Enhancing Feature Extraction for VF Detection using Data Mining Techniques

A Rosado-Muñoz, G Camps-Valls, J Guerrero-Martínez,
JV Francés-Villora, J Muñoz-Marí, AJ Serrano-López

Grup Processament Digital de Senyals (GPDS), Universitat de València. València, Spain.

Abstract

Previous studies developed by the authors proposed VF detection algorithms, including VT discrimination, based on time-frequency distributions. However, due to the large number of parameters extracted from the distributions, efficient schemes for parameter selection and significance estimation are needed. This study proposes a combined strategy of classical and modern techniques for the selection of parameters to develop improved VF detection algorithms. We show how exhaustive exploration of the input space using data mining techniques simplifies and improves the solution and reduces the computational cost of detection algorithms. Jointly with classical selection techniques (correlation, Wilks' Lambda, statistical significance), other approaches are used (PCA, SOM-Ward and CART). We show that better results are achieved using less number of parameters than previous VF detection algorithms.

1. Introduction

Detection of VF at an early stage is a crucial point in order to lower the risk of sudden death and allow the specialist to have greater reaction time to give the patient a good recovering therapy. Previous studies developed by the authors proposed Ventricular Fibrillation (VF) detection algorithms, including Ventricular Tachycardia (VT) discrimination, based on time-frequency (t-f) distributions [1, 2]. Time-frequency distributions provide information in both time and frequency domains, they show spectral changes along time, which is useful in non-stationary signals as in ventricular arrhythmias. Fig. 1 illustrates a time-frequency representation of a VF segment. Once the distribution is obtained, different parameters are calculated. They take advantage of simultaneous time and frequency measures, giving indication of power in time and frequency.

Detection algorithms are based on a reduced set of t-f parameters. However, as new parameters are extracted from a t-f distribution, efficient schemes for parameter selection and significance estimation are needed in order to avoid redundancy. In this sense, a principled statistical framework for data analysis with high input space has become necessary. This need is covered by fields such as data mining (DM) and

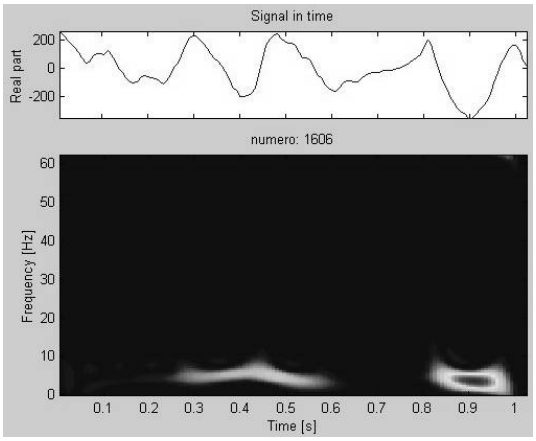

Figure 1. Time–frequency representation of a VF signal.

Knowledge Discovery in Databases (KDD). They are new frameworks for the extraction of knowledge from data and have recently captured the interest of scientists in many areas. These approaches combine classical statistical tools with state-of-the-art signal processing techniques such as neural networks, genetic algorithms, kernel-based methods, fuzzy logic and advanced visualization tools like Self-Organizing Maps and clustering algorithms. A full detailed overview of all the steps comprising the KDD process can be retrieved from [3].

We propose a combined strategy of some classical and modern techniques for the selection of time-frequency extracted parameters to accomplish effective discrimination of VF pathologies. This communication is outlined as follows. In Section II, data collection and the scope of our study are presented. In Section III methods are detailed. Results in Section IV will precede some concluding remarks and a proposal for further work.

2. Data collection and feature extraction

Data from 29 patient recordings were analyzed, each containing an average of 30 minutes of continuous ECG, of which 100 minutes contained VF. Data were processed to obtain 25 time-frequency parameters from the Pseudo Wigner-Ville (PWV) distribution calculated over 128 point

segments at a 125Hz sampling frequency. Some basic statistics of the extracted parameters are shown in Table 1. A full detailed description of parameters can be found in [1, 2].

Table 1. Some basic statistics of the calculated parameters.

Variable	Normal	Other	VT	VF-Flutter
MINIMFREC	0.73 ± 0.49	0.63 ± 0.38	0.64 ± 0.35	0.64 ± 0.34
MAXIMFREC	21.9 ± 7.7	20.1 ± 7.3	15.4 ± 7.8	14.1 ± 5.6
PMXFRQ	5.51 ± 3.16	4.01 ± 2.47	2.80 ± 2.00	2.56 ± 1.24
AREA (50%)	133 ± 107	126 ± 96	186 ± 136	173 ± 110
LFREC (50%)	9.47 ± 5.01	7.81 ± 3.68	5.49 ± 3.89	4.66 ± 2.17
LTMP (50%)	13.83 ± 11.65	15.43 ± 12.65	33.81± 20.73	34.97 ± 21.80
MINFREC (50%)	3.28 ± 3.35	3.07 ± 3.85	2.44 ± 1.75	2.36 ± 1.55
MAXFREC (50%)	12.76 ± 6.04	10.88 ± 4.65	7.94 ± 4.32	7.03 ± 2.13
NAREAS (50%)	1.46 ± 0.81	1.55 ± 1.59	1.76 ± 0.93	1.64 ± 0.86
DIFFTMP (50%)	5.72 ± 6.04	5.96 ± 6.57	8.80 ± 7.88	9.89 ± 7.93
TMY (50%)	158.6 ± 72.2	158.3 ± 62.6	292 ± 124	251 ± 113
TE	6.5E+08± 1.0E+09	1.9E+09±5.1E+10	1.2E+10±2.0E+11	9.9E+08±1.7E+09
TEH	5.8E+07± 1.0E+08	3.7E+08±1.7E+10	2.9E+09±7.2E+10	1.9E+07±1.0E+08
TEL	5.0E+08±7.2E+08	1.3E+09±2.9E+10	7.4E+09±1.0E+11	8.7E+08±1.4E+09
QTEL	75.4 ± 10.6	76.7 ± 11.6	84.8 ± 9.5	86.0 ± 10.1
QTEH	12.1 ± 9.9	7.9 ± 7.2	4.0 ± 6.0	2.8 ± 4.4
CT8	3.6 ± 1.6	3.8 ± 1.5	6.3 ± 1.3	6.0 ± 1.3
TSNZ	1048 ± 635	1104 ± 638	1596 ± 477	1496 ± 492
TSNZL	686 ± 338	723 ± 332	1250 ± 345	1197 ± 346
TSNZH	172 ± 232	161 ± 226	125 ± 177	104 ± 182
QTL	68.53 ± 9.70	68.86 ± 10.17	79.25 ± 10.01	81.21 ± 10.66
QTH	15.02 ± 10.04	12.30 ± 8.69	6.43 ± 7.61	5.35 ± 7.02
MDL8	93.4 ± 44.3	86.7 ± 39.5	69.2 ± 36.8	64.0 ± 25.4
VDL8	99.2 ± 43.8	87.9 ± 38.6	50.3 ± 29.9	46.7 ± 21.2
CURVE	0.112 ± 0.123	0.134 ± 0.117	0.038 ± 0.202	-0.008 ± 0.208

