Available online at www.sciencedirect.com
SCIENCE DIRECT
Pattern Recognition 39 (2006) 737–746
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Non-parametric classifier-independent feature selection

Naoto Abe*, Mineichi Kudo

Division of Computer Science, Graduate School of Information Science and Technology, Hokkaido University, Sapporo, Hokkaido 060-0814, Japan

Received 2 July 2005; received in revised form 10 November 2005; accepted 10 November 2005

Abstract

Feature selection is used for finding a feature subset that has the most discriminative information from the original feature set. In practice, since we do not know the classifier to be used after feature selection, it is desirable to find a feature subset that is universally effective for any classifier. Such a trial is called *classifier-independent feature selection*. In this study, we propose a novel classifier-independent feature selection method on the basis of the estimation of Bayes discrimination boundary. The experimental results on 12 real-world datasets showed the fundamental effectiveness of the proposed method.

Keywords: Classifier-independent feature selection; Bayes classifier; Garbage feature; Non-parametric; Two-stage feature selection

1. Introduction

In pattern recognition, the goal of feature selection is to find a feature subset that has the most discriminative information from a given set of candidate features. The main benefits of feature selection are follows: (i) reducing the measurement cost and storage requirements, (ii) coping with the degradation of the classification performance due to the finiteness of training sample sets, (iii) reducing training and utilization time and, (iv) facilitating data visualization and data understanding. Feature selection has been applied mainly for the improvement of the discriminating performance and for the reduction of the computational cost in a wide range of applications such as text classification, DNA micro-array data analysis and image recognition/retrieval, and so on.

Algorithms for feature selection can be divided into two groups. One group is called *classifier-specific feature selection* (CSFS). CSFS selects a feature subset that maximizes the value of a given criterion function such as the estimated recognition rate for a specified classifier. Therefore, CSFS is useful when we know in advance what classifier will be used for a given problem. A large number of CSFS algorithms have been proposed both in pattern recognition [1–5] and in machine learning [6–12], and some comparative studies of CSFS have been reported for large-scale feature selection problems [13–16]. In addition, many CSFS algorithms based on the support vector machine (SVM) [17] have been proposed recently [18–22]. Those algorithms usually use a linear kernel. There are also algorithms using a non-linear kernel, while the main purpose of them is not feature selection but the improvement of the performance of SVM classifiers. In machine learning, the so-called "wrapper approach" and "filter approach" are proposed. The former belongs to CSFS group. The latter does not assume a specific classifier, but needs a specific measure. Once a specific measure is assumed, it implies the existence of a hypothetical classifier with which the measure works best. Hence, in this study, we regard the filter approach as belonging to CSFS, too.

The other group is called *classifier-independent feature selection* (CIFS) for which only a few algorithms have been proposed so far [23–26]. In CIFS, we seek a feature subset that is universally effective for all classifiers at the same time. To do this, we have to find a feature subset that contributes the largest separation between class-conditional probability densities. It means that we have to estimate class-conditional probability densities as precisely as possible, which leads Bayes optimal classifier. In other words, CIFS is equivalent to CSFS for Bayes classifier.

* Corresponding author. Tel.: +81 11 706 6852; fax: +81 11 706 7393.
E-mail address: chokujin@main.ist.hokudai.ac.jp (N. Abe).

0031-3203/$30.00
doi:10.1016/j.patcog.2005.11.007

In that sense the goal of CIFS is equivalent to removing only *garbage features* which have no discriminative information. To perform CIFS, we can take one of two approaches on the basis of training samples: (1) try to estimate the class-conditional density of each class so that we can have a quasi Bayes classifier, and (2) try to estimate Bayes classification boundary. We evaluate a feature subset in the degree of contribution for density separation in (1) and in the degree of contribution for constructing the boundary in (2). The divergence method [23] belongs to the first approach, and the subclass method [24], RFI [25] and PRISM [26] belong to the second approach. It would be better to take the latter approach than the former approach because, in general, the estimation of the boundary is easier than that of the densities [27].

Along to this direction, we propose a non-parametric CIFS algorithm on the basis of k-nearest neighbor method. In Section 2, we describe the idea of the proposed method. In Section 3, we investigate the effectiveness of the proposed method mainly from the viewpoint whether the proposed method can find a feature subset that is universally effective for any classifier. To confirm this, we compare the recognition rates of classifiers before and after the garbage feature removal on 12 real-world datasets. Finally, we discuss and conclude the usefulness of the proposed method.

2. Feature selection on the basis of Bayes boundary estimation

2.1. Key idea for feature selection

Our approach is twofold: (1) to estimate Bayes discrimination boundary as precisely as possible and (2) to evaluate the effectiveness of features in terms of the boundary. However, in fact, it is difficult to find the true discrimination boundary under a limited number of training samples. To overcome this problem, we try to find only some points on the discrimination boundary (Fig. 1). This is enough, because the normal vectors on those points show which features are necessary and which are not for discriminating classes. If the discrimination boundary is linear (Fig. 1(a)), the normal vectors on points A, B and C are the same. Therefore, we only need one normal vector for feature evaluation. In the case of non-linear boundary (Fig. 1(b)), at points A and C, the normal vectors indicate that feature x_1 is more important. On the other hand, at point B, the normal vector says that feature x_2 is more important. These normal vectors reflect the importance of features locally but faithfully. By collecting those local evidences, in Fig. 1(a) and (b), we can finally know that both features are necessary.

2.2. Normal vectors on Bayes discrimination boundary

In this paper, we consider two-class problem only, but it is easy to extend the method to multi-class cases. Let a sample $\boldsymbol{x}$ be a vector $\boldsymbol{x}=(x_1, x_2, \ldots, x_d)$ and H be the sets of boundary points. Then, the boundary H of Bayes classifier is shown as

$$H = \{\boldsymbol{x} | P(\omega_1)p(\boldsymbol{x}|\omega_1) - P(\omega_2)p(\boldsymbol{x}|\omega_2) = 0\}. \quad (1)$$

Here, $P(\omega_i)$ is the priori probability of class ω_i, and $p(\boldsymbol{x}|\omega_i)$ is the class-conditional probability density function of class ω_i. By the assumption that $\boldsymbol{x}, \boldsymbol{x}+\Delta\boldsymbol{x}$ are on the boundary, the normal vector $\boldsymbol{u}$ at $\boldsymbol{x}$ is given, with approximation $p(\boldsymbol{x}+\Delta\boldsymbol{x}|\omega_i) \cong p(\boldsymbol{x}|\omega_i) + (\Delta\boldsymbol{x})^{\mathrm{T}}(\nabla p(\boldsymbol{x}|\omega_i))$, by

$$\boldsymbol{u} = P(\omega_1)\nabla p(\boldsymbol{x}|\omega_1) - P(\omega_2)\nabla p(\boldsymbol{x}|\omega_2). \quad (2)$$

2.3. Gradient vectors using k-nearest neighbors

From Eq. (2), the gradient vector of $p(\boldsymbol{x}|\omega)$ needs to be calculated. In this section, a non-parametric way is adopted to this goal. According to Ref. [28], we estimate it on the basis of given training samples.

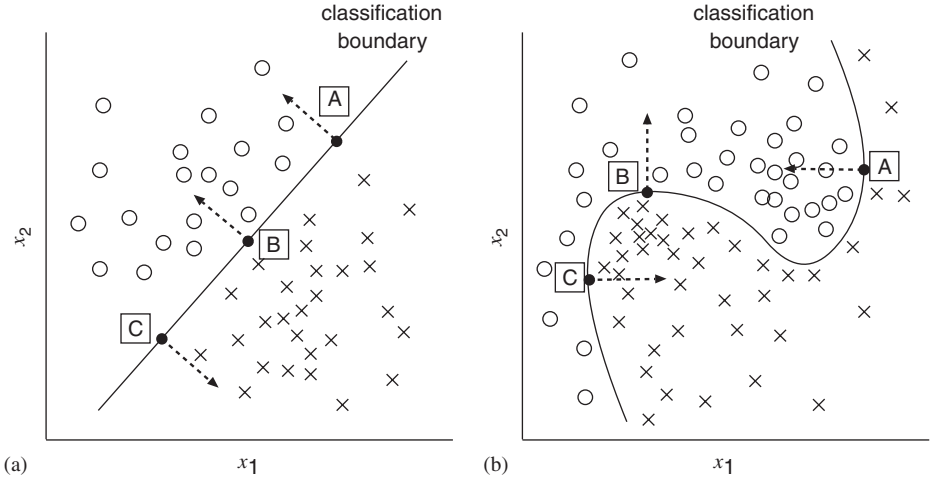

Fig. 1. Local discriminative importance of features: (a) linearly separable case, (b) linearly non-separable case.