Sensors and Actuators B 212 (2015) 353–363

Contents lists available at ScienceDirect

Sensors and Actuators B: Chemical

journal homepage: www.elsevier.com/locate/snb

Feature selection and analysis on correlated gas sensor data with recursive feature elimination

Ke Yan[a], David Zhang[b,*]

[a] Department of Electronic Engineering, Graduate School at Shenzhen, Tsinghua University, Shenzhen 518055, China
[b] Biometric Research Centre, Department of Computing, The Hong Kong Polytechnic University, Hung Hom, Kowloon, Hong Kong

ARTICLE INFO

Article history:
Received 18 August 2014
Received in revised form 1 February 2015
Accepted 6 February 2015
Available online 16 February 2015

Keywords:
Feature selection
Feature ranking
SVM-RFE
Correlation bias
Breath analysis
Transient feature

ABSTRACT

Support vector machine recursive feature elimination (SVM-RFE) is a powerful feature selection algorithm. However, when the candidate feature set contains highly correlated features, the ranking criterion of SVM-RFE will be biased, which would hinder the application of SVM-RFE on gas sensor data. In this paper, the linear and nonlinear SVM-RFE algorithms are studied. After investigating the correlation bias, an improved algorithm SVM-RFE + CBR is proposed by incorporating the correlation bias reduction (CBR) strategy into the feature elimination procedure. Experiments are conducted on a synthetic dataset and two breath analysis datasets, one of which contains temperature modulated sensors. Large and comprehensive sets of transient features are extracted from the sensor responses. The classification accuracy with feature selection proves the efficacy of the proposed SVM-RFE + CBR. It outperforms the original SVM-RFE and other typical algorithms. An ensemble method is further studied to improve the stability of the proposed method. By statistically analyzing the features' rankings, some knowledge is obtained, which can guide future design of e-noses and feature extraction algorithms.

1. Introduction

Feature selection (FS) is a widely-used technique in pattern recognition applications. By removing irrelevant, noisy, and redundant features from the original feature space, FS alleviates the problem of overfitting and improves the performance of the model. The time and space cost of the learning algorithm can also be reduced. More importantly, we can gain a deeper insight of the data by analyzing the importance of the features [1,2]. Many researchers have explored the use of FS techniques in electronic nose (e-nose) systems and achieved good results [3–9].

In the context of classification, FS algorithms can be roughly divided into three categories: filters, wrappers and embedded methods, based on how they interact with classifiers [1,2]. Filters evaluate each feature by predefined criteria, such as correlation criteria and information theoretic criteria [1], which are independent from classifiers. Wrappers treat classifiers as black boxes and aim at finding a feature subset that has the minimum cross-validation error on the training data. Examples of wrappers include sequential forward selection [3], genetic algorithms, and simulate annealing [4]. Embedded methods generally include two kinds of approaches. In some methods, such as a decision tree [7], the training of the classifier intrinsically selects a subset of features. Some methods estimate the importance of the features from the coefficients in the classifiers, e.g. the algorithm in [5].

Support vector machine recursive feature elimination (SVM-RFE) is an embedded FS algorithm proposed by Guyon et al. [10]. It uses criteria derived from the coefficients in SVM models to assess features, and recursively removes features that have small criteria. It has both linear and nonlinear versions. The nonlinear SVM-RFE uses a special kernel strategy [10,11] and is preferred when the optimal decision function is nonlinear. As a backward elimination method, SVM-RFE is able to model the dependencies among features. Compared to wrappers, SVM-RFE does not use the cross-validation accuracy on the training data as the selection criterion, thus is (1) less prone to overfitting; (2) able to make full use of the training data; (3) much faster, especially when there are a lot of candidate features. As a result, it has been successfully applied in many problems, especially in gene selection [10–15].

However, there is still one problem in SVM-RFE that has not been addressed. When some of the candidate features are highly correlated, the assessing criteria of these features will be influenced, and their importance will be underestimated. Inspired by [16], we call this phenomenon "correlation bias". It is a crucial problem especially for gas sensor features that are often correlated. In this paper,

[*] Corresponding author. Tel.: +852 27667271.
E-mail addresses: yank10@mails.tsinghua.edu.cn (K. Yan), csdzhang@comp.polyu.edu.hk (D. Zhang).

http://dx.doi.org/10.1016/j.snb.2015.02.025
0925-4005/© 2015 Elsevier B.V. All rights reserved.

a simulated experiment is first employed to illustrate this problem. Then a novel strategy, correlation bias reduction (CBR), is proposed to reduce this potential bias in both linear and nonlinear SVM-RFE. Finally, an ensemble method is suggested to improve the stability of the feature selection results.

It is known that human breath contains biomarkers that can be used for disease diagnosis [17]. E-nose systems have been applied to analyze breath samples. In this paper, the proposed method is evaluated on two breath analysis datasets. The first breath analysis dataset was collected by an e-nose with 10 gas sensors, three of which were metal oxide semiconductor (MOS) sensors under temperature modulation (TM) [18]. The dataset contains 295 samples from healthy subjects and 279 from diabetics. The second dataset was collected by an e-nose with 12 MOS sensors [19]. The breath samples were from healthy subjects and also subjects with diabetes, renal disease, and airway inflammation, respectively. Over 1000 features are extracted from the gas sensors' responses. The comprehensive feature set contains seven kinds of transient features. Experimental results show that the Gaussian SVM-RFE is better than the linear one, as well as other typical algorithms. The proposed CBR strategy further enhances the accuracy. The ensemble method is proved to have better stability. Furthermore, systematic statistical analysis on the features' rankings reveals useful information about which sensors, feature types and TM voltages are more important. For example, TM sensors significantly outperform the ones operated under constant temperature. Phase feature extracted from TM sensors is proved to be the most effective feature. The information provides guidance for future e-nose and feature designing.