Four classes are labeled with different prior probabilities: 'NORMAL SINUS RHYTHM' (p_1=40.25%), 'VF-FLUTTER' (p_2=10.66%), 'OTHER RHYTHMS' (p_3=40.25%) and 'VT' (p_4=8.84%).

3. Methods

Previous parameter reduction techniques were based on Wilks' lambda, correlation analysis and discriminant analysis. In this work, three approaches were used for feature selection: Principal Component Analysis (PCA), which provides a preliminary insight in data distribution, Self-Organizing Maps (SOM-Ward) to get qualitative information about the structure of data, and Classification and Regression Trees (CART) to determine the parameter importance analyzing the surrogate and main splits.

3.1. PCA

This classical linear method is extensively used in variable selection. Normalization of data to zero mean and unit variance was strictly necessary to obtain significant results. PCA considerably improved the significance analysis of parameters.

3.2. Self-Organizing Maps (SOM-Ward)

Clustering data is a useful technique to identify homogeneous groups of variables or cases. The usual methods are based on calculating geometric distances between patterns such as K-means, hierarchical clustering and discriminant analysis. This processing pursues that similar input patterns self–organize in the output space. This technique, rather than offering a ranking of relevance variables, yields qualitative information about the structure itself of the data. A useful technique for assessing clusters confidence is to inspect the quantization error of a Self-Organizing Map (SOM) [4] with the classical hierarchical cluster algorithm of Ward (SOM-Ward-clustering) [5].

3.3. CART

CART, Classification and Regression Tree, is a binary decision tree algorithm [6], which has two branches at each internal node. Based on a decade of machine learning and statistical research, CART provides stable performance and reliable results [5, 7]. Its proven methodology is characterized by:

- *A pruning strategy.* Training the trees does not follow any stopping rule but an over-growing and then pruning back methodology.
- *A Binary-Split Search Approach.* CART's binary decision trees are sparing with data and detect efficiently structure in small data sets.
- *Automatic Self-Validation Procedures.* In the search for patterns in databases it is essential to avoid "overfitting,". CART's embedded test disciplines ensure that the patterns found will hold up when applied to new data.
- *Splitting Criteria* CART includes five single-variable splitting criteria - Gini, Symgini, Twoing, Ordered Twoing and Class Probability for classification trees. The default Gini method typically performs best, but, given specific circumstances, other methods can generate more accurate models. CART's unique "Twoing" procedure, for example, is tuned for classification problems with many classes. To deal more effectively with select data patterns, CART also offers splits on linear combinations of continuous predictor variables.

4. Feature selection

4.1. PCA

The analysis of the first three principal components (67% of explained variance) suggests that variables `TE, TEL, TEH,` and `CURVE` are very poor representations of the data (Fig. 2a). Accordingly, variables `TSNZL, CT8, QTL, LFREC, MAXFREC` and `LTMP` contain relevant information in the first principal component.

However, when working with high inter- and intra-subjects variability, it becomes necessary to assess global PCA results performing individual PCA. Individual PCA is illustrated in Fig. 2b. In this case, we make estimations on individual PCA and extract "mean" information over the population. Therefore, we can conclude that it is necessary to use, as

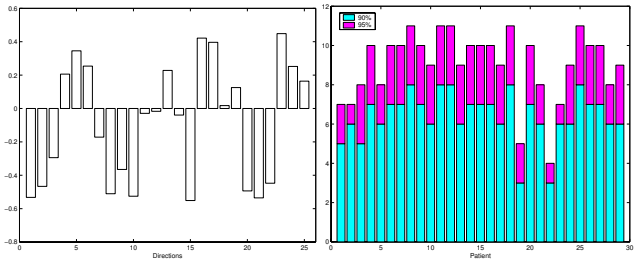

Figure 2. **Left:** Eigenvector components of the highest eigenvalue (34% of explained variance). **Right:** Number of necessary principal components to explain a given % of variance.

a mean, six principal components to explain the 90% of variance and 9 for the 95%. Additionally, we inspected the correlation matrix between eigenvectors as a function of the patient and found that most of them were highly correlated. This procedure helped us to identify three patients with significantly different features from the rest of the population. Consequently, results on these patients are undertaken in a cautious way.

4.2. SOM-Ward Maps

Visualization of SOM-Ward clustering determined relevant features and data distