06
Breast cancer Wisconsin (Diagnostic)	**1.33**	1.56	1.34	1.50	1.34	1.50	1.34	1.56	1.34	1.42
Glass identification	0.27	0.25	0.25	0.25	0.25	**0.23**	0.25	**0.23**	**0.23**	0.25
Letter recognition	331.53	340.69	330.16	328.73	354.52	**327.86**	332.80	328.75	329.16	341.53
Statlog (Landsat satellite)	**12.48**	12.67	12.66	13.25	12.73	12.52	14.05	13.36	13.19	13.91
Wine	0.16	**0.13**	**0.13**	0.16	**0.13**	0.19	0.17	0.16	**0.13**	0.16
Statlog (Vehicle silhouettes)	3.48	3.56	3.52	**3.39**	3.42	3.44	3.42	3.56	3.66	3.69
Housing	1.20	1.22	1.23	**1.17**	1.19	1.25	**1.17**	**1.17**	**1.17**	**1.17**
Machine	**0.19**	0.20	**0.19**	0.20	0.20	**0.19**	**0.19**	**0.19**	0.28	0.28
Mammographic mass	**1.88**	1.91	2.16	2.08	2.19	1.92	1.92	1.92	1.98	1.97
Seismic-bumps	1.19	1.91	1.19	**1.17**	1.22	1.22	1.22	1.20	1.20	1.25
Cardiotocography	3.17	3.81	3.11	3.06	3.09	3.20	3.27	**3.06**	**3.06**	3.08
Forest type mapping	1.38	**1.34**	1.39	1.50	1.38	1.38	1.56	1.50	1.38	1.36
Image segmentation	23.16	25.02	24.20	23.14	23.41	22.98	23.66	**22.75**	23.84	24.44

4.3. The comparison performance of our new feature selection based on RBF and polynomial kernel function with several other feature selection methods

In this section, the performance metrics that consist of accuracy and F1-Score is compared with the RBF and polynomial kernel function. From each dataset, their performance is extracted which delivers the best value. In the case of the polynomial kernel function that performs similarly for every polynomial degree, we choose the polynomial degree that performs faster in the running time. The comparison associated with the new feature selection is based on RBF and polynomial kernel function for every dataset. The performance of the proposed feature selection algorithm was also compared with the other well-established feature selection methods, such as Fisher score [10], Chi-Square test [11], and Laplacian score [12], as shown in Table 8.

From Table 8, it can be concluded that both kernel functions perform similarly in almost every dataset that was evaluated. The running time is slower when using the polynomial kernel function. However, the polynomial kernel function is higher in the performance of accuracy and F1-Score than RBF. Compared to the Fisher score, Chi-Square test, and Laplacian score algorithm as the feature selection, our proposed method was delivered higher accuracy and F1-Score until 40 percent difference, for example in the Credit Score, Letter Recognition, Statlog (Landsat Satellite), Forest Type Mapping, and Image Segmentation dataset.

Table 8. The comparison of the proposed method with Fisher's score, Chi-Square Test, and Laplacian score algorithm

