Signal Processing 93 (2013) 2956–2968

Contents lists available at SciVerse ScienceDirect

Signal Processing

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

Resampling methods for quality assessment of classifier performance and optimal number of features

Raquel Fandos [a,*], Christian Debes [b], Abdelhak M. Zoubir [a]

[a] Signal Processing Group, Institute of Telecommunications, Technische Universität Darmstadt, Merckstr. 25, 64283 Darmstadt, Germany
[b] AGT Group (R&D) GmbH, 64295 Darmstadt, Germany

ARTICLE INFO

Article history:
Received 10 June 2012
Received in revised form
28 April 2013
Accepted 4 May 2013
Available online 15 May 2013

Keywords:
Pattern recognition
Resampling
Bootstrap
Classifier design and evaluation
Feature evaluation and selection
Optimal dimensionality

ABSTRACT

We address two fundamental design issues of a classification system: the choice of the classifier and the dimensionality of the optimal feature subset. Resampling techniques are applied to estimate both the probability distribution of the misclassification rate (or any other figure of merit of a classifier) subject to the size of the feature set, and the probability distribution of the optimal dimensionality given a classification system and a misclassification rate. The latter allows for the estimation of confidence intervals for the optimal feature set size. Based on the former, a quality assessment for the classifier performance is proposed. Traditionally, the comparison of classification systems is accomplished for a fixed feature set. However, a different set may provide different results. The proposed method compares the classifiers independently of any pre-selected feature set. The algorithms are tested on 80 sets of synthetic examples and six standard databases of real data. The simulated data results are verified by an exhaustive search of the optimum and by two feature selection algorithms for the real data sets.

1. Introduction

When pattern recognition practitioners are required to design a classifier for a specific problem, they are generally provided with a data set of S observations. Each observation s, $1 \le s \le S$, has associated a feature vector $\mathbf{t} = \{t_1, t_2, \ldots, t_N\}$ and a class label $c \in \{1, \ldots, C\}$, where C is the number of classes. Besides other design decisions [1], the pattern recognition expert needs to select:

1. a classification system such as k-Nearest Neighbor (k-NN) [2], neural networks [3], Mahalanobis distance classifier [4], decision trees [5], Fisher's linear discriminant [2], Support Vector Machines (SVM) [6], among others.
2. an n^*-element subset of features, $\mathbf{t}^* = \{t_1^*, t_2^*, \ldots, t_{n^*}^*\}$ with $n^* \le N$, that optimizes a certain figure of merit f for a given classification system. Typically, f corresponds to the misclassification rate, and it must be evaluated on the available S observations.

Indeed, both choices are interrelated. There is no overall optimal classifier, and the suitability of one or another is application dependent. Numerous examples in the literature provide comparisons of classification systems for different applications, e.g., [7–11]. Typically, a feature set is chosen beforehand and all classifier candidates are tested on it. The classifier providing the lowest f is adopted.

The selection of a feature subset can reduce not only the cost of recognition by reducing the number of features to be collected, but it also provides a better classification accuracy due to finite sample size effects ($S < \infty$), i.e., overfitting or the so-called curse of dimensionality [12]. The estimation of the optimal feature subset by feature

[*] Corresponding author. Tel.: +49 6151 16 70804; fax: +49 6151 16 3778.
E-mail addresses: rfandos@spg.tu-darmstadt.de (R. Fandos), cdebes@agtgermany.com (C. Debes), zoubir@spg.tu-darmstadt.de (A.M. Zoubir).

0165-1684/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.sigpro.2013.05.004

selection algorithms is a classical pattern recognition problem and is still an active field of research. A short summary of the most significant works in this respect is included in Section 4.1.

The optimal feature subset depends on the classification system. Therefore, a subset that performs well for one classifier might provide poor results for another one, but a second subset could outperform it. In short, it is not fair to compare different classifiers with the same feature subset. However, this is normally the case. In this paper, we propose a novel method that overcomes this issue by assessing the classifier performance without constraints to any specific feature set.

The prediction of the optimal number of features, n^*, for a given problem, has been an active field of research for several decades (see Section 4.1). Most approaches assume Gaussianity for the features and a common covariance matrix. Thus, to the best of our knowledge, no conclusive work exists. In practice, a rule of thumb suggesting that n^* should be between six and ten times smaller than S is generally applied [13]. This rule considers neither the quality of the available features nor the possible unbalance between the classes.

Knowing n^* beforehand allows for saving computational time when a feature selection algorithm is employed in order to decide for the optimal feature subset $\mathbf{t}^*$. Furthermore, if n^* is close to N, this might indicate that the available N features do not describe the problem sufficiently, and if possible, more or better features should be extracted.

