Proceedings of the 28th IEEE
EMBS Annual International Conference
New York City, USA, Aug 30-Sept 3, 2006

SaD01.4

Feature Selection in Pathology Detection using Hybrid Multidimensional Analysis

G. Castellanos, E. Delgado, G. Daza, L. G. Sánchez, J. F. Suárez

Control and Digital Signal Processing Group - National University of Colombia
{cgcastellanosd, edelgadot, gdazas, lgsanchezg, jfsuarezc}**@unal.edu.co**

Abstract— **Heuristical algorithms can reduce the computational complexity. Such methods require of some stoping criteria (cost function). Some of these cost functions are based on statistics like univariate and multivariate methods of analysis. Dimensional reduction techniques such as *Principal Component Analysis* (PCA) allow to find a lower dimension transformed space based on data variance, but this procedure does not take into account information about classes separability, the direction of maximum variance does not necessarily correspond to the direction of maximum separability. 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 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 the comparison with an alternate feature selection method founded on unfolding the multivariate analysis into univariate and bivariate analysis.**

I. Introduction

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 of the collected biosignal records since more observations are needed, and the well known *curse of dimensionality* phenomena [1]. 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.

Effective feature selection can be carried out by means of extensive search or heuristic search algorithms [2]. Extensive search algorithms are supposed to obtain an optimal feature subset after searching the input training space thoroughly, being computationally expensive. On the other hand, heuristical algorithms which are based on empirical rules can reduce the computational complexity, though the final subset could not be optimal but enough for the classification purpose. Such methods require of some stoping criteria (cost function). Often, these cost functions are based on statistics, like the univariate and multivariate methods of analysis. Dimensional reduction techniques such as *Principal Component Analysis* (PCA), which is an orthogonal representation of data, that in some way, allow to find a lower dimension transform space based on data variance, but this procedure does not take into account information about classes separability, thus, PCA is not suitable because the direction of maximum variance does not necessarily correspond to the direction of maximum separability [3].

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 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 the comparison with an alternate feature selection method founded on univariate and bivariate analysis.

II. Dimensional reduction by heuristic multivariate analysis

Given a p-dimensional initial training feature space, a sequential multivariate analysis is done. This is how an heuristic search method and its cost function are structured. They provide the search direction and subset construction rules.

Among heuristic search methods, forward sequential search is preferred in which each current stage adds one feature at the time. In general, floating techniques might be considered; they based on combining forward and backward selection try to provide a better subset optimizing a cost function. For instance, Δp_+ features are added and $\Delta p_- > \Delta p_+$ discarded [4].

A. Statistical relevance measures

Let $\{k = 1, \dots, L\}$ be the set of patterns, where each pattern has n observations. Let $\boldsymbol{\xi} \in \mathbb{R}^p$ be the set of features, for which the subset $\hat{\boldsymbol{\xi}}_i = \boldsymbol{\xi} \cap \overline{\xi_i}$ is constructed, being $\overline{\xi_i}$ the complement of ξ_i in $\boldsymbol{\xi}$. A feature ξ_i is strongly relevant for the given cost function $f_{\hat{\boldsymbol{\xi}}}$, if and only if,

$$f_{\hat{\boldsymbol{\xi}}}\left(k|\xi_i, \hat{\boldsymbol{\xi}}_i\right) \neq f_{\hat{\boldsymbol{\xi}}}\left(k|\hat{\boldsymbol{\xi}}_i\right) \quad (1)$$

Multivariate analysis of variance (MANOVA) can be used as the cost function (1), here the relevance measure is the separability between patterns for a given feature space. This is an hypothesis test about the equality or inequality of the average values $\boldsymbol{m}_l$ for each one of the features. In this case, the model of multiple analysis of statistical relevance for each p-dimensional vector $\boldsymbol{x}_{ij}$

$$\boldsymbol{x}_{kj} = \boldsymbol{m}_k + \varepsilon_{kj},\ \boldsymbol{m}_{ki} = \boldsymbol{m} + \alpha_k$$

1-4244-0033-3/06/$20.00 ©2006 IEEE.

where j is the observation and k the patterns; $\boldsymbol{m}_k$ are the mean vectors for each pattern and ε_{kj} is the model perturbation, $\boldsymbol{m}$ is the overall mean and α is the perturbation across the overall mean.