Dataset	Feature selection method	Accuracy (%)	F1-Score (%)	Running time (s)
Iris	New feature selection based on RBF kernel function with σ = 0.05	98.00	98.04	**0.11**
	New feature selection based on 3rd polynomial kernel function	98.00	98.04	**0.11**
	Fisher Score	**100.00**	**100.00**	0.17
	Chi-Square Test	**100.00**	**100.00**	0.22
	Laplacian Score	**100.00**	**100.00**	7.03
Thyroid disease	New feature selection based on RBF kernel function with σ = 1000	98.47	99.09	**0.22**
	New feature selection based on 3rd polynomial kernel function	98.51	99.11	**0.22**
	Fisher Score	**100.00**	**100.00**	0.28
	Chi-Square Test	**100.00**	**100.00**	0.36
	Laplacian Score	**100.00**	**100.00**	3.42
Credit score	New feature selection based on RBF kernel function with σ = 1000	97.44	94.29	0.06
	New feature selection based on 4th polynomial kernel function	97.78	95.00	0.05
	Fisher Score	98.81	98.60	**0.03**
	Chi-Square Test	98.81	98.60	0.06
	Laplacian Score	**100.00**	**100.00**	1.86
Breast cancer Wisconsin (Diagnostic)	New feature selection based on RBF kernel function with σ = 0.0001	**90.00**	87.71	1.30
	New feature selection based on 1st polynomial kernel function	**90.00**	87.71	1.33
	Fisher Score	87.50	**90.91**	**0.20**
	Chi-Square Test	87.50	**90.91**	0.25
	Laplacian Score	88.24	86.36	1.34
Glass identification	New feature selection based on RBF kernel function with σ = 1000	95.14	96.79	0.27
	New feature selection based on 6th polynomial kernel function	95.24	96.86	**0.23**
	Fisher Score	98.46	98.26	1.03
	Chi-Square Test	98.46	98.26	1.19
	Laplacian Score	**100.00**	**100.00**	13.09
Letter recognition	New feature selection based on RBF kernel function with σ = 1000	99.40	99.42	280.41
	New feature selection based on 6th polynomial kernel function	**99.69**	**99.68**	327.86
	Fisher Score	99.64	99.64	32.50
	Chi-Square Test	99.39	99.39	**28.98**
	Laplacian Score	99.31	99.30	313.06
Statlog (Landsat satellite)	New feature selection based on RBF kernel function with σ = 1000	93.63	93.65	10.94
	New feature selection based on 1st polynomial kernel function	**93.89**	**93.85**	12.48
	Fisher Score	33.33	50.00	**0.14**
	Chi-Square Test	66.67	75.00	0.17
	Laplacian Score	80.95	75.00	1.97
Wine	New feature selection based on RBF kernel function with σ = 0.0001	91.12	91.66	**0.13**
	New feature selection based on 2nd polynomial kernel function	91.12	91.66	**0.13**
	Fisher Score	**100.00**	**100.00**	0.27
	Chi-Square Test	**100.00**	**100.00**	0.36
	Laplacian Score	**100.00**	**100.00**	3.95
Statlog (Vehicle silhouettes)	New feature selection based on RBF kernel function with σ = 1000	87.42	85.29	3.16
	New feature selection based on 5th polynomial kernel function	87.45	85.32	3.42
	Fisher Score	85.00	87.78	**0.38**
	Chi-Square Test	**89.66**	**88.86**	**0.38**
	Laplacian Score	87.47	83.32	5.08
Housing	New feature selection based on RBF kernel function with σ = 1000	85.60	85.12	**1.13**
	New feature selection based on 4th polynomial kernel function	85.62	85.18	1.17
	Fisher Score	96.97	95.24	1.19
	Chi-Square Test	95.15	95.56	1.33
	Laplacian Score	**98.74**	**99.14**	14.72
Machine	New feature selection based on RBF kernel function with σ = 0.0001	85.46	83.78	**0.19**
	New feature selection based on 3rd polynomial kernel function	85.14	83.41	**0.19**
	Fisher Score	94.77	94.42	0.52
	Chi-Square Test	94.41	94.20	0.52
	Laplacian Score	**98.10**	**98.09**	6.97
Mammographic mass	New feature selection based on RBF kernel function with σ = 0.05	**75.33**	**75.74**	1.81
	New feature selection based on 1st polynomial kernel function	**75.33**	**75.74**	1.88
	Fisher Score	66.67	75.00	**0.17**
	Chi-Square Test	50.00	66.67	**0.17**
	Laplacian Score	71.67	74.63	1.81
Seismic-bumps	New feature selection based on RBF kernel function with σ = 1000	80.85	73.05	1.11
	New feature selection based on 3rd polynomial kernel function	80.85	73.05	1.19
	Fisher Score	72.73	80.00	**0.25**
	Chi-Square Test	**90.91**	**94.12**	0.30
	Laplacian Score	78.13	82.93	2.02
Cardiotocography	New feature selection based on RBF kernel function with σ = 1000	90.49	91.74	2.84
	New feature selection based on 4th polynomial kernel function	90.68	91.86	3.06
	Fisher Score	89.56	85.86	**0.59**
	Chi-Square Test	**96.67**	**95.24**	0.61
	Laplacian Score	92.09	91.67	4.38
Forest type mapping	New feature selection based on RBF kernel function with σ = 1000	95.21	95.09	1.28
	New feature selection based on 2nd polynomial kernel function	95.83	95.54	1.34
	Fisher Score	96.63	96.30	**0.59**
	Chi-Square Test	94.89	93.45	**0.59**
	Laplacian Score	**100.00**	**100.00**	9.33
Image segmentation	New feature selection based on RBF kernel function with σ = 1000	98.14	98.14	22.39
	New feature selection based on 8th polynomial kernel function	98.35	98.33	22.75
	Fisher Score	**100.00**	**100.00**	2.20
	Chi-Square Test	98.41	98.10	**2.14**
	Laplacian Score	98.72	98.67	22.53

