Feature Selection in Pathology Detection Using Hybrid Multidimensional Analysis

Edilson Delgado-Trejos
National University of Colombia, Colombia

Germán Castellanos
National University of Colombia, Colombia

Luis Gonzalo Sánchez
National University of Colombia, Colombia

Julio F. Suárez
National University of Colombia, Colombia

INTRODUCTION

The basic problem of pathology detection is to classify a given observation in one of *m* known classes. A set of features presumably contains enough information to distinguish among the classes. Traditional statistical methods break down partly because of the increase in the number of observations, but mostly because of the increase in the number of variables associated with each observation (Castellanos, Daza, Sánchez, Castrillón, & Suárez , 2006).

Performance in training of pattern recognition systems to detect pathologies can be increased, if proper feature extraction is done. Training procedures usually deal with a high number of features; nevertheless a high dimension input space means significant processing time, higher cost in the collection of biosignal records since more observations are needed, and the well known curse of dimensionality phenomena (Lee, Lendasse, & Verleysen, 2003). As a result, the whole training performance declines. In this sense, effective feature selection should be carried out to select those features with higher discriminant capability, keeping or even increasing the accuracy of classification procedures (Yu & Liu, 2004). Literature shows wrapper type procedures of feature selection that search in a considerably smaller number of subsets. Sequential Forward Floating Selection (SFFS) is a common example found in this type of procedure (Webb, 2002). High-dimensional datasets present many mathematical challenges as well as some opportunities, and are bound to give rise to new theoretical developments (Fodor, 2002). Dimensional reduction techniques such as Principal Component Analysis (PCA), which is an orthogonal representation of data, allow us in some way to find a lower dimension in transformed space based on data variance, but this procedure does not take into account information about class separability, therefore PCA is not suitable because the direction of maximum variance does not necessarily correspond to the direction of maximum separability (Hyvarinen, 1999).

Dimensionality reduction procedures perform well on sets of correlated features while variable selection methods perform poorly. These methods fail to pick relevant variables because the score they assign to correlated features is too similar, and none of the variables is strongly preferred over another. Hence, feature selection and dimensionality reduction algorithms have complementary advantages and disadvantages. Dimensionality reduction algorithms thrive on the correlation between variables but fail to select informative features from a set of more complex features. Variable selection algorithms fail when all the features are correlated, but succeed with informative variables (Wolf & Bileschi, 2005).

In this work, we propose a feature selection algorithm with heuristic search that uses Multivariate Analysis of Variance (MANOVA) as the cost function. This technique is put to the test by classifying hypernasal from normal voices of CLP (cleft lip and/or palate) patients. The classification performance, computational time, and reduction ratio are also considered by comparing with an alternate feature selection method based on the unfolding of multivariate analysis into univariate and bivariate analysis. The methodology is effective because it has in mind the statistical and geometrical relevance present in the features, which does not summarize the analysis of the separability among classes, but searches a quality level in signal representation.

BACKGROUND

Dimensionality reduction is the representation of high dimension patterns in a subspace of less dimensionality based on either a linear or nonlinear transformation. This transformation optimizes a specific criterion in the subspace (Cohen, Tian, Zhou, & Huang, 2002). There is a concept responsible for directing the context of reduction called *relevance*, according to a given *cost function*.

Definition 1 (Cost function): Let the initial set of features be $\boldsymbol{\xi} = \{\xi_i : i = 1, \ldots, p\}$, where the estimated values for ξ_i constitute an element of multidimensional representation called observation. The set of observations is composed by $\mathbf{X} = \{\mathbf{x}_\ell \in \mathbb{R}^p : \ell = 1, \ldots, n\}$ contained in $span\{\mathbf{X}\}$; where with linear or nonlinear transformations spaces of representation $\boldsymbol{\mathcal{X}}_j = \mathcal{G}_j\{\mathbf{X}\} = \{\boldsymbol{x}_\ell \in \mathbb{R}^M : \ell = 1, \ldots, n\}$ are obtained. These spaces of representation are composed by subsets of features $\hat{\boldsymbol{\xi}}_j = \{\zeta_i : i = 1, \ldots, M\}$, such that, $M \le p$ y $j \in \mathbb{N}$. Let $\mathbf{k} = \{k_r : r \in \mathbb{N}\}$ be the set of class labels such that each observation $\boldsymbol{x}_\ell$ is assigned one and only one class label. A function that according to an associated metric finds a corresponding value to the evaluated data is called cost function. This function can be described as:

$$f_{\hat{\xi}_j} : \mathbb{N} \times \mathbf{H}_x \to \mathbb{R}$$
$$(\mathbf{k}, \boldsymbol{\mathcal{X}}_j) \mapsto f_{\hat{\xi}_j}(\mathbf{k}, \boldsymbol{\mathcal{X}}_j)$$

where $\mathbf{H}_x \subseteq \mathbb{R}^M$. In this article, we will take $f_{\hat{\xi}_j}(\mathbf{k}, \boldsymbol{\mathcal{X}}_j) \equiv f_{\hat{\xi}_j}(\mathbf{k}, \hat{\xi}_j)$ to aid notation.

Definition 2 (Relevance): Let the observation set be $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_\ell \in \mathbb{R}^M : \ell = 1, \ldots, n\}$ composed by M features, $\hat{\boldsymbol{\xi}} = \{\xi_i : i = 1, \ldots, M\}$. Let $\mathbf{k} = \{k_r : r \in \mathbb{N}\}$ be the set of class labels, such that each observation $\boldsymbol{x}_\ell$ is assigned one and only one class label. Let δ be a threshold of significance according to a metric (geometrical, statistical, etc). $\boldsymbol{\mathcal{X}}$ is relevant according to a metric of significance, if there is a cost function, $f_{\hat{\mathbf{i}}}$ with the associated metric of significance, such that, the obtained value for $f_{\hat{\xi}}(\mathbf{k}, \hat{\boldsymbol{\xi}}) > \delta$.

Observation 1 (Feature selection): It is a particular form of reducing dimensionality, when the obtained cost function value, $f_{\hat{\xi}}$, associated with a metric, is involved in the relevance criteria used in a mapping or transformation function and a feature subset is obtained. This feature subset directly corresponds to the initial space ξ and is oriented towards maximizing the representation capability as well as minimizing cost.

In general, given a set of observations where each observation is associated to one and only one class label from a set of labels $\mathbf{k}$, the problem of feature selection is to find a subset $\hat{\boldsymbol{\xi}}_i \subseteq \boldsymbol{\xi}$, such that, if the cardinal of $\hat{\boldsymbol{\xi}}_i$ is M, and all the M-cardinal subsets are in $\breve{\boldsymbol{\xi}}$, the subset $\hat{\boldsymbol{\xi}}_i$ is that which optimizes a cost function $f_{\hat{\xi}_i}$, such as (Webb, 2002):

$$f_{\hat{\xi}_i}(\mathbf{k}, \hat{\boldsymbol{\xi}}_i) = \max_{\breve{\xi} \subseteq \xi} f_{\breve{\xi}}(\mathbf{k}, \hat{\boldsymbol{\xi}}_i) \qquad (1)$$

Additionally, feature selection according to the cost function can be carried out in two ways (Yu & Liu, 2004):

1. *Wrapper:* When $f_{\hat{\mathbf{i}}_i}$ uses information of the classification procedure in an attempt to minimize classifier error with validation observations (different to the observations used in training).
2. *Filter:* It is a preprocessing of the data based on the optimization of $f_{\hat{\mathbf{i}}_i}$ with regard to a given metric. Thus, the irrelevant, redundant, and non discriminant variables are discarded. This process is independent of the classification procedure and has advantages such as simple implementation.

Feature selection can be carried out by means of extensive or heuristic search algorithms. Extensive search algorithms are supposed to obtain an optimal feature subset after searching the input training space thoroughly, being computationally expensive (Dash & Liu, 2003). On the other hand, heuristical algorithms which are based on empirical rules can reduce the computational complexity, even though the final subset may not be optimal but enough for the classification purpose. Such methods require some stopping criteria (cost function) (Webb, 2002).

Related Content