Feature Weighting and Selection Using a Hybrid Approach Based on Rademacher Complexity Model Selection

LF Giraldo[1,2], E Delgado[1], CG Castellanos[1]

[1]Control and Digital Signal Group, National University of Colombia
[2]Research Center for Control, Automation and Production, Andes University, Colombia

Abstract

This study proposes a hybrid feature weighting and selection model for reducing the system dimensionality, improving the classification accuracy. The hybrid selection model is tuned by means of genetic algorithms, where the involved evaluation uses the Rademacher complexity using the k-nearest neighbors classifier. This approach simultaneously minimizes the feature number and training error and provides information about the relevance of each feature. The model was tested on artificial databases as well as by using features extracted from cardiac signals. The used ECG records for ischemic detection correspond to the E-STT database and the used heart sound database for cardiac murmur detection corresponds to phonocardiographic (PCG) records assembled in the National University of Colombia. The classification error result in the ischemic detection was 1.3% with 50.7% of dimensionality reduction rate, while in the cardiac murmur detection was 6.9% with 87.3% of dimensionality reduction rate.

1. Introduction

The automatic detection of cardiac pathologies strongly depends on the appropriate feature selection (effective data representation), which mostly are related to timing, morphology and spectral properties of cardiac signals. Moreover, the cardiac signals have high within-class variability. Unfortunately, many of the candidate features are irrelevant to the target concept [1]. This procedure is known as feature selection, where the main purpose is to select the best subset of the input feature set, which has a problem related to the relevance measure election, since the reduced feature space should ideally contain the total intrinsic information, in such a way that the generalization capacity does not decrease. Feature weighting is a more general method which the original feature set is multiplied by a weight value proportional to the ability of the feature to distinguish pattern classes [2]. Modern heuristic search procedures, such as genetic algorithms, have been found effective to obtain near-optimal solutions in large-sized feature sets but nonlinear interactions add more complexity to the evaluation function design. In [3], a hybrid system is proposed using genetic algorithms and decision trees in order to reduce the feature number, obtaining error rates up to 16.9% and a feature reduction around 61%. In [2], a dimensionality reduction is performed using genetic algorithms and the k-nearest neighbor classifier on biochemical data, the feature space is geometrically transformed, obtaining error rates up to 1% and feature reduction of 85%. A simultaneous feature selection and feature weighting is performed in [4], using Tabu search and the k-nearest neighbor classifier. The main objective of this study is to find a reduced representation space of the normal and pathological cardiac dynamic that allows processing time reduction and improves the classification accuracy using the Rademacher model. According to this, a hybrid system formed by genetic algorithms, decision trees and the k nearest neighbor rule for developing feature selection and feature weighting is proposed. The genetic algorithm will generate parameter evolution by means of the Rademacher complexity minimization using a k-NN classifier. Thus, a feature subset (with low order and high discriminatory capability, both in parallel) is searched. The inclusion of Rademacher complexity included in the evaluation function will increase the generalization capacity of the classifiers, since an uncertainty component is added in the subset evaluation stage. Experiments and comparisons between the proposed model and well-known methods for feature selection have been performed on synthetic databases and cardiac-signal features, showing the effectiveness of this approach in pathologies recognition tasks.

2. Overview of Rademacher Complexity

Rademacher complexity is a measure proposed in [5] which attempts to balance the complexity of the model with its fit to the data by minimizing the sum of the training error and a penalty term. The computation of the Rademacher complexity is *data driven* which means that it depends on the distribution of the data and hence one can expect better performance for particular instances of learning problems.

Let $\{\mathbf{x}_i, y_i\}_{i=1}^n$ be a set of training instances, where $\mathbf{x}_i$ is the pattern or example associated with features $\{F_j\}_{j=1}^q$, and y_i is the label of the example $\mathbf{x}_i$. Let $h(\mathbf{x}_i)$ be the class obtained by the classifier h, trained using $\{\mathbf{x}_i, y_i\}_{i=1}^n$. Then, the training error is defined as:

$$\hat{e}(h) = \frac{1}{n}\sum_{i=1}^{n} I_{\{h(\mathbf{x}_i) \neq y_i\}}$$

where,

$$I_{\{h(\mathbf{x}_i) \neq y_i\}} = \begin{cases} 1, & \text{when } h(\mathbf{x}_i) \neq y_i \\ 0, & \text{when } h(\mathbf{x}_i) = y_i \end{cases}$$

Let $\{\sigma_i\}_{i=1}^n$ be a sequence of Rademacher random variables (i.i.d.) independent of the data $\{\mathbf{x}_i\}_{i=1}^n$ and each variable takes values +1 and -1 with probability $1/2$. According to this, the computation of the Rademacher complexity involves de following steps:

– Generate $\{\sigma_i\}_{i=1}^n$.
– Get a new set of labels, doing $z_i = \sigma_i y_i$.
– Train the classifier $h_{\mathcal{R}}$ using $\{\mathbf{x}_i, z_i\}_{i=1}^n$.
– Compute the Rademacher penalty, given by

$$\mathcal{R}_n = \left| \frac{1}{n}\sum_{i=1}^{n} \sigma_i I_{\{h_{\mathcal{R}} \neq y_i\}} \right|$$

– Train the classifier h, using $\{\mathbf{x}_i, y_i\}_{i=1}^n$.
– Compute the training error $\hat{e}(h)$.
– The Rademacher complexity RC is given by

$$RC = \hat{e}(h) + \mathcal{R}_n$$

3. Experimental setup

3.1. Dataset

3.1.1. Artificial data sets

A number of experiments and comparisons on well-known benchmark data sets (where the truly relevant features are known) were performed.

Monk-1. It is a two-classes, six discrete features $\{F_i\}_{i=1}^6$ and 432 instances dataset. Only F_1, F_2 and F_5 are relevant to the target concept.

Monk-3. It is a two-classes, six discrete features $\{F_i\}_{i=1}^6$ and 432 instances dataset. Only F_2, F_4 and F_5 are relevant to the target concept.

Syndata. It was generated by a dataset generator, available in [6]. Thus, a two-classes, fourteen real features $\{F_i\}_{i=1}^{14}$ and 500 instances dataset was obtained. Only F_1 to F_{10} are relevant to the target concept.

3.1.2. Real data sets: cardiac signal characterization

PCG database. Corresponds to features extracted from phonocardiographic (PCG) records: 50 subjects without valve disorders and 98 with evidence of cardiac murmurs, 8 records per subject (different auscultation areas). Signals were acquired at a sampling rate of 44.1 kHz with 16 bits per sample. 360 representative beats were chosen by 3 specialists: 180 normal and 180 with evidence of cardiac murmur. 93 features are derived from acoustical, spectral and fractal analysis. The PCG records belong to the National University of Colombia.

E-STT database. Corresponds to features extracted from the ECG records of the E-STT database, available in [7]. Signals were acquired at a sampling rate of 250 Hz with 12 bits per sample. 1800 representative beats were chosen: 900 considered normal beats and 900 beats with evidence of ischemia. 85 features are derived from wavelet analysis, diagnostic measures and nonlinear analysis.

3.2. Proposed method

The proposed method for feature selection is shown in Figure 1. The feature space reduction is carried out by inducing a decision tree (ID3 algorithm) [8] on the whole sample set and selecting only features used to build it. Then, an heuristic approach is applied, with the purpose of minimizing the classification error rate and searching a low order feature subset with high discriminatory power, both in parallel. This is done by using a genetic algorithm, which generates and allows the evolution of the method parameters using the classifier error (with the Rademacher penalty included) as evaluation function. The classifier is based on the k-nearest neighbor rule. The method parameters are:

– $\{w_j\}_{j=1}^q$: feature weights into the interval $[0,1]$, where w_j is the weight of the feature F_j. The feature space is geometrically transformed, improving the classification accuracy and giving a relevance level of each feature.
– θ_w: It is a decision threshold into the interval $[0,1]$. If $w_j < \theta_w$, then F_j is discarded.
– k: Number of nearest neighbors.
– $\mathcal{P}_m$: Mutation probability of the genetic algorithm.

Additionally, the genetic algorithm specifications are listed below:

Encoding. Taking into account the features set $\{F_i\}_{i=1}^q$, the method parameters are encoded into binary chromosomes, as follow:

$$\underbrace{0\ldots10}_{\substack{w_1 \\ 20\,bits}}\underbrace{0\ldots10}_{\substack{w_2 \\ 20\,bits}}\cdots\underbrace{0\ldots10}_{\substack{w_q \\ 20\,bits}}\underbrace{0\ldots10}_{\substack{\theta_w \\ 20\,bits}}\underbrace{0\ldots10}_{\substack{k \\ 4\,bits}}\underbrace{0\ldots10}_{\substack{\mathcal{P}_m \\ 20\,bits}}$$