The comparison of the mean vectors for the k patterns to find significant differences is carried out through the hypothesis test:

$$H_0 : \boldsymbol{m}_1 = \boldsymbol{m}_2 = \cdots = \boldsymbol{m}_L,$$
$$H_1 : \exists \boldsymbol{m}_l \neq \boldsymbol{m}_m; \forall l, m \in 1, \ldots, L$$

MANOVA does not find the features per se; it is only a way for validating a given set of features, that is why this is only a cost function. Consequently, the following algorithm to construct and evaluate feature subsets is proposed.

B. MANOVA progressive algorithm

The Wilks' test is often used for MANOVA. It consists on the likelihood ratio test of H_0 given by:

$$\Lambda = \frac{|\mathbf{E}|}{|\mathbf{E} + \mathbf{H}|} \tag{2}$$

which is know as Wilks' Λ; being $\mathbf{H}$ the hypothesis matrix, which can be understood as a measure of dispersion among the mean values of the patterns. While, $\mathbf{E}$ is the error matrix; that relates to a dispersion measure among observations within the patterns. H_0 is rejected if the dispersion among patterns is greater than the dispersion among observations within patterns, and so $\mathit{\Lambda} \in [0, 1]$ tends to be a zero. On the other hand, Wilks' $\mathit{\Lambda}$ can be similar to an F-statistic though in an inverse manner. A large F-statistic rejects H_0 [5].

These are the steps for a forward sequential search algorithm based on multivariate statistical relevance function:

1) Calculate the F-statistic (transformation of Wilks' $\mathit{\Lambda}$) for one-feature subsets. From these values, the feature with the largest F-statistic is chosen and its cumulative probability value is computed using the F distribution.
2) Construct 2-dimensional subsets, combining the feature previously chosen in step 1 with the remaining features. Each one of these subsets is evaluated through the Wilks' test, and its respective F-statistic is updated.
3) Select the 2-dimensional subset with the largest F-statistic. Calculate its respective cumulative probability using the F-distribution. This value must exceed the calculated value in step 1, to update the subset. Otherwise, terminate the search.
4) Construct the feature subset adding one feature to the updated subset. These new analysis groups correspond to the subset selected in step 3 and each one remaining feature.
5) Return to step 3 and update the subset using the same criteria. Continue to step 4 and repeat the updating process over and over. The algorithm stops as shown in step 3 when added features do not increases the cumulative probability. So, the final size of the subset is p' where $p' \leq p$. In this manner, we can select those features that joint are more discriminant.

Note that if during the calculations $\mathit{\Lambda} \to 0/0$, there is a linear dependency in the current subset. Then its F-statistic is forced to zero and so the evaluated subset is rejected.

III. Dimensional reduction by unfolding multivariate analysis

In this case, the multivariate analysis of statistical relevance is unfolded in multiple analyses of smaller dimension, particularly univariate and bivariate analysis are carried out, with these kind of analyses an alternative feature selection methodology is constructed, which connect in cascade 3 blocks as follows: univariate analysis of separability, bivariate analysis of correlation, and heuristic search using a classifier as the cost function.

A. Univariate analysis of separability

Because it is important to assure that each one of the features $\xi_i \subset \boldsymbol{\xi}; \forall i = 1, 2, \ldots, p$ have enough discriminant power, as the statistical relevance measure is proposed an hypothesis test based on significant differences of each feature between patterns:

H_0: There is not a significant difference on the observations in the ξ_i feature between patterns. Consequently, the overall difference is zero.

H_1: There is a significant difference on the observations in the ξ_i feature between patterns. Consequently, the overall difference is not zero.

In particular, the hypothesis test is based on the t-Student distribution. Let $\mathbf{x}_k \in \xi_i = \{x_j : j = 1, \ldots, n; k = 1, 2\}$ be the observation vectors for each k pattern, and unknown mean $m_{\mathbf{x}_k}$ and variance $\sigma^2_{\mathbf{x}_k}$.

From the set of n observations per pattern $\tilde{m}_{\mathbf{x}_k}$ and $\tilde{\sigma}^2_{\mathbf{x}_k}$, $k = 1, 2$ are estimated. Assuming gaussian distributions for $\mathbf{x}_k, k = 1, 2$ the confidence interval of the respective mean estimations, for a α significance level are given by,

$$\begin{aligned}(\tilde{m}_{\mathbf{x}_1} - \tilde{m}_{\mathbf{x}_2}) - t_{1-\alpha/2}\{n_1 + n_2 - 2\}\, \tilde{\sigma}_{\Delta \tilde{m}_{\mathbf{x}}} \leq \\ \leq m_{\mathbf{x}_1} - m_{\mathbf{x}_2} \leq (\tilde{m}_{\mathbf{x}_1} - \tilde{m}_{\mathbf{x}_2}) + \\ + t_{1-\alpha/2}\{n_1 + n_2 - 2\}\, \tilde{\sigma}_{\Delta \tilde{m}_{1\mathbf{x}}}\end{aligned} \tag{3}$$

being

$$\tilde{\sigma}^2_{\Delta \tilde{m}_{1\mathbf{x}}} = \frac{n_1 \tilde{\sigma}^2_{\mathbf{x}_1} + n_2 \tilde{\sigma}^2_{\mathbf{x}_2}}{n_1 + n_2 - 2} \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

where $t_{1-\alpha/2}\{n_1 + n_2 - 2\}$ represents the percentile value of t-distribution with $n_1 + n_2 - 2$ degrees of freedom (in this case, $n_1 = n_2 = n$). When the confiedence inteval given by (3) contains zero value, H_0 is not rejected. Otherwise, the alternative hypothesis H_1 is accepted. Thus, it is assumed there is a significant difference between patterns measures if H_0 is rejected.

As result, the dimension of the initial feature set is reduced to $\xi \subset \boldsymbol{\xi} \in \mathbb{R}^{p_1}$, $p_1 \leq p$.

B. Bivariate Analysis - Component Analysis

Let $\boldsymbol{\xi}_{1\times p_1}$ be a random vector of features with mean value $E\{\boldsymbol{\xi}\} = \boldsymbol{m}_\xi$, so that the covariance matrix $\boldsymbol{\Sigma_\xi}$, of size $p_1 \times p_1$ is,

$$\boldsymbol{\Sigma_\xi} = E\left\{(\boldsymbol{\xi} - \boldsymbol{m_\xi})(\boldsymbol{\xi} - \boldsymbol{m_\xi})^{\mathsf{T}}\right\}$$

and the correlation matrix P from $\boldsymbol{\Sigma_\xi}$ is defined as:

$$P(m,q) = \frac{\Sigma_\xi(m,q)}{\sqrt{\Sigma_\xi(m,m)\,\Sigma_\xi(q,q)}}$$

A null value in the correlation function between the variables ξ_m y ξ_q implies both variables are linearly independent. In the opposite case, the correlation matrix P tends to be singular.

Thence, we search for pairs of features that expose the larger correlation indexes than 0.5. From each one of this pairs we choose the feature with the smaller overall correlation. Therefore, the new feature subset has dimension p_2, where $p_2 \leq p_1 \leq p$.

C. Heuristic search

With the aim of finding a feature subset that minimizes the classification error, an heuristic search is structured. Precisely, we use a forward sequential search algorithm with a cost function based on a bayesian classifier. Hence, the new subset has dimension p', where $p' \leq p_2 \leq p_1 \leq p$

IV. Experimental background

A. Database and data preprocessing

The sample contains 90 children recordings (observations) labelled as *normal* and *hypernasal* (45 patients per pattern) and evaluated by specialists. Each recording is composed of 5 words of Spanish language: */coco/*, */gato/*, */jugo/*, */mano/* y */papá/*. Signals were acquired under controlled conditions for low ambient noise using a dynamic microphone (cardioid). All signals range between $(-1, 1)$. The extracted features are based on statistical moments relative to position, scale, and shape [6] per voice parameter. A total of 128 features per word were extracted. Consequently, the initial feature set is 640-dimensional. We do data preprocessing to reduce the influence of factors such as systematic acquisition errors, ocasional failures of recording devices, etc. Besides this guaranties the homogeneity of the statistical properties of the training features.

Data preprocessing consist in an hypothesis test (t-student) to generate a confidence interval for data should lie within; we discard features when less than 90% of the observations lie within the interval. Moreover, we look for normality in data; so we perform a Kolmogorov-Smirnov test to verify normality. Those features rejected by this test are transform using the Box-Cox transform and tested again. Features that are not normal after transformation are rejected.

B. Performance Analysis

Both methods proposed (section II and III) were compared in 3 ways: computational time T, feature reduction ratio RR, and classification accuracy CA. To prove the separability of the chosen subsets we used a bayesian classifier in cross-validation. Eventually, we could compare the results with the extensive search, but having 640 variables the number of subsets is $4.5624e + 192$.

V. Results

From the data preprocessing 36.7% of the original set was discarded by the first test. The percentage of features that could not be normalized was 25.9%. Once preprocessing was done, the initial set was reduce to 47.2%. If we go directly to classification after preprocessing the dimension is still too large and the bayesian classifier does not converge. Feature selection is needed.

Results of comparison between the two methods are shown in Table I, in this Table we can see how the computational time for the first method is longer than the second. Nevertheless, these times just illustrate the computational complexity, but this is not a parameter for choosing a method that is performed off-line. Even though the reduction ratio for the first method is a little less than the second, the classification accuracy is higher. It depends on the implementation requirements to choose any of the methods. Accuracy versus complexity should be the criteria.

TABLE I
Comparison of the 2 methods

Method	T [seg]	RR [%]	CA [%]
Dimensional reduction by heuristic multivariate analysis	19.7	91.26	90
Dimensional reduction by unfolding multivariate analysis	3.81	92.66	87.26

VI. Conclusions

The proposed hybrid methodologies (joint work among multivariate statistical analysis and features heuristic search methods), to reduce the initial training space dimension, showed their performance to identify pathologies. The methodology is effective because has in mind the statistical and geometrical relevancy present in the features, which does not resume the analysis to the separability among classes, but also searches a quality level in signals representation. La methodology employed by MANOVA progressive algorithm was the best, and analyze subsets generated by heuristic principles, to avoid exhaustive processes, proposing a decision based on a hypothesis testing with low computational cost.

The data preprocessing was a basic stage systems training to identify pathologies. In this particular case, the classifier performance does not converge, even employing the best kind of feature selection. In order to comply with all the assumptions of the statistical model imposes, data distribution must be properly analyzed forehand.

REFERENCES

[1] J. Lee, A. Lendasse, and M. Verleysen, “Nonlinear projection with curvilinear distances: Isomap versus curvilinear distance analysis,” *Neurocomputing*, vol. 57, p. 49–76, 2004.

[2] M. Dash and H. Liu, “Consistency-based search in feature selection,” *Artificial Intelligence*, vol. 151, pp. 155–176, 2003.

[3] A. Hyvarinen, “Survey on independent component analysis,” *Neural Computing Surveys*, vol. 2, pp. 94–128, 1999.

[4] P. A. Devijver and J. Kittler, *Pattern Recognition: A Statistical Approach*. Prentice Hall, 1982.

[5] A. C. Rencher, *Methods of Multivariate Analysis*, 2nd ed. New York: Wiley Series in Probability and Statistics, 2002.

[6] G. Castellanos, O. D. Castrillón, and E. Guijarro, “Multivariate analysis techniques for effective feature selection in voice pathologies,,” in CASEIB, 2004.