In this paper, we present a novel algorithm, which is based on resampling techniques. Its purpose is twofold: on one hand, it assesses the performance of a classifier avoiding bonds to any feature subset. By doing so the best classifier out of a set of possible classifiers can be determined without restricting the procedure to a specific set of features that is suboptimal for most classifiers. On the other hand, it estimates the probability distribution of the optimal number of features n^* subject to a certain figure of merit f. This allows to predict the region in which the optimal number of features will be with a preset confidence. It further allows to infer confidence information on the figure of merit. Unlike previous works, no assumption for the features distribution is required.

Resampling techniques, e.g. the bootstrap, are computationally intensive tools for statistical inference in situations when either little is known about the data statistics or the available amount of data is too small to allow asymptotics based tools [14]. In the field of pattern recognition, the bootstrap has been thoroughly employed for addressing a variety of issues. A common application is the estimation of a reliable misclassification rate from a small number of observations [15–19,8] or when analytic expressions cannot be obtained. Bootstrap techniques have also been used in the context of feature selection [20–22].

Neither error estimation nor feature selection is the objective of the resampling algorithm proposed in this paper. Furthermore, there is a fundamental difference between error estimation bootstrapping and the method proposed hereafter. While the former resamples the data observations, our method resamples the features. To the best knowledge of the authors, there exists no previous work where resampling has been employed for classifier quality assessment and estimation of the optimal feature set dimensionality.

The paper is organized as follows. Section 2 is devoted to the resampling principle. In Section 3, we employ resampling to estimate the distribution of the figure of merit subject to the dimensionality of the feature set. Based on this, a quality assessment for classifier systems is proposed. An algorithm that predicts n^* is provided in Section 4.2, after a description of the state of the art in Section 4.1. Results are presented in Sections 5 and 6, for synthetic and real data, respectively. For each data set, the performance of three classifiers is compared according to the proposed quality assessment. Subsequently, confidence intervals for the optimal dimensionality are estimated. In order to verify the effectiveness of the proposed techniques, an exhaustive search of the overall optimal feature subset is performed for the synthetic data. For the real data, two well-established feature selection techniques, the Sequential Forward Selection (SFS) and the Sequential Floating Forward Selection (SFFS), have been applied. The dimensionality and performance of $\mathbf{t}^*$ is compared with the predicted ones. We conclude the paper with Section 7.

2. Resampling techniques

The resampling technique proposed in this paper is similar to the bootstrap. In this section, the standard bootstrap is described. Subsequently, the differences between it and the resampling method employed thereafter are highlighted.

The bootstrap is a technique that allows for statistical inference of parameters when few data observations are at hand or too little is known about the statistics of the problem. Despite being computationally demanding, it has gained importance in the last years, as the available computer power is exponentially increasing [23].

Let $\mathbf{x}=\{x_1,x_2,\dots,x_N\}$ be a set of measurements, which are realizations of a random variable set $\mathbf{X}=\{X_1,X_2,\dots,X_N\}$, drawn from a distribution $p_{\mathbf{X}}$. Typically, one is interested in the distribution of some parameter estimator $\hat{\theta}=\hat{\theta}(\mathbf{X})$. For example for the mean $\mu_{\mathbf{X}}$, we could be interested in the distribution of the sample mean: $\hat{\theta}=\hat{\mu}_{\mathbf{X}}=(1/N)\sum_{j=1}^{N}X_j$. If $p_{\mathbf{X}}$ is known, it is possible to exactly evaluate the distribution of the parameter estimator $\hat{\theta}$, $p_{\hat{\theta}}$. However, if $p_{\mathbf{X}}$ is unknown or $\hat{\theta}$ is some complicated estimator, its distribution cannot be derived in a closed form. Provided that enough data is available, asymptotic arguments could be used and the distribution of $\hat{\theta}$ could be approximated. If this is not the case, we may apply the bootstrap.

The bootstrap paradigm dictates that the unknown distribution $p_{\mathbf{X}}$ is approximated by the empirical distribution of the data $\hat{p}_{\mathbf{X}}$. Hence, B bootstrap samples $\mathbf{x}'_b=\{x'_1,x'_2,\dots,x'_n\}$, $1\le b\le B$, are generated from $\mathbf{x}$ by drawing at random with replacement. For each sample $\mathbf{x}'_b$ a bootstrap parameter estimate, $\hat{\theta}'_b=\hat{\theta}(\mathbf{x}'_b)$, is obtained. Thus, the distribution of $\hat{\theta}$, $p_{\hat{\theta}}$, is approximated by the distribution of $\hat{\theta}'$, $p'_{\hat{\theta}}$, provided a large number B of