Expert Systems With Applications 90 (2017) 40–49

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A semi-supervised learning approach for model selection based on class-hypothesis testing

Juan M. Gorriz[a,c,*], Javier Ramirez[a], John Suckling[c], F.J. Martinez-Murcia[a], I.A. Illán[a], F. Segovia[a], A. Ortiz[b], D. Salas-González[a], D. Castillo-Barnés[a], C.G. Puntonet[d]

[a] Department of Signal Theory, Networking and Communications, Avda. Fuentenueva s/n 18071, University of Granada, Spain
[b] Department of Communication Engineering, Campus de Teatinos s/n 29071, University of Málaga, Spain
[c] Department of Psychiatry, Robinson Way, CB2 0SZ, University of Cambridge, UK
[d] Department of Computer Architecture and Technology, C/ Daniel Saucedo s/n 18071, University of Granada, Spain

ARTICLE INFO

Article history:
Received 3 April 2017
Revised 13 July 2017
Accepted 1 August 2017
Available online 5 August 2017

Keywords:
Statistical learning and decision theory
Semi-supervised learning
Support vector machines (SVM)
Hypothesis testing
Partial least squares

ABSTRACT

This paper deals with the topic of learning from unlabeled or noisy-labeled data in the context of a classification problem. In the classification problem the outcome yields one of a discrete set of values thus, assumptions on them could be established to obtain the most likely prediction model at the *training stage*. In this paper, a novel case-based model selection method is proposed, which combines hypothesis testing from a discrete set of expected outcomes and feature extraction within a cross-validated classification stage. This wrapper-type procedure acts on fully-observable variables under hypothesis-testing and improves the classification accuracy on the test set, or keeps its performance at least at the level of the statistical classifier. The model selection strategy in the cross validation loop allows building an ensemble classifier that could improve the performance of any expert and intelligence system, particularly on small sample-size datasets. Experiments were carried out on several databases yielding a clear improvement on the baseline, i.e., SPECT dataset $Acc = 86.35 \pm 1.51$, with $Sen = 91.10 \pm 2.77$, and $Spe = 81.11 \pm 1.61$. In addition, the CV error estimate for the classifier under our approach was found to be an almost unbiased estimate (as the baseline approach) of the true error that the classifier would incur on independent data.

1. Introduction

Statistical learning theory (SLT) is a recently developed area in statistics that has been successfully applied to several fields including machine learning and artificial intelligence (Hastie, Tibshirani, & Friedman, 2001; James, Witten, Hastie, & Tibshirani, 2013; Vapnik, 1998). From least squares methods for linear regression, proposed in the very beginning of the nineteenth century, to the novel advances in machine learning such as random forests, Support Vector Machines (SVM), bagging or boosting in the early 90s (Breiman, Friedman, Olshen, & Stone, 1984; Hastie et al., 2001; Vapnik, 2000), SLT has become a new paradigm focusing on supervised and unsupervised modeling and prediction, i.e., the development of Computer-Aided Diagnosis (CAD) systems (Gur et al., 2004; Illán et al., 2010; Padilla, Lopez, Gorriz, & 2012; Suzuki, Li, Sone, & Doi, 2005). On the other hand, *decision theory* is the application of statistical hypothesis testing to the detection of signals in noise (Kay, 1993). Because under hypothesis testing we are essentially attempting to determine a desired pattern or to classify it as one of a set of possible patterns, it is also referred to as a pattern recognition or classification problem (Fukunaga, 1990).

