Pattern Recognition 43 (2010) 3712–3729

Contents lists available at ScienceDirect

Pattern Recognition

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

SVM-FuzCoC: A novel SVM-based feature selection method using a fuzzy complementary criterion

S.P. Moustakidis, J.B. Theocharis *

Aristotle University of Thessaloniki, Department of Electrical & Computer Engineering, Auth University Campus, 54124 Thessaloniki, Greece

ARTICLE INFO

Article history:
Received 27 January 2009
Received in revised form
5 February 2010
Accepted 4 May 2010

Keywords:
Feature selection
Fuzzy sets
Feature redundancy
Fuzzy complementary criterion
Support vector machines

ABSTRACT

An efficient filter feature selection (FS) method is proposed in this paper, the SVM-FuzCoC approach, achieving a satisfactory trade-off between classification accuracy and dimensionality reduction. Additionally, the method has reasonably low computational requirements, even in high-dimensional feature spaces. To assess the quality of features, we introduce a local fuzzy evaluation measure with respect to patterns that embraces fuzzy membership degrees of every pattern in their classes. Accordingly, the above measure reveals the adequacy of data coverage provided by each feature. The required membership grades are determined via a novel fuzzy output kernel-based support vector machine, applied on single features. Based on a fuzzy complementary criterion (FuzCoC), the FS procedure iteratively selects features with maximum additional contribution in regard to the information content provided by previously selected features. This search strategy leads to small subsets of powerful and complementary features, alleviating the feature redundancy problem. We also devise different SVM-FuzCoC variants by employing seven other methods to derive fuzzy degrees from SVM outputs, based on probabilistic or fuzzy criteria. Our method is compared with a set of existing FS methods, in terms of performance capability, dimensionality reduction, and computational speed, via a comprehensive experimental setup, including synthetic and real-world datasets.

1. Introduction

The exponential growth of computer technology in recent years has led to the proliferation of vast amount of data. Since continued data accumulation is inevitable, preprocessing techniques are essential to keep pace with data collection rate. In this context, feature selection (FS) algorithms have become indispensable components of data preprocessing in the field of machine learning [1], statistical pattern recognition [2,3], and data mining [4,5]. The primary goal of FS is to select a subset of valuable input variables by discarding irrelevant or redundant features. Beneficial effects of FS techniques are learning acceleration and improvement of the classifiers' performance. Additionally, dimensionality reduction of feature spaces enhances interpretability of obtained models, since classifiers involving many features are less comprehensive.

A typical FS process initiates by producing candidate feature subsets based on a search strategy. Each candidate subset is evaluated and compared with the previous best one according to a particular evaluation criterion. Depending on the evaluation criterion used, FS methods are divided into three main categories: wrapper, filter, and hybrid models. In wrapper methods, FS adheres to an induction algorithm and the candidate subsets are validated in terms of accuracy provided by a classifier [6]. Many wrapper methods follow a backward process where at each stage, one of the features is excluded from the feature set, the removal of which yields the least reduction in training accuracy [7]. Although high classification rates are usually achieved, wrappers are computationally intense due to their coupling with the classifier. An additional shortcoming is that when dealing with small size datasets, these methods suffer from overfitting and inferior generalization results.

Filter methods are independent of the classification model used. FS in these methods relies on intrinsic characteristics of features to reveal their discriminating power. Along this direction, several measures of relevance have been employed to carry out FS. Correlation criteria are the simplest measures used [8,9], which can detect only linear dependencies of features. Recently, some FS methods are suggested [10–12], utilizing the mutual information metric, to assess the feature relevance to target classes and redundancy between features, although they require greater computational efforts for their implementation. Moreover, *FFSEM* [13] and filter methods presented in [14,15] use class similarity measures with respect to the selected subset as evaluation criteria. *ReliefF* [16] validates the importance of features according to the separability of neighboring patterns. Zhang et al. [17] proposed a feature selection method according to features' constraint preserving ability. More concrete, a 'good'

* Corresponding author.
E-mail address: theochar@eng.auth.gr (J.B. Theocharis).

0031-3203/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2010.05.007

feature should be the one on which two samples from the same class (must-link constraint) are close to each other, whereas samples from different classes (cannot-link constraint) are far away from each other. A recently proposed unsupervised FS method, referred to here as *Mitra* ([18]), partitions the initial feature set into a number of homogeneous subsets and proceeds to selecting a representative feature from each subset. Despite their low complexity, filter approaches do not always succeed with high classification rates as the evaluation criterion used for FS is not necessarily associated with the classifiers to be applied. Hybrid methods [19,20] integrate FS within the learning algorithm, with the goal to exploit the advantages of both wrapper and filter approaches.

Different strategies have been recently investigated for subset generation. *Branch and Bound* (BB) [21] performs an almost exhaustive search but is exponentially prohibitive even with a moderate number of features. *Sequential forward selection* (SFS) [22], which is simpler and faster for moderate dimensionality problems, iteratively adds features to an initial subset so that a given evaluation criterion is maximized, whereas *sequential backward selection* (SBS) eliminates one feature at a time that exhibits the smallest criterion decrease. *Sequential floating forward selection* (SFFS) [23] and *Plus l–take away r* [24] are more sophisticated versions. At each stage, these methods enlarge feature set using forward selection and then, discard features using backward selection. The relevant metrics used in the existing FS methods are geared towards identifying informative features individually and minimizing redundancy present in the reduced feature subset. Nevertheless, they manipulate information globally, in that a single scalar metric is usually employed (relevance index, correlation, redundancy), integrating all patterns of every class.