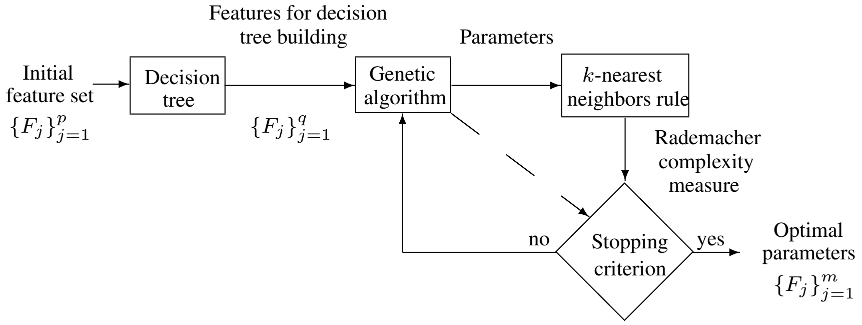

Figure 1. Proposed model for feature selection

Evaluation function. In this approach for evaluating feature subsets is used the Rademacher complexity as a classification error measure (from k-NN classifier).
Mutation. In literature it is found that varying the mutation rate during the run better results are given [9, 10]. Therefore, an isotropic self-adaptive mutation is used, where the mutation probability is encoded into each chromosome.
Other parameters. In [11] and [12] are recommended experimental values for adjusting the genetic algorithm parameters: crossover rate of 0.95, population size of 40 chromosomes, a generation gap of 95% and 300 generations. The algorithm will end when all generations are accomplished.

The proposed model is tested using an accuracy measure (i.e., the classification error rate) as evaluation function, in order to observe the ability of the Rademacher complexity for addressing the genetic algorithm search. Also, it is estimated the predictive accuracy by five replications of two-fold cross-validation ($5 \times 2cv$) [13]. Each time a two-crossvalidation is performed, the procedure is executed independently in each fold and it has no access to the another fold: the accuracy of the selected subset in each execution is measured in the no accessed fold. In this way, the reported accuracy will be the mean of the ten accuracies and the standard deviation, since the $5 \times 2cv$ scheme is used. With the aim of observe the generalization ability of the model, a classification noise on the training folds with probability of 15% is applied. At least, all dataset elements were divided into two groups: Seventy percent for the processing stage and thirty percent for method verification. In this way, the method is verified by completely unknown elements.

3.3. Comparison with other methods

In order to evaluate the effectiveness of the proposed method on cardiopathies detection, experiments over artificial data sets and cardiac-signal feature sets were carried out. Initially, the classification results (using k-NN and ID3 algorithm) were obtained over all data sets without feature selection. After, comparisons between the proposed method and conventional methods were also performed. The conventional methods are listed below [14]:

Sequential Forward Selection (SFS). This procedure selects the best single feature and adds one feature at a time which in combination with the selected features maximizes the criterion function. Once a feature is retained, it cannot be discarded. This routine stops when criterion function cannot be maximized adding another feature.

Sequential Forward Floating Search (SFFS). First enlarge the feature subset by *l* features using forward selection and then delete *r* features using backward selection. The values of *l* and *r* are determined automatically and updated dynamically. It provides a result close to the optimal solution.

Branch and Bound Search(B&B). Uses the branch-and-bound search method; only a fraction of all possible feature subsets need to be enumerated to find the optimal subset.

4. Results

Tables 1 and 2 show the experimental results: the classification error rate using $5 \times 2cv$ and the dimensionality reduction rate, respectively. Where $\mathrm{PM}_{\not{\mathcal{R}}}$ is the proposed methodology (ID3/GA/k-NN) using the training error as evaluation function, while $\mathrm{PM}_{\mathcal{R}}$ is the proposed methodology (ID3/GA/k-NN) using the Rademacher complexity as evaluation function. Table 3 shows the verification results of the proposed scheme with/without the Rademacher complexity. It is notable that the generalization capability was increased when the Rademacher penalty was included in cardiac pathology detection (ischemia and cardiac murmurs). In this table, ε_v is the verification error.