The manuscript is organized as follows. Section 2 describes the details of the linear and nonlinear SVM-RFE algorithm. Section 3 investigates the correlation bias problem and proposes SVM-RFE + CBR. Section 4 introduces the breath analysis datasets and feature extraction methods. Section 5 shows the results of the FS experiments and provides the feature analysis results. Section 6 concludes the paper.

2. SVM-RFE

2.1. Linear SVM-RFE

The output of SVM-RFE is a ranked feature list. Feature selection can be achieved by choosing a group of top-ranked features. The ranking criterion of SVM-RFE is closely related to the SVM model. SVM is a popular algorithm for classification partially due to its high accuracy and good generalization ability. It has been successfully applied in many e-nose applications [9]. Therefore, ranking criterion derived from its model will probably have good performance.

The intuition of SVM is to find a separating hyperplane with the largest margin. In linear separable cases, the margin is twice the distance between the separating hyperplane and the training sample closest to it [20]. Given a set of training samples $\{\boldsymbol{x}_i, y_i\}$, $\boldsymbol{x}_i \in \mathbf{R}^d, y_i \in \{-1, 1\}, i = 1, \ldots, n$, the decision function of a linear SVM is

$$f(\boldsymbol{x}) = \boldsymbol{w} \cdot \boldsymbol{x} + b. \tag{1}$$

It can be proved that the margin M is simply $2/\|\boldsymbol{w}\|$, thus maximizing the margin is equivalent to minimizing $\|\boldsymbol{w}\|^2$ under constraints. The dual form of the Lagrangian formulation of the problem can be written as [20]:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i \cdot \boldsymbol{x}_j, \tag{2}$$

where α_i are the Lagrange multipliers. Solutions of α_i can be found by maximizing L_D under constraints $\alpha_i \geq 0$ and $\sum_{i=1}^{n} \alpha_i y_i = 0$. The samples corresponding to nonzero α's are known as support vectors. Then the weight vector $\boldsymbol{w}$ can be obtained by

$$\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i. \tag{3}$$

The ranking criterion for feature k is the square of the kth element of $\boldsymbol{w}$,

$$J(k) = w_k^2. \tag{4}$$

In each iteration of the recursive feature elimination (RFE), a linear SVM model is trained. The feature with the smallest ranking criterion is removed since it has the least effect on classification [13]. The remaining features are kept for the SVM model in the next iteration. This process is repeated until all the features have been removed. Then the features are sorted according to the order of removal. The later a feature is removed, the more important it should be. When the feature dimension is high, removing features one by one will be time-consuming. In such cases, more than one feature can be removed in each iteration [10]. However, this strategy may influence the precision [13] and cause the correlation bias problem, which will be described in Section 3.1.

2.2. Nonlinear SVM-RFE

Most gene selection problems have much more features (several thousand) than samples (less than 100), so linear SVM-RFE is more suitable in these cases to avoid overfitting. But in many other situations where the number of samples is larger, nonlinear SVM-RFE can be expected to outperform the linear one since it can fit the data with less bias.

Nonlinear SVM considers to map the features into a new space with higher dimension:

$$\boldsymbol{x} \in \mathbf{R}^d \mapsto \Phi(\boldsymbol{x}) \in \mathbf{R}^h. \tag{5}$$

In the new space, the samples are expected to be linearly separable. Thus Eq. (2) can be rewritten as

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j \Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j). \tag{6}$$

Note that the only form that $\Phi(\boldsymbol{x})$'s are involved in the training algorithm is their inner product. So we can replace $\Phi(\boldsymbol{x}_i) \cdot \Phi(\boldsymbol{x}_j)$ with a kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ without knowing the explicit form of Φ. This is a particularly useful trick because it is hard to determine the form of Φ in real-world problems. There are several choices for kernel functions, though, one common choice being the Gaussian kernel

$$K(\boldsymbol{x}_i, \boldsymbol{x}_j) = e^{-\gamma \|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}. \tag{7}$$

Since the form of Φ is unknown, the weight vector $\boldsymbol{w}$ cannot be obtained. However, linear SVM-RFE can be extended to nonlinear cases via a special strategy. If the removal of a feature causes only small changes in the objective function Eq. (6), the feature should be removed [10,11]. This leads to the following ranking criterion for feature k:

$$J(k) = \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j) - \frac{1}{2}\sum_{i,j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i^{(-k)}, \boldsymbol{x}_j^{(-k)}). \tag{8}$$

The notation $(-k)$ means the feature k has been removed, i.e. $\boldsymbol{x}^{(-k)} \in \mathbf{R}^{d-1}$. The above criterion is the difference of Eq. (6) before and after removing feature k while keeping the α's unchanged. The features with small J's will be eliminated in each iteration of RFE. This criterion is applicable for all kinds of kernels. When the linear