Soft Computing (2021) 25:6391–6400
https://doi.org/10.1007/s00500-021-05630-7

METHODOLOGIES AND APPLICATION

A noise-aware feature selection approach for classification

Mostafa Sabzekar[1] · Zafer Aydin[2]

Accepted: 26 January 2021 / Published online: 17 February 2021

Abstract
A noise-aware version of support vector machines is utilized for feature selection in this paper. Combining this method and sequential backward search (SBS), a new algorithm for removing irrelevant features is proposed. Although feature selection methods in the literature which utilize support vector machines have provided acceptable results, noisy samples and outliers may affect the performance of SVM and feature selections method, consequently. Recently, we have proposed relaxed constrains SVM (RSVM) which handles noisy data and outliers. Each training sample in RSVM is associated with a degree of importance utilizing the fuzzy c-means clustering method. Therefore, a less importance degree is assigned to noisy data and outliers. Moreover, RSVM has more relaxed constraints that can reduce the effect of noisy samples. Feature selection increases the accuracy of different machine learning applications by eliminating noisy and irrelevant features. In the proposed RSVM-SBS feature selection algorithm, noisy data have small effect on eliminating irrelevant features. Experimental results using real-world data verify that RSVM-SBS has better results in comparison with other feature selection approaches utilizing support vector machines.

Keywords Feature selection · Noisy data · Importance degree · Sequential backward search

1 Introduction

Feature selection (Blum and Langley 1997; Guyon and Elisseeff 2003) is a critical step in knowledge discovery applications and is considered as one of the essential steps in data mining (Maldonado et al. 2011; Dash and Liu 1997; He and Wu 2011). The goal of feature selection is to provide a reduced subset of features from a data set to simplify the model for easier interpretation as well as enhance the model's generalizability by reducing overfitting. It also tries to solve the curse of dimensionality problem by ignoring irrelevant features during the construction of a model for prediction or classification. This is computationally intractable due to its combinatorial nature in the number of primary features. Therefore, a direct result of choosing the best and smallest number of features is that we can reduce the time of building the model, significantly, without any changes or even may be with better performance. Consequently, feature selection provides the following benefits: (1) it decreases the computational time for building the model, (2) it tries to avoid overfitting, (3) it enhances the generalizability of the model, (4) it improves the model accuracy by removing misleading features.

The feature selection method studies generally can be grouped into the following categories (Shieh and Yang 2008; Liu and Zheng 2006):

a) First them out, *filter methods* identify less informative features in a data set and filter out them by considering their statistical properties. This approach typically solves the problem in two steps. First, a criterion is used to rank each feature and then the high ranked features are selected for model construction. For example, Duda and Stork (2001) introduced an importance score for each feature by calculating its correlation to the class labels for two-class classification problems as follows:

✉ Mostafa Sabzekar
sabzekar@birjandut.ac.ir

[1] Department of Computer Engineering, Birjand University of Technology, Birjand, Iran

[2] Department of Computer Engineering, Abdullah Gül University, Kayseri, Turkey

$$F(j) = \left| \frac{m_j^+ - m_j^-}{\left(s_j^+\right)^2 + \left(s_j^-\right)^2} \right|, \quad (1)$$

where m_j^+(m_j^-) and s_j^+(s_j^-) are the mean and the standard deviation values for jth feature in the first (second) class, respectively. Yan et al. (2019) suggested a filter-based method for selecting the best features on biomedical data using an improved Coral Reefs Optimization (CRO) algorithm. A new criterion for each feature based on information theory was proposed in Hancer et al. (2018). The proposed approach used a nearest neighbor strategy and also Fisher score, and then high ranked features were chosen as the best subset. In another study (Zheng and Wang 2018), information entropy was utilized to score each feature of data set. Finally, it should be noted that the efficiency of a filter-based method is directly dependent on choosing the scoring function which represents the relevance between each feature and the class labels. In the literature, information gain, principal component analysis (PCA), statistical approaches are used for this purpose more than other approaches.

b) The *wrapper methods*, instead, search the feature space and evaluate each subset by a classifier. In general, evolutionary methods are utilized to search the feature space. The methods in this category typically have better accuracy than filter-based methods, but they usually are computationally more complex. Mafarja and Mirjalili (2018) proposed a novel Whale optimization algorithm (WOA) with tournament selection and utilized it for feature selection. In another study Mafarja et al. (2018), a search strategy utilizing a novel Grasshopper optimization algorithm (GOA) is proposed for feature selection approach. They utilized KNN classifiers for evaluating each possible solution. In Zakeri and Hokmabadi (2019) a real-valued GOA was proposed to solve the problem. The method was evaluated on available data sets with various dimensionalities, and it has shown better performance in 7 out of 10 data sets. The most efforts in this category lie on utilizing and developing new evolutionary methods to search the feature space more efficiently.

c) Finally, *embedded methods* try to take the advantages of two other categories. Filter-based approaches do not incorporate the learning phase. Instead, wrapper-based approaches utilize a learning classifier to assess the overall quality of a chosen feature subset. On the contrary of two mentioned categories, embedded methods do not consider the feature selection and the learning as separate steps and choose the best features during the model construction. For example, Lu (2019) proposed an embedded method from sparse learning perspective for unknown data heterogeneity. In another work, Faris et al. (2019) proposed an embedded feature selection method for a spam detection application. They utilized an intelligent approach based on Genetic Algorithm (GA) and Random Weight Network (RWN).

The main contribution of this paper is to investigate the effect of applying a noise-aware version of SVM (our proposed RSVM) on feature selection. The details of RSVM and the proposed method will be discussed in the next sections. Further discussions about RSVM appeared in Sabzekar et al. (2011).

The rest of this paper is organized as follows. In the next section, we briefly review the main published papers in the literature. The structure of SVM and RSVM followed by the proposed RSVM-SBS algorithm is discussed in Sect. 3. The assessment of the proposed method is given in Sect. 4. Section 5 contains some conclusions and discussions about the paper.

2 Literature review

Many efforts are performed by researchers to improve the performance of feature selection. Among various algorithms, feature selection using support vector machine (SVM) approaches has proven their effectiveness not only for classification tasks (Pławiak et al. 2019; Abdar and Makarenkov 2019) but also as feature selection methods (Neumann et al. 2005; Mundra and Rajapakse 2010; Torres-Valencia et al. 2017; Abdoos et al. 2016). In these methods no assumption is made about data distribution. The rest of this section reviews some studies which utilize SVMs for feature selection. The main advantages of SVMs, in comparison with other classification methods, that motivate the researchers to utilize them for feature selection are (Vapnik 1998): good generalization ability, escaping from the local minimum and few parameters in training.

Maldonado et al. (2011) introduced a novel embedded feature ranking and selection method which determines the best feature subset without any validation step. The main idea of their method is to optimize the kernel function of SVM and at the same time select the best feature subset. However, the method suffers from tuning several parameters.

In another study (Xia and Hu 2006), the authors tried to decrease the effect of noise and outliers in data utilizing fuzzy support vector machines (FSVMs). The main idea of

FSVM is to assign a fuzzy membership to each training sample based on a predefined membership function.

Another embedded method for SVM-based feature selection was proposed in Benítez-Peña et al. (2019). This method simultaneously maximizes the margin and minimizes the selected features. False-positive and false-negative rates were also imposed as upper bonds.

Hold-out SVM (HO-SVM) (Maldonado and Weber 2009), a wrapper-based feature selection, used sequential backward selection (SBS) approach to identify and eliminate the least informative feature in each step using error rate of SVM in validation phase.

A combination of a metaheuristic method, namely Taguchi genetic algorithm (TGA), and FSVMS was proposed for solving the feature selection problem in Tang (2010). On the one hand, TGA tries to produce better chromosomes, and in another hand, the authors suggested a new fuzzy membership function for FSVM to achieve more accurate classification results. Each feature set was evaluated using this new classifier as a wrapper scheme. In a similar study (Xiong and Wang 2008), the authors investigated the effect of ant colony optimization (ACO) and support vector machines for selecting the minimal set of features.

Zaman and Karray (2009) proposed an enhanced version of SVMs for feature selection in intrusion detection systems. First, the proposed method gives a weight to each input feature which is considered as a rank value. Then, correlation between the features is calculated by either forward selection or backward elimination.

Aljarah et al. (2018) made an effort to introduce a hybrid method that optimizes the SVM parameters such as kernel parameters and simultaneously finds suitable number of features. To achieve this goal, Grasshopper optimization algorithm was adopted.

Maldonado and López (2018) introduced a feature selection method based on support vector machines that can be applied to class-imbalanced data sets. Their feature selection approach is based on gradient descent for feature penalization. To deal with the class-imbalance problem, their method was combined with cost-sensitive SVM and also support vector data description (SVDD).

Although SVM-based feature selection provides a promising and powerful algorithm, the noisy input features and outliers may cause negative effects on its performance. The main reason is that the optimal separating hyperplane (OSH) in SVM is determined by support vectors that include only a small part of the training samples (Sabzekar and Naghibzadeh 2013a). Consequently, the performance of feature selection method is directly affected by this matter. In this paper we used a noise-aware version of SVMs, namely relaxed constraints SVMs (RSVMs) (Sabzekar et al. 2011), for feature selection. It considers noisy data and gives less degrees of importance to them. Furthermore, it has proven its capability in the presence of noisy and outlier samples in the input space and introduced two new concepts, namely tolerance and uncertainty in a given data set.

3 The proposed method

In this section, we describe the details of our proposed RSVM-SBS algorithm for feature selection in classification problems. But before, let us briefly review the structure of SVM and our proposed RSVM in the two next subsections.

3.1 The structure of SVM

Support vector machines (SVMs) (Vapnik 1995) were originally introduced to solve two-class classification problems and widely studied and utilized for different pattern recognition applications. SVM tries to find a separating hyperplane, usually in an infinite space, in such a way that maximizes the margin between two classes of input data. The good generalization ability and remarkable achievement results of SVM have convinced the researchers that it can be applied to any classification problems.

Let $\{x_i, y_i\}_{i=1}^{n}$ be n training data vectors, where $x_i \in R^m$ is the input data and $y_i \in \{-1, +1\}$ is the associated class label. The SVM searches for the optimal separating function between two classes of input data achieved by a quadratic programming (QP) problem:

$$\begin{aligned} &\text{Minimize } Q(w, b, \xi) = \frac{1}{2} w^2 + C \sum_{i=1}^{n} \xi_i \\ &\text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i \\ &\qquad \xi_i \geq 0, \quad i = 1, \ldots, n \end{aligned} \tag{2}$$

where w and b are the weight vector and bias term of the OSH, respectively, C is the trade-off parameter between maximizing the margin and misclassification error, and finally, ξ_i is the slack variable that allows some misclassification errors in the training phase. For nonlinearly separable problems, a promising trick is to map the training sample into a higher-dimensional space, namely feature space, using function $\varphi : R^m \rightarrow R^s$ and try to find a linear separating hyperplane between the classes. Optimization problem (2) is usually solved by converting to the Lagrangian dual formulation and introducing the kernel function $K(x_i, x_j) = \varphi(x_i).\varphi(x_j)$ which satisfies Mercer's conditions as follows:

$$\begin{aligned} &\text{Minimize } Q(\alpha) = \frac{1}{2}\sum_{i,j=1}^{n}\alpha_i\alpha_j y_i y_j K\left(x_i, x_j\right) - \sum_{i=1}^{n}\alpha_i \\ &\text{subject to } \sum_{i=1}^{n}\alpha_i y_i = 0, \quad 0 \le \alpha_i \le C \end{aligned} \tag{3}$$

The non-negative Lagrange multipliers α_i, corresponding to each data point x_i, are the only unknown parameters and are calculated by solving (3). For most data vectors, α_i are zero. The samples with nonzero α_i are called support vectors. The final separating hyperplane can be written as

$$f(x) = w^T.\varphi(x) + b = \sum_{i=1}^{n}\alpha_i y_i K(x_i, x) + b \tag{4}$$

where the bias term b is given by

$$b = y_j - \sum_{i\in S}\alpha_i y_i K(x_i, x_j) \tag{5}$$

and the decision function is to form

$$D(x) = sign(f(x)) = sign\left(\sum_{i=1}^{n}\alpha_i y_i K(x_i, x) + b\right) \tag{6}$$

As we can conclude from (6), the final classifier is calculated by only data points with nonzero Lagrange multipliers, namely support vectors (SVs). Consequently, support vector machines are very sensitive to noisy data and outliers.

3.2 The structure of RSVM

Despite many advantages and opening new opportunities for researchers, support vector machines face some challenges. As mentioned before, the SVM classifier is determined by a small portion of input data (support vectors) only. Thus, it becomes very sensitive to noisy data as well as outliers. Another challenge in training of SVM is that the input data points have equal effects on training the classifier. To tackle this challenge, fuzzy support vector machine (FSVM) (Benítez-Peña et al. 2019) assigns different membership values to each training sample. However, choosing the membership function in FSVM is a critical issue because these membership values appear in its cost function. Instead, we proposed relaxed constraint support vector machines (RSVMs) (Sabzekar et al. 2011) that insert the fuzzy memberships in the constraints of SVM formulation (2). It has been shown that this idea causes the RSVM to be more resistant to noise and produces more reliable results. This noise-aware version of SVM is extended to solve one-class (GhasemiGol et al. 2010) and multi-class (Sabzekar et al. 2009) classification problems and has shown its performance on different applications such as clustering (Sabzekar and Naghibzadeh 2013b), TCP traffic classification problem (Sabzekar et al. 2013) and arrhythmia detection using ECG (Nasiri et al. 2009). The fuzzy inequality of RSVM gives more flexibility to SVM. The RSVM formulation is as follows:

$$\begin{aligned} &\text{Minimize } Q(w, b, \xi) = \frac{1}{2}w^2 + C\sum_{i=1}^{n}\xi_i \\ &\text{subject to } y_i(w^T x_i + b) \underset{\sim}{\geq} 1 - \xi_i \\ &\xi_i \ge 0, \quad i = 1, \ldots, n \end{aligned} \tag{7}$$

We define a membership function for the fuzzy inequality in Eq. (7) as follows:

$$\mu : R^{m+1+n} \to [0, 1), i = 1, 2, \ldots, n,$$

$$\mu_i(w, b, \xi) = \begin{cases} 1, & if y_i(w^T x_i + b) \ge 1 - \xi_i \\ \dfrac{(w^T x_i + b) - 1 + \xi_i + d_i}{d_i}, & if\, 1 - (\xi_i + d_i) \le y_i(w^T x_i + b) \le 1 - \xi_i \\ 0, & if y_i(w^T x_i + b) < 1 - (\xi_i + d_i) \end{cases} \tag{8}$$

and

$$P_i = \left\{(w, b, \xi) \in R^{m+1+n} | y_i\left(w^T x_i + b\right) \underset{\sim}{\geq} 1 - \xi_i, \xi_i \ge 0\right\}. \tag{9}$$

Let $P = \bigcap_{i\in I} P_i$, where $I = \{1,2,\ldots,n\}$, then Eq. (7) can be written as follows:

$$Minimize\left\{Q(w, b, \xi) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\xi_i \middle| (w, b, \xi) \in P\right\} \tag{10}$$

As we know, each α-cut on (7) (as shown in Fig. 1) forms the classical set $P(\alpha)$ as follows:

$$P(\alpha) = \left\{(w, b, \xi) \in R^{m+1+n} \middle| \mu_P(w, b, \xi) \ge \alpha\right\}, \tag{11}$$

where $\alpha \in (0, 1]$ and $\mu_P(x) = \inf\{\mu_i(x), i \in I\}$.

For a given α, the optimal solution of each constraint will be:

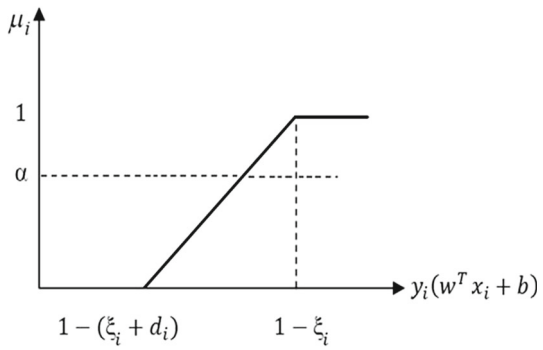


Fig. 1 Membership function μ_i

$$S(\alpha) = \left\{ (w, b, \xi) \in R^{m+1+n} | \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \xi_i \right\} = \left\{ Min \frac{1}{2}\|w'\|^2 + C\sum_{i=1}^{n} \xi_i', (w', b', \xi') \in P(\alpha) \right\}, \tag{12}$$

where

$$P(\alpha) = \bigcap_{i \in I} \{ (w, b, \xi) \in R^{m+1+n} | y_i (w^T x_i + b) \geq r_i(\alpha), \xi_i \geq 0 \}. \tag{13}$$

Taking $r_i(\alpha) = 1 - \xi_i - d_i(1-\alpha)$, RSVM formulation can be written as:

$$\begin{aligned} &\text{Minimize } \frac{1}{2}w^2 + C\sum_{i=1}^{n} \xi_i \\ &\text{subject to } y_i(w^T x_i + b) \geq 1 - \xi_i - d_i(1-\alpha) \\ &\qquad \xi_i \geq 0, \quad i = 1, 2, \ldots, n \end{aligned} \tag{14}$$

Then, we can convert optimization problem (14) to its dual form as follows:

$$\begin{aligned} &\text{Maximize } Q(\beta) = \sum_{i=1}^{n} \beta_i (1 - d_i + d_i\alpha) - \frac{1}{2}\sum_{i,j=1}^{n} \beta_i \beta_j y_i y_j K(x_i, x_j) \\ &\text{subject to } \sum_{i=1}^{n} \beta_i y_i = 0, 0 \leq \beta_i \leq C, \quad i = 1, 2, \ldots, n, \end{aligned} \tag{15}$$

where β_i are Lagrange multipliers. It should be noted that the decision function of RSVM is given such as (6).

The RSVM method introduces two new parameters, namely importance degrees d_i and uncertainty factor α. The first parameter is assigned to each training sample. A greater value of d_i gives more flexibility to corresponding x_i to violate its constraints. Thus, for noisy data and outliers, we can assign a greater d_i to allow a larger margin. The second parameter represents our certainty about the whole data set. A greater value of α means less violation from the constraints. In the presence of high level of noise, a smaller α causes the constraints to be more relaxed. For more details about these parameters see Sabzekar et al. (2011).

3.3 The proposed RSVM-SBS algorithm

The feature selection methods are directly affected by noisy data and outliers. The wrapper-based methods utilize a classifier for evaluation of each feature subset. Therefore, this problem is more critical for these types of feature selection methods. Among different classification methods, SVM and its extensions have been widely used for feature selection. However, as we discussed in Sabzekar and Naghibzadeh (2013b), the optimal separating hyperplane in SVM is simply affected by noisy data and outliers. The robustness of an algorithm is defined as its capability to overcome noisy data and outliers. Accordingly, our main goal is to propose a robust SVM-based feature selection approach.

In this section, we propose a new method for eliminating irrelevant features by combining RSVM and sequential backward search (SBS) (Marill and Green 1963) approaches and call it RSVM-SBS. The SBS method evaluates each feature subset and removes one feature in each iteration of the algorithm using the following error function:

$$S_j = ERR\left(x^{(-j)}\right) - ERR \tag{16}$$

where ERR is the error of training data and $ERR\left(x^{(-j)}\right)$ is the test error when the jth feature is removed:

$$ERR\left(x^{(-j)}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\underset{\sim}{y}\left(x_i^1, \ldots, x_i^{j-1}, x_i^{j+1}, \ldots, x_i^m\right) \neq y_i \right), \tag{17}$$

where $\underset{\sim}{y}$ is the class label for the ith input data which is predicted by RSVM when the jth feature is removed. It should notable that the feature with the least S_j is introduced as the least important feature and is chosen to be removed. It should be noted that the SBS method is a greedy algorithm. Thus, the main challenge of such an algorithm is its high time complexity. However, detecting one feature in each iteration of the SBS as an irrelevant feature using our noise-aware RSVM as objective function, helps us to identify noisy features. Thus, to decrease the effect of noisy data and outliers on feature selection in classification problems, we utilize our proposed noise-aware version of SVM, namely, relaxed constraints support vector machine as the evaluation function. The proposed RSVM-SBS pseudocode is as follows.

RSVM-SBS Algorithm

Inputs: The training samples x_t with m features and desired class labels y_t
Output: The best feature subset

1. Let the remaining feature subset be $F = \{1, \ldots, m\}$ and $R = E = \emptyset$ be the list of removed features and test error, respectively
2. Form the training samples set x_t with only the feature set F as $x_t = x_t(:,F)$
3. Train RVSM using training data (x_t, y_t)
4. Test the model and compute the error rate for $(x_t(:,F), y_t)$
5. For all training data x_t, calculate S_j and find the feature u with the smallest S
6. $E = [E_t, E]$, where E_t is classification error on test sample x_u
7. $R = [F(u), R]$
8. Remove the feature with the smallest selection criterion $F = F(1:u-1, u+1:length(F))$
9. If $length(F) > 1$ go to 2
10. The best features for a given data set will be the subset which has the least test error

As described in the proposed RSVM-SBS algorithm, one feature is removed in each step. In fact, this feature is the most irrelevant among the given features. Let F denotes the set of surviving features. At the beginning of the algorithm, it initialized by all m features of a given data set. Moreover, R and E are the set of discarded features and error lists, respectively. They are empty sets at the start of the RSVM-SBS algorithm. In each iteration of the algorithm, one feature is identified as an irrelevant feature. To do this, we train the RSVM in the absence of each feature in F using training data. Then, the feature with smallest S (namely, u) is selected for removing from F and is added to R:

$$u = argmin(S) \tag{18}$$

Finally, it should be noted that if we have a data set with more than two classes of data, the problem would be a multi-class problem. In this case, the one-against-all RSVM is used for training. In the next section, the proposed RSVM-SBS approach is evaluated using real-world data sets.

4 Experimental results

Table 1 gives some details about the data sets used for the assessment of the proposed RSVM-SBS and its comparisons to other state-of-the-art methods in the literature. All data sets are obtained from the UCI Repository (Blake et al. 1998).

For all data sets, the value of attributes is normalized into the interval $[-1,1]$ using an affine transformation. Classification accuracy is calculated using tenfold cross-validation. We used the RBF kernel function $K(x,y) = e^{\|x-y\|^2/\sigma^2}$ with kernel parameter $\sigma = 0.5$ and $C = 100$ for all methods to compare their results in the same conditions. It gives us the opportunity to evaluate these methods in terms of their differences (membership function for FSVM and relaxed constraints for the proposed RSVM).

In evaluation of each method, the best feature subset on training data is selected based on the test error on the selected features by the following cost function as in Xia (2008):

$$ERR(x_s) = \frac{1}{n_s}\sum_{i=1}^{n_s}(\tilde{y}(x_{si}) \neq y_i) \tag{19}$$

where n_s is the number of the test samples and $\tilde{y}(x_i)$ is the predicted class label of x_i. Table 2 summarizes the results of the experiments in which the proposed RSVM-SBS is compared with SVM-SBS, FSVM-SBS, HO-SVM (Maldonado and Weber 2009) and the situation when we do not use any feature selection.

As shown in Table 2, in all of the data sets except for the Iris data set (on which the error rate for RSVM-SBS is same as the SVM-SBS, FSVM-SBS and HO-SVM error rates) and Lymphography data set (that HO-SVM has shown better results) the feature selection using RSVM reported lower error rate. Moreover, the proposed method has shown greater F1-score for all data sets except for the Image Segmentation data set. Furthermore, the proposed method eliminates more features as irrelevant ones. In view of the fact that in many data mining applications there is a direct relationship between having a high number of features in a data set and the overfitting problem, the proposed RSVM-SBS algorithm shows more accurate results in different applications. Table 3 shows the minimum number of selected features using different methods after 10 distinct iterations. As described in Table 3 the proposed method selects the minimum number of features.

In the assessment of each feature selection method for classification tasks, it is important to select the minimum subset of features and simultaneously achieve higher classification accuracy. Thus, we come up with a measure that considers both the error rate and the number of features selected such as the product of the error rate and the percentage of features selected. A lower value for this criterion means better performance of feature selection method. Figure 2 compares the results of each method with this measure.

As shown in Fig. 2, the proposed method tries to select a small number of features and at the same time tries to eliminate irrelevant features.

Table 1 Details of data sets

Data set	Number of samples	Number of features	Number of classes
Heart	270	13	2
Pima	768	8	2
Iris	150	4	3
Lymphography	148	18	4
Glass	214	9	6
Image segmentation	210	19	7
Dorothea	1950	100,000	2

Table 2 The average of error rates and F1-scores

Data set	All features		SVM-SBS		FSVM-SBS		HO-SVM		RSVM-SBS	
	Error rate	F1-score	Error rate	F1-score	Error rate	F1-score	Error rate	F1-score	Error rate	F1-score
Heart	21.85	73.21	19.26	78.67	18.52	80.23	18.52	80.23	17.73	87.43
Pima	23.44	68.04	22.79	70.63	22.14	73.41	21.50	77.80	21.12	80.93
Iris	2.70	80.51	2.01	91.85	2.01	91.85	2.01	91.85	2.01	91.85
Lymphography	17.57	72.68	14.86	72.89	13.51	79.65	13.43	81.64	13.51	84.12
Glass	29.44	66.67	29.44	66.67	27.57	70.54	25.67	72.76	23.46	75.91
Image segmentation	9.52	71.43	8.10	77.12	7.14	82.42	8.41	80.65	7.66	84.73
Dorothea	41.30	52.49	34.56	55.73	33.67	58.62	30.15	61.02	25.26	67.49

Table 3 The minimum number of features selected in different methods

Data set	All features	SVM-SBS	FSVM-SBS	HO-SVM	RSVM-SBS
Heart	13	12	11	11	10
Pima	8	8	8	7	7
Iris	4	3	3	3	3
Lymphography	18	16	16	17	15
Glass	9	8	6	7	6
Image segmentation	19	16	16	14	14
Dorothea	100,000	1025	820	2945	936

In another experiment, we performed the following grid search for trade-off parameter C and kernel parameter for each kernel function by choosing geometrically increasing values and repeat the feature selection for each C and kernel parameter pair and choose the best values for the hyper-parameters and the best number of features for that hyper-parameter combination:

$$\begin{aligned} C &\in \{2^{-5}, 2^{-3}, 2^{-1}, 2^{1}, 2^{3}, 2^{5}, \ldots 2^{13}, 2^{15}\}, \\ \sigma &\in \{2^{-15}, 2^{-13}, \ldots, 2^{-1}, 2^{1}, 2^{3}, 2^{5}\}, \\ d &\in \{2.3\ldots 100\}, \end{aligned} \tag{20}$$

where σ and d are the parameters of RBF and polynomial kernel functions, respectively. However, the linear kernel function is parameter-free. Figure 3 reports the product of the error rate and the percentage of features selected using the best SVM hyper-parameters.

As shown in Fig. 3, the proposed method outperforms other methods.

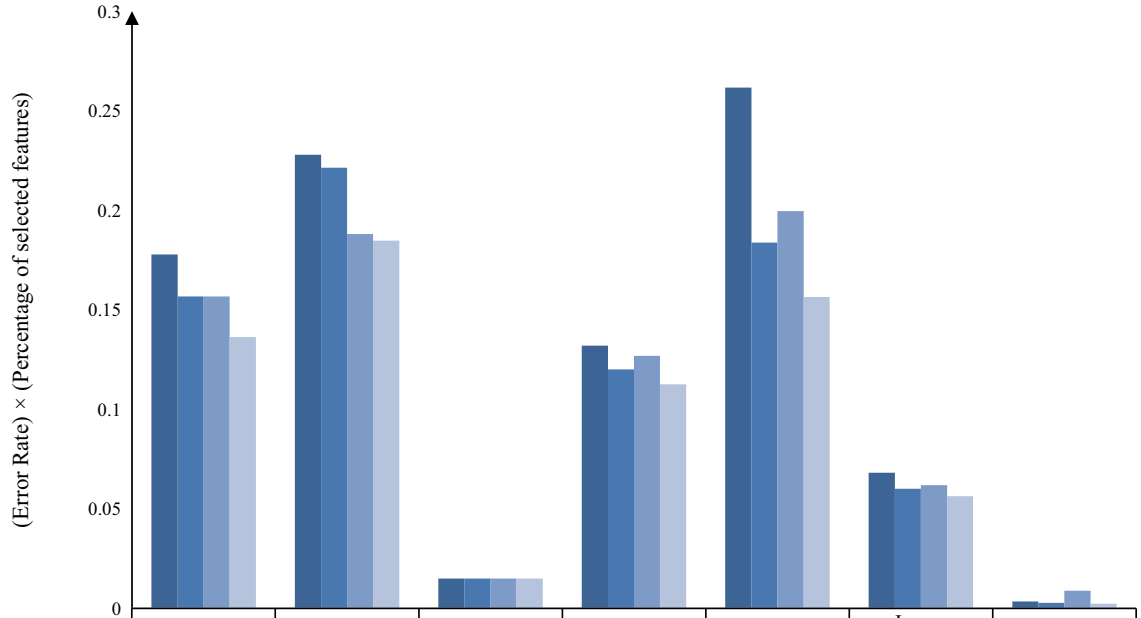
For the last experiment, we compared our proposed RSVM-SBS (with optimized hyper-parameters) with Boruta method (Kursa and Rudnicki 2010). It deals with the "all-relevant" feature selection instead of "minimal-optimal" problem. Therefore, it labels each feature as important or unimportant feature. Table 4 summarizes the results in terms of the number of selected features and the product of error rate (E) and the percentage of selected features (P).

As shown in Table 4, with increasing the number of features, the Boruta algorithm selects more features as important ones in comparison with RSVM-SBS. The results confirm the superiority of the proposed RSVM-SBS.

5 Conclusions and discussions

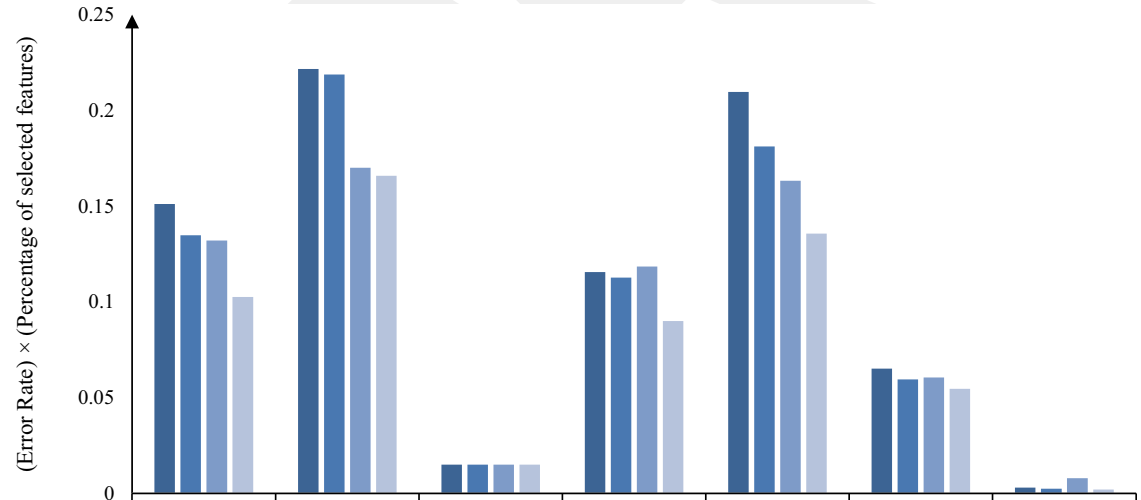
Feature selection is one of the most important steps in data mining applications. Increasing the number of feature vectors in a data set causes different problems such as high complexity of time, overfitting and decreasing the efficiency of the method through the impact of irrelevant features. There are many efforts in the literature aiming to select the best feature subset in a data set. Feature selection approaches which utilized support vector machines have produced promising results in recent years. However, one drawback of such algorithms is that they are very sensitive to noisy data and outliers, because the optimal separating hyperplane obtained by SVM depends on only a small portion of the input samples, namely support vectors.

In this paper, a new method for feature selection is proposed that uses relaxed constraints support vector machines (RSVMs). RSVM is a recently presented algorithm by the authors that has a good ability in the presence of noisy data and outliers. The RSVM converts the constraints of SVM formulation to fuzzy inequalities and allow



	Heart	Pima	Iris	Lymphography	Glass	Image Segmentation	Dorothea
SVM-SBS	0.177784615	0.2279	0.015075	0.132088889	0.261688889	0.068210526	0.0035424
FSVM-SBS	0.156707692	0.2214	0.015075	0.120088889	0.1838	0.060126316	0.00276094
HO-SVM	0.156707692	0.188125	0.015075	0.126838889	0.199655556	0.061968421	0.008879175
RSVM-SBS	0.136384615	0.1848	0.015075	0.112583333	0.1564	0.056442105	0.002364336

Fig. 2 The product of error rate and the percentage of selected features for different data sets



	Heart	Pima	Iris	Lymphography	Glass	Image Segmentation	Dorothea
SVM-SBS	0.151201	0.2217	0.015075	0.1156221	0.2096499	0.06521456	0.003024
FSVM-SBS	0.134802	0.218712	0.015075	0.112771	0.18125	0.059484	0.002412
HO-SVM	0.132089	0.170012	0.015075	0.118552	0.163201	0.0605671	0.007963
RSVM-SBS	0.102548	0.165878	0.015075	0.090013	0.13561	0.054589	0.002045

Fig. 3 The product of error rate and the percentage of selected features for different data sets with SVM hyper-parameter optimization

them to have more relaxation and flexibility. Thanks to this method, we can handle tolerance and uncertainty in our given data. Finally, it can be used efficiently for feature selection because in any SVM-based feature selection method noisy samples and outliers may affect the

Table 4 Comparison of the Boruta algorithm with the proposed method

Data set	# features	Boruta		RSVM-SBS	
		# selected features	$E \times P$	# selected features	$E \times P$
Heart	13	8	0.1274	9	0.1025
Pima	8	6	0.1828	6	0.1659
Iris	4	4	0.1645	3	0.0151
Lymphography	18	13	0.0941	14	0.9001
Glass	9	8	0.1712	6	0.1356
Image segmentation	19	17	0.0512	14	0.0546
Dorothea	100,000	2141	0.0158	924	0.0020

performance of SVM and feature selection method, consequently.

For feature selection, we combine RSVM with a sequential backward search that eliminates one less important feature in each iteration of the algorithm and name the new method as RSVM-SBS. The proposed method has lower error rate and higher F1-score than other methods. Moreover, to confirm the improvements are statistically significant, we applied a two-tailed Student's t test on F1-scores and compared the proposed RSVM-SBS with the other methods. The performance of RSVM-SBS is significantly better than the other methods, with a p-value smaller than 0.05. Therefore, we confirmed the effectiveness of the proposed approach. The proposed method is not dependent on SBS or any similar algorithm. SBS is chosen because it is a simple and straightforward algorithm for feature selection. Moreover, it is easy to understand and interpret. However, the proposed RSVM-based method is not restricted to SBS and one can easily extend it to other feature selection methods. Thus, we can summarize the strengths of the proposed method as follows. The main advantage of the proposed method in comparison with other feature selection method is that it can deal with noises in data and feature set. The reason is that it utilizes a noise-aware objective function (RSVM) that decides about the quality of each feature, separately. Moreover, it utilizes SBS method for feature selection that is a simple method and can easily replace by sequential forward selection as well as sequential backward elimination methods. It should be noted that the SBS method is a greedy algorithm. Thus, the main challenge of such an algorithm is its high time complexity. However, detecting one feature in each iteration of the SBS as an irrelevant feature using our noise-aware RSVM as objective function, helps us to identify noisy features. Although the wrapper-based feature selection methods report more accurate results, their main drawback is high computational complexity. Thus, they are not recommended such methods for large data sets. As a future work, we may investigate a heuristic method for searching the solution space. Furthermore, our backward elimination methods can be combined with forward selection methods to achieve better performance.

Compliance with ethical standards

Conflict of interest The authors declared no conflicts of interest with respect to the authorship and/or publication of this article.

References

Abdar M, Makarenkov V (2019) CWV-BANN-SVM ensemble learning classifier for an accurate diagnosis of breast cancer. Measurement 146:557–570

Abdoos A, Khorshidian Mianaei P, Rayatpanah Ghadikolaei M (2016) Combined VMD-SVM based feature selection method for classification of power quality events. Appl Soft Comput 38:637–646

Aljarah I, Al-Zoubi A, Haris F, Hassonah M, Mirjalili S, Saadeh H (2018) Simultaneous feature selection and support vector machine optimization using the grasshopper optimization algorithm. Cognit Comput 10:478–495

Benítez-Peña S, Blanquero R, Carrizosa E, Ramírez-Cobo P (2019) Cost-sensitive feature selection for support vector machines. Comput Oper Res 106:169–178

Blake C, Keogh E, Merz CJ (1998) UCI repository of machine learning databases. Technical report, Department of Information and Computer Science, University of California, Irvine, CA, http://www.ics.uci.edu/~mlearn/MLRepository.htm

Blum A, Langley PP (1997) Selection of relevant features and examples in machine learning. Artif Intell 97:245–271

Dash M, Liu H (1997) Feature selection for classification. Intell Data Anal 1:13–156

Duda PEHRO, Stork DG (2001) Pattern classification. Wiley-Interscience Publication, Hoboken

Faris H, Al-Zoubi A, Heidari A, Aljarah I, Mafarja M, Hassonah M, Fujita H (2019) An intelligent system for spam detection and identification of the most relevant features based on evolutionary random weight networks. Inf Fusion 48:67–83