5. CONCLUSION

Feature selection is a crucial issue in machine learning, which makes users refuse to use the redundant features not correlated to the target of class in the dataset. There are two types of feature selection; however, it tends to filter, wrapper, or ensemble of both. In this paper, a new feature selection based on kernel function was introduced and applied to 16 real-world datasets from UCI data repository. K-means clustering was utilized as the classifier and only used 75 percent of the number of features that were sorted using this method. The performance was evaluated using RBF and polynomial kernel function with 10-fold cross-validation used to determine its accuracy and F1-Score as the performance comparison. The running time was also examined as consideration and analyzed.

From the experiments, it is concluded that when the new feature selection uses RBF kernel function, the performances varied according to the value of kernel parameter σ. The majority performed its best when using the kernel parameter σ=1000, while the feature selection based on polynomial kernel function was not affected by the use of the value of polynomial degree. In conclusion, the new feature selection based on RBF kernel function has a faster running time compared to the polynomial kernel function. For future work, the invention of new feature selection is still widely accessible for development. Other kernel functions and the evaluation techniques can be used for comparison. Moreover, utilize other classifiers can also be considered.

ACKNOWLEDGEMENTS

This research was financially supported by the University of Indonesia, with DRPM PUTI Q2 2020 research grant scheme. The authors also would like to express gratitude for all the reviewers involved in the improvement of this article

REFERENCES

[1] F. Benoit, M. V. Heeswijk, Y. Miche, M. Verleysen, and A. Lendasse, “Feature selection for nonlinear models with extreme learning machines,” *Neurocomputing*, vol. 102, pp. 111-114, Feb. 2013

[2] B. Seijo-Pardo, A. Alonso-Betanzos, K. P. Bennett, V. Bolón-Canedo, J. Josse, M. Saeed, and I. Guyon, “Biases in feature selection with missing data,” *Neurocomputing*, vol. 342, pp. 97-112, May 2019

[3] P. Drotár, J. Gazda, and Z. Smékal, “An experimental comparison of feature selection methods on two-class biomedical datasets,” *Computers in Biology and Medicine*, vol. 66, pp. 1-10, Nov 2015.

[4] R. Kohavi and G. H. John, “Wrappers for feature subset selection,” *Artificial Intelligence*, vol. 97, pp. 273-324, 1997.

[5] S. Das, “Filters, wrappers and a boosting-based hybrid for feature selection,” *Proceedings of the 18th International Conference on Machine Learning*, pp. 74-81, 2001.

[6] H. Liu and H. Motoda, “Computational methods of feature selection: First edition,” *Chapman and Hall/CRC Press*, Taylor and Francis Group, pp. 1-33, 2007.

[7] A. A. Yahya, “Feature selection for high dimensional data: An evolutionary filter approach,” *Journal of Computer Science*, vol. 7, pp. 800-820, May 2011.

[8] J. Tang, S. Alelyani, and H. Liu, “Feature selection for classification: A review,” *Data Classification: Algorithms and Applications, CRC Press*, pp. 37-64, 2014.

[9] F. Fleuret, “Fast binary feature selection with conditional mutual information,” *Journal of Machine Learning Research*, vol. 5, pp. 1531-1555, 2004.

[10] R. O. Duda, P. E. Hart, and D. G. Stork, “Pattern classification: Second edition,” *Wiley-Interscience Publication*, 2001.

[11] X. Jin, A. Xu, R. Bie, and P. Guo, “Machine learning techniques and chi-square feature selection for cancer classification using SAGE gene expression profiles,” *In: Li J., Yang Q., Tan AH. (eds) Data Mining for Biomedical Applications. BioDM 2006. Lecture Notes in Computer Science*, vol. 3916, pp. 106-115, 2006.

[12] X. He, D. Cai, and P. Niyogi, “Laplacian score for feature selection,” *Proceedings of the 18th International Conference on Neural Information Processing Systems*, vol. 18, pp. 507-514, 2005.

[13] K. Miyahara and M. J. Pazzani, “Collaborative filtering with the simple Bayesian classifier,” *In: Mizoguchi R, Slaney J. Editors. PRICAI 2000 Topics in Artificial Intelligence*, vol. 1886, pp. 679-689, 2000.

[14] R. Battiti, “Using mutual information for selecting features in supervised neural net learning,” *IEEE Transactions on Neural Networks*, vol. 5, no. 4, pp. 537-550, July 1994.

[15] B. Waad, “New approach for wrapper feature selection using genetic algorithm for Big Data,” *In: Lavangnananda K., Phon-Amnuaisuk S., Engchuan W., Chan J. (eds) Intelligent and Evolutionary Systems. Proceedings in Adaptation, Learning and Optimization*, vol. 5, pp. 75-83, 2016.

[16] A. Al-Ani, “Ant colony optimization for feature subset selection,” *The Second World Enformatika Conference*, vol. 1, pp. 35-38, 2005.

[17] T. Jirapech-Umpai and S. Aitken, “Feature selection and classification for microarray data analysis: Evolutionary methods for identifying predictive genes,” *BMC Bioinformatics*, vol. 6, no. 1, pp. 1-11, June 2005.

[18] Y. Zhang, S. Wang, and P. Phillips, “Binary PSO with mutation operator for feature selection using decision tree applied to spam detection,” *Knowledge-Based Systems*, vol. 64, pp. 22-31, 2014.

[19] K. Jain, “A survey on feature selection techniques,” *International Journal of Innovations in Engineering Research and Technology*, vol. 4, no. 5, pp. 1-4, 2017.
[20] A. Blum and P. Langley, “Selection of relevant features and examples in machine learning,” *Artificial Intelligence*, vol. 97, pp. 245-271, 1997.
[21] J. Zhang, Y. Xiong, and S. Min, “A new hybrid filter/wrapper algorithm for feature selection in classification,” *Analytica Chimica Acta*, vol. 1080, pp. 43-54, 2019.
[22] M. Yuan, Z. Yang, and G. Ji, “Partial maximum correlation information: A new feature selection method for microarray data classification,” *Neurocomputing*, vol. 323, pp. 231-243, 2019.
[23] C. Liu, W. Wang, Q. Zhao, X. Shen, and M. Konan, “A new feature selection method based on a validity index of feature subset,” *Pattern Recognition Letters*, vol. 92, pp. 1-8, 2017.
[24] K. L. Chiew, C. L. Tan, K. Wong, K. S. C. Yong, and W. K. Tiong, “A new hybrid ensemble feature selection framework for machine learning-based phishing detection system,” *Information Sciences*, vol. 484, pp. 153-166, 2019.
[25] M. Moran and G. Gordon, “Curious feature selection,” *Information Sciences*, vol. 485, pp. 42-54, 2019.
[26] S. P. Lloyd “Least squares quantization in PCM,” *IEEE Trans. on Infor. Theory*, vol. 28, no. 2, pp. 129-137, 1982.
[27] D. Dua and C. Graff, “UCI Machine Learning Repository,” University of California, School of Information and Computer Science, [Online], Available at: http://archive.ics.uci.edu/ml/index.php.
[28] D. M. W. Powers, “Evaluation: from precision, recall, and f-measure to ROC, informedness, markedness & correlation,” *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37-63, 2011.
[29] T. Fawcett, “An introduction to ROC analysis,” *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861-874, 2006.

BIOGRAPHIES OF AUTHORS

Zuherman Rustam is an Associate Professor and a lecturer of the intelligence computation at the Department of Mathematics, University of Indonesia. He obtained his Master of Science in 1989 in informatics, Paris Diderot University, French, and completed his Ph.D. in 2006 from computer science, University of Indonesia.
Assoc. Prof. Dr. Rustam is a member of IEEE who is actively researching machine learning, pattern recognition, neural network, artificial intelligence.

Sri Hartini is a Bachelor of Science from the Department of Mathematics, University of Indonesia, who is also completing the Master of Science at the University of Indonesia and is currently pursuing a Ph.D. in intelligence computation.
Ms. Hartini is passionately researching machine learning, computer vision, neural networks, and deep learning in various fields.