The most common form of machine learning is the supervised learning (LeCun, Bengio, & Hinton, 2015). In this case, a quantitative response Y_k and several predictors $\{\mathbf{X}_k\}$ for $k = 1, \ldots, p$ are observed, and the aim is to discover the relationship among them, which can be written in a general form $Y = f(\mathbf{X}) + \epsilon$, where f is an unknown function of the predictors and ϵ is a random error term. In this way, supervised learning refers to a set of approaches for estimating f based on a set of known predictors and responses (James et al., 2013). When the supervised learning does not involve predicting a quantitative value but a qualitative response or *class*, this is known as a classification problem. In the latter case, once a

* Corresponding author at: Department of Signal Theory, Networking and Communications, Avda. Fuentenueva s/n 18071, University of Granada, Spain.
E-mail addresses: gorriz@ugr.es (J.M. Gorriz), javierrp@ugr.es (J. Ramirez), js369@cam.ac.uk (J. Suckling), fjesusmartinez@ugr.es (F.J. Martinez-Murcia), illan@ugr.es (I.A. Illán), fsegovia@ugr.es (F. Segovia), aortiz@ic.uma.es (A. Ortiz), dsalas@ugr.es (D. Salas-González), fjesusmartinez@ugr.es (D. Castillo-Barnés), carlos@atc.ugr.es (C.G. Puntonet).

http://dx.doi.org/10.1016/j.eswa.2017.08.006
0957-4174/

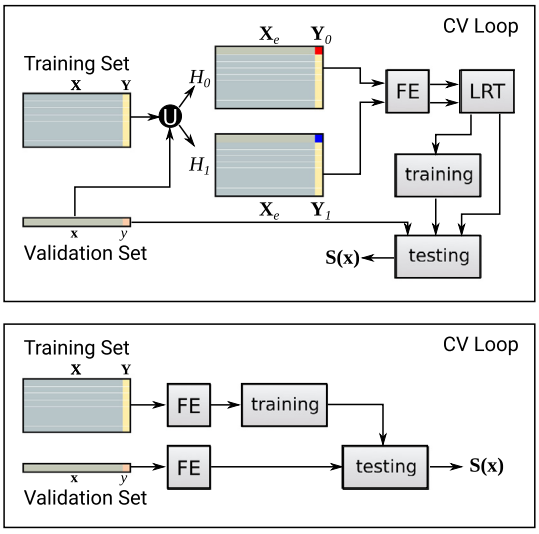

Fig. 1. Diagram of the model selection approach (up) versus the common supervised FE learning approach (bottom) on the training set.

final classifier $\hat{f}$ has been estimated, it can be used to predict the classes of the test samples.

Another field of research that have drawn the attention in the machine learning community is the semi-supervised learning (SSL) (Chapelle, Scholkopf and Zien, 2006). SSL belongs to the supervised category but in this case we have access to an additional unlabeled sample $\{\mathbf{X}_{p+j}\}$ for $j = 1, \ldots, m$ or samples with a few noisy initial labels (Lu, Gao, Wang, Wen, & Huang, 2015). In general, the solution to this subcategory is broadly based on two approaches: avoiding the use of unlabeled data or treating unobserved Y variables as a latent-class variables in the estimation of the system parameters (Chapelle et al., 2006). In recent years, several feature selection methods have been proposed based on Information Theory, filter (the one used in this paper as a baseline), embedded and wrapper methods (Guyon, Gunn, Nikravesh, Zadeh, & Eds, 2006) in fully-supervised data (Brown, Pocock, Zhao, & Luján, 2012). Under these approaches the features $\mathbf{X}$ are selected by quantifying the information that they share with the class variable Y. However, on partially-labeled datasets surrogate variables can be introduced to derive ranking equivalent approaches using all available information in an entirely classifier-independent and inference-free fashion (Sechidis, 2015). Some surrogate approaches assume the label of the unobservable variable Y and are found to be valid and informed to perform hypothesis testing for feature selection (Sechidis, Calvo, & Brown, 2014).

1.1. General outline

Before estimating the classifier f, relevant and non-redundant features are usually extracted from the raw data to facilitate the subsequent learning and generalization steps (Varol, Gaonkar, Erus, Schultz, & Davatzikos, 2012). Based on the previous ideas and feature extraction (FE) schemes, we investigate the possibility of using a semi-supervised model selection algorithm based on hypothesis testing applied to the responses or outcomes.