Table 3. Verification error rate

		$5 \times 2cv$ (%)	ε_v (%)
$\mathrm{PM}_{\mathcal{R}}$	Ischemia detection	1.3±0.3	2.1
	Cardiac murmur detection	6.9±1.9	8.2
$\mathrm{PM}_{\not{\mathcal{R}}}$	Ischemia detection	2.0±0.6	10.8
	Cardiac murmur detection	7.2±3.8	18.3

Table 1. Classification error rate

Database	1-NN	ID3	SFS/1-NN	SFFS/1-NN	B&B	$\mathrm{PM}_{\mathscr{R}}$	$\mathrm{PM}_{\mathcal{R}}$
Monk1	29.4±2.7	27.8±2.2	14.7±3.6	16.1±2.7	22.1±3.7	9.1±5.7	8.5±2.2
Monk3	31.7±0.6	36.1±1.7	15.6±3.8	20.0±5.7	18.3±6.8	4.1±2.1	3.3±1.2
SynthData	15.0±1.7	24.4±2.1	16.0±6.2	14.6±1.5	17.2±3.5	4.5±1.8	3.6±1.5
E-STT	15.6±1.3	28.6±1.2	14.9±1.2	14.5±0.9	15.9±2.3	2.0±0.6	**1.3±0.3**
PCG	18.3±4.5	26.6±0.3	18.2±0.9	18.9±3.2	23.7±2.9	7.2±3.8	**6.9±1.9**

Table 2. Dimensionality reduction rate

Database	1-NN	ID3	SFS/1-NN	SFFS/1-NN	B&B	$\mathrm{PM}_{\mathscr{R}}$	$\mathrm{PM}_{\mathcal{R}}$
Monk1	0.0	50.0	31.7	31.7	57.1	59.2	53.3
Monk3	0.0	33.3	35.0	30.0	57.1	68.5	65.0
SynthData	0.0	14.3	55.3	48.0	33.3	35.7	29.3
EST-T	0.0	72.9	77.1	71.9	80.1	89.1	**50.7**
PCGset	0.0	35.7	74.7	77.4	87.7	69.8	**87.3**

5. Conclusions

– The effectiveness of error penalization by using the Rademacher model has been proved to be of high value for automatic detection of cardiac pathologies, giving the best classification results among the considered models.
– The results show that the Rademacher penalty adds generalization capacity to the classifier, which is a necessary constraint due to the high within-class variability of cardiac signals. This uncertainty included in the feature selection allows an effective dimensionality reduction.
– The resultant weighting vector allows to determine the relevance level on each selected feature, which helps to obtain a physical interpretation.

References

[1] Jain AK, Duin RPW, Mao J. Statistical pattern recognition: a review. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(1):4–37.

[2] Raymer M, Punch W, Goodman E, Kuhn L, Jain A. Dimensionality reduction using genetic algorithms. IEEE Transactions on Evolutionary Computation 2000;4(2):164–171.

[3] Bala J, Huang J, Vafaie H, DeJong K, Wechsler H. Hybrid learning using genetic algorithms and decision trees for pattern classification. In IJCAI. 1993; 719–724.

[4] Tahir MA, Bouridane A, Kurugollu F. Simultaneous feature selection and feature weighting using hybrid tabu search/k-nearest neighbor classifier. Pattern Recogn Lett 2007; 28(4):438–446.

[5] Koltchinskii V. Rademacher penalties and structural risk minimization. IEEE Transactions on Information Theory July 2001;47(2):1902–1914.

[6] Melli G. Data set generator. http://www. datasetgenerator.com/.

[7] PhysioNet. European ST-T database. http://www. physionet.org/physiobank/database/edb/.

[8] Quinlan JR. Induction of decision trees. Machine Learning 1986;1(1):81–106.

[9] Eiben A, Hinterding R, Michalewichz. Z. Parameter control in evolutionary algorithms. IEEE Transactions on Evolutionary Computation 1999;3(2):124–141.

[10] Bäck T, Eiben AE, van der Vaart NAL. An empirical study on gas ẅithout parameters.. In PPSN VI: Proceedings of the 6th International Conference on Parallel Problem Solving from Nature. London, UK: Springer-Verlag. ISBN 3-540-41056-2, 2000; 315–324.

[11] DeJong KA. An analysis of the Behavior of a class of Genetic Adaptative System. Ph.D. thesis, University of Michigan, 1975.

[12] Grefenstette JJ. Optimization of control parameters for genetic algorithms. IEEE transactions on Systems Man and Cybernetics 1986;16(1):122–128.

[13] Alpaydin E. Combined 5 x 2 cv f test for comparing supervised classification learning algorithms. Neural Computation 1999;11(9):1885–1892.

[14] Dash M, Liu H. Feature selection for classification. Intelligent Data Analysis Elsevier 1997;1:131–156.

Address for correspondence:

Name: Edilson Delgado-Trejos
Full postal address: Universidad Nacional de Colombia. Campus La Nubia. Vía al aeropuerto. Oficina V-212. Manizales - Caldas. Colombia. Tel: +57 3007809495.
E-mail address: edelgadot@unal.edu.co