In this paper, we suggest the SVM-FuzCoC approach, suitable for high-dimensional feature sets. The proposed FS is a filter method and its main characteristics are described as follows. (i) We introduce the notion of fuzzy partition vector (FPV) associated with each feature, which comprises fuzzy membership grades of training patterns (projected on that feature) to their own classes. FPV treats each feature on a pattern-wise base, allowing us to assess redundancy between features. (ii) Exploiting high generalization capabilities of kernel-based support vector machines (K-SVM), a fuzzy output K-SVM (FO-K-SVM) scheme is developed and applied on each single feature, to construct the associated FPV. (iii) The proposed method performs a forward selection guided by a fuzzy complementary criterion (FuzCoC). Particularly, FuzCoC operates on the pre-computed feature FPVs, paying due attention on complementary characteristics between the features. As a result, we obtain small co-operative subsets of discriminating (highly relevant) and non-redundant features, each one covering better a different pattern region. It is analytically shown that FuzCoC acts like a minimal-redundancy–maximal-relevance (mRMR) criterion used by some existing methods of the literature. (iv) FuzCoC-based feature selection does not adhere only to the FO-K-SVM approach. Therefore, several variants of SVM-FuzCoC are developed, whereby apart from the proposed FO-K-SVM, seven other membership degree determination methods are considered. These methods include parametric and non-parametric probabilistic approaches to convert SVM outputs to well calibrated posterior probabilities and methods based on fuzzy criteria. Experimental investigation on a set of benchmark problems of varying complexities shows that computational load of SVM-FuzCoC is reasonably low, while achieving at the same time, a good trade-off between dimensionality reduction and classification accuracy.

The remainder of the paper is organized as follows. Section 2 presents the fuzzy K-SVM classifier used for determining membership grades of patterns over classes. In Section 3, we elaborate on the proposed FS method, including the FPV properties, useful definitions in the fuzzy domain, and presentation of the SVM-FuzCoC algorithm. Section 4 includes illustrative results to highlight attributes of the suggested approach. Comparative results are given in Section 5, contrasting our method with other FS techniques. Additionally, we analyze the performance of SVM-FuzCoC using different schemes for derivation of pattern classification grades. Finally, conclusions are drawn in Section 6.

2. Fuzzy output kernel-based SVM

A fuzzy output kernel-based SVM (FO-K-SVM) approach is suggested in this section, providing both the decision class and the membership grades of patterns to their classes. Initially, we give an outline of K-SVM principles and proceed to a brief review of traditional multiclass extensions. To determine the fuzzy degrees apportioned to classes, binary K-SVM outputs are fuzzified via a properly designed membership function, applied on boundary surfaces.

2.1. Multiclass kernel-based SVM

Consider a dataset comprising labeled training patterns: $D=\{(\boldsymbol{x}_i,y_i), i=1,\ldots,N\}$. Each pattern $\boldsymbol{x}_i=[x_{i,\ 1},x_{i,\ 2},\ldots,x_{i,\ n}]^T\in\Re^n$ belongs to one of two classes, with its class label given as $y_i\in\{+1,-1\}$. Given a nonlinear mapping $\boldsymbol{\Phi}:\Re^n\to\mathcal{F}$, each vector $\boldsymbol{x}_i$ in the original feature space is transformed to a potentially higher dimensional feature space $\mathcal{F}$: $\boldsymbol{x}_i\to\boldsymbol{\Phi}(\boldsymbol{x}_i)$, $i=1,\ldots,N$. The embedding of feature mapping and the associated kernel function lead to a powerful nonlinear scalar-product-based algorithm (K-SVM), executed in$\mathcal{F}$. K-SVM seeks for a suitable separating hyperplane in the transformed space $\mathcal{F}$: $f(x)=((\boldsymbol{w}\cdot\boldsymbol{\Phi}(x))+b)=0$, parameterized by the pair $(\boldsymbol{w},b)$, $\boldsymbol{w}\in\mathcal{F}$, $b\in\Re$ [25,26]. The optimal decision function is obtained by

$$f(\boldsymbol{x})=\mathrm{sgn}\left(\sum_{i\in S}a_iy_iK(\boldsymbol{x}_i,\boldsymbol{x})+b\right) \tag{1}$$

where $S=\{i:0<a_i^*\le C\}$. Coefficient a_i is non-zero when $\boldsymbol{x}_i$ is a support vector; otherwise it is zero. Any function satisfying Mercer's theorem can be used as scalar product, thus serving as a kernel function. In this research, we employ the Gaussian RBF kernel:

$$K(\boldsymbol{x}_i,\boldsymbol{x}_j)=\exp\left(\frac{-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2}{2\sigma^2}\right) \tag{2}$$

where σ denotes the variance along the feature axis. Selection of the RBF kernel is dictated by its ability to handle high-dimensional data. In our simulations, the parameters C and σ are heuristically determined from a grid of pre-selected values $C\in\{10,30,50,100\}$ and $\sigma\in\{0.005,0.01,0.1,0.5,1\}$.

Since K-SVMs were originally designed for binary classification, a decomposition scheme should be devised to tackle multiclass problems [27]. One-versus-all (OVA) is a common approach, accomplished by combining several binary SVM classifiers. For a M-class problem, OVA proceeds to construct a set of binary classifiers $\{f_1,\ldots f_k,\ldots f_M\}$. Each f_k, $k=1,\ldots,M$, is trained individually to separate class c_k from the rest of the classes, included in $\overline{C}_k=\{1,\ldots,\ell,\ldots,M|\ell\neq k\}$. Following the *winner-takes-all* principle, an unknown pattern $\boldsymbol{x}$ is then assigned to the class that exhibits the maximum decision function value $f_k(\boldsymbol{x})$:

$$c_k=\arg\max_k\left\{f_k(\boldsymbol{x})=\sum_{i\in S}a_i^*y_iK(\boldsymbol{x}_i,\boldsymbol{x})+b^*\right\} \tag{3}$$