In Fig. 1, we show the differences between our methodology (up) and the baseline (bottom) at the training stage to derive the classifier $\hat{f}$. Learning from data samples involves, at this stage, model fitting by the use of observed variables and their labels (outcomes) that are grouped into two groups, the training and the validation sets. The use of a validation set at the training stage allows to select the classifier whose actual risk $S(x)$ is minimal, i.e., by parameter tuning. Finally, a test set can be employed only to assess the performance (generalization) of a fully-specified classifier and to avoid overfitting (Ripley, 1996).

FE may be applied to surrogate variables, shown in the latter figure as the class-information $Y_{0,1}$, within the Cross Validation (CV)-loop to obtain extended feature datasets by hypothesizing on the unknown outcomes of the validation patterns. The statistical consequences in each hypothesis could be analyzed in terms of probability within a Bayesian framework as proposed in this paper (the *likelihood ratio test* (LRT) block in Fig. 1), that is, in a classifier-independent fashion unlike embedded or wrapper methods. Other possibility is to evaluate the classifier configuration derived from the feature datasets, i.e., probability map of the support vectors (Padilla, et al., 2012). The influence of the validation pattern on the prediction models, i.e., a trained SVM, will depend on the *relevance* of the features that represent the validation samples in the feature space (Chapelle et al., 2006). Assuming the feature to be relevant, a decision function can be formulated in terms of class-conditional probabilities following the Neyman–Pearson (NP) lemma. The resulting LRT is similar to the one achieved by the classical linear discriminant (LDA) or quadratic DA analysis, but evaluated on two different feature datasets. Finally, the overall system can be seen as a wrapper-type method since although the feature selection is classifier-independent the system builds the final ensemble classifier based on a maximization process on several feature subsets (Martinez-Murcia, Górriz, & Ramírez, 1999).

This paper is organized as follows. In Section 2, a background to the NP approach to signal detection is provided. In the following Section 3 classical FE methods, such as Least Squares (LS) and Partial LS methods, are applied on semi-supervised datasets to obtain two feature extractions of the training database. As a result, hypothesis testing theory is employed to provide a novel framework for model selection as a part of the general tools of assessing statistical accuracy, such as CV or Bootstrap methods (Varma & Simon, 2006). The resulting LRT is the optimal tradeoff between type *I&II* errors which is employed for FE under certain assumptions. The set of assumptions comprises Gaussian modeling for conditional probabilities, feature relevance, and statistical independence among feature components. Finally, in Section 5, a fully experimental framework is provided to demonstrate the benefits of the proposed approach acting on baseline filter-based approaches, i.e., using LS and PLS FE methods and a SVM learning algorithm that minimizes the *leave-one-out* (LOO)-CV error. In Section 6, conclusions are drawn.

2. NP approach to signal detection: A background

Assume we observe a training set of random variables $\mathbf{Z} = \{\mathbf{X} \in \mathbb{R}^p, Y \in \mathbb{R}\}$. The realization of the outcome variable Y is modeled by normal distributions $\mathcal{N}(\mu_i, \sigma_i)$, with mean μ_i and variance σ_i for $i = 1, \ldots, C$, where C denotes the number of outcomes or classes. In general, we must therefore determine if $\mu = \mu_i$ for a single observation under a multiple C-ary hypothesis testing using a NP criterion. However, this is hardly used in practice, and the minimum probability error criterion is used instead (Kay, 1993). For a binary problem the test is defined as:

$$H_0 : \mu = 0; \quad H_1 : \mu = 1 \tag{1}$$

where every possible value of μ is thought as one of two competing hypotheses. In terms of the observed variable $\mathbf{x}$, the hypothesis can be reformulated as:

$$H_0 : \mathbf{x}; Y = 0 \quad H_1 : \mathbf{x}; Y = 1 \tag{2}$$

Thus, we are implicitly assuming that the hypothesis testing on the unobserved class Y can be reformulated in terms of the observed pattern value $\mathbf{x}$, via the joint pdf, $p(\mathbf{X}, Y)$ or an unknown function, f: $\mathbf{X} \mapsto Y$, an issue that is common in signal detection problems