GhasemiGol M, Sabzekar M, Monsefi R, Naghibzadeh M, Sadoghi Yazdi H (2010) Support vector data description with fuzzy constraints. In: First international conference on intelligent systems, modelling and simulation (ISMS), pp 10–14, Liverpool, England

Guyon I, Elisseeff A (2003) An introduction to variable and feature selection. J Mach Learn Res 3:1157–1182

Hancer E, Xue B, Zhang M (2018) Differential evolution for filter feature selection based on information theory and feature ranking. Knowl Based Syst 140:103–119

He Q, Wu C (2011) Membership evaluation and feature selection for fuzzy support vector machine based on fuzzy rough sets. Soft Comput 15:1105–1114

Kursa M, Rudnicki W (2010) Feature selection with the Boruta package. J Stat Softw 36:1–13

Liu Y, Zheng YF (2006) FS-SFS: a novel feature selection method for support vector machines. Pattern Recogn 39:1333–1345

Lu M (2019) Embedded feature selection accounting for unknown data heterogeneity. Expert Syst Appl 119:350–361

Mafarja M, Mirjalili S (2018) Whale optimization approaches for wrapper feature selection. Appl Soft Comput 62:441–453

Mafarja M, Aljarah I, Heidari A, Hammouri A, Faris H, Al-Zoubi A, Mirjalali S (2018) Evolutionary population dynamics and grasshopper optimization approaches for feature selection problems. Knowl Based Syst 145:25–45

Maldonado S, Weber R (2009) A wrapper method for feature selection using support vector machines. Inf Sci 179:2208–2217

Maldonado S, Weber R, Basak J (2011) Simultaneous feature selection and classification using kernel-penalized support vector machines. Inf Sci 181:115–128

Maldonadoa S, López J (2018) Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification. Appl Soft Comput 67:94–105

Marill T, Green DM (1963) On the effectiveness of receptors in recognition system. IEEE Trans Inf Theory 9:11–17

Mundra PA, Rajapakse JC (2010) SVM-RFE with MRMR filter for gene selection. IEEE Trans Nanobiosci 9(1):31–37

Nasiri JA, Sabzekar M, Sadoghi Yazdi H, Naghibzadeh M, Naghibzadeh B (2009) Intelligent arrhythmia detection using genetic algorithm and emphatic SVM (ESVM). In: Third UKSim European symposium on computer modeling and simulation (EMS), pp 112–117, Athens, Greece

Neumann J, Schnorr C, Steidl G (2005) Combined SVM-based feature selection and classification. Mach Learn 61:129–150

Pławiak P, Abdar M, Acharya UR (2019) Application of new deep genetic cascade ensemble of SVM classifiers to predict the Australian credit scoring. Appl Soft Comput 84:105740

Sabzekar M, Naghibzadeh M (2013a) Fuzzy c-means improvement using relaxed constraints support vector machines. Appl Soft Comput 13:881–890

Sabzekar M, Naghibzadeh M (2013b) Fuzzy c-means improvement using relaxed constraints support vector machines. Appl Soft Comput 13(2):881–890

Sabzekar M, Naghibzadeh M, Sadoghi Yazdi H, Effati S (2009) Emphatic constraints support vector machines for multiclass classification. In: Third UKSim European symposium on computer modeling and simulation (EMS), pp 118–123, Athens, Greece

Sabzekar M, Sadoghi Yazdi H, Naghibzadeh M (2011) Relaxed constraints support vector machines for noisy data. Neural Comput Appl 20:671–685

Sabzekar M, Hossein Yaghmaee Moghaddam M, Naghibzadeh M (2013) TCP traffic classification using relaxed constraints support vector machines. In: Fathi M (ed) Integration of practice-oriented knowledge technology: trends and prospectives. ISBN 978–3–642–34470–1, pp 129–141

Shieh MD, Yang CC (2008) Multiclass SVM-RFE for product form feature selection. Expert Syst Appl 35:531–541

Tang W (2010) Feature selection using hybrid Taguchi genetic algorithm and fuzzy support vector machine. In: Sixth international conference on natural computation, pp 2348–2355

Torres-Valencia C, Álvarez-López M, Orozco-Gutiérrez Á (2017) SVM-based feature selection methods for emotion recognition from multimodal data. J Multimodal User Interfaces 11:9–23

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vapnik V (1998) Statistical learning theory. Wiley, New York

Xia H (2008) Feature selection based on fuzzy SVM. In: Fifth international conference on fuzzy systems and knowledge discovery (FSKD), vol 1, pp 586–589

Xia H, Hu BQ (2006) Feature selection using fuzzy support vector machines. Fuzzy Optim Decis Mak 5:187–192

Xiong W, Wang C (2008) Feature selection: a hybrid approach based on self-adaptive ant colony and support vector machine. In: International conference on computer science and software engineering, pp 751–754

Yan C, Ma J, Luo H, Patel A (2019) Hybrid binary coral reefs optimization algorithm with simulated annealing for feature selection in high-dimensional biomedical datasets. Chemom Intell Lab Syst 184:102–111

Zakeri A, Hokmabadi A (2019) Efficient feature selection method using real-valued grasshopper optimization algorithm. Expert Syst Appl 119:61–72

Zaman S, Karray F (2009) Features selection for intrusion detection systems based on support vector machines. In: 2009 6th IEEE consumer communications and networking conference, pp 1–8

Zheng K, Wang X (2018) Feature selection method with joint maximal information entropy between features and class. Pattern Recogn 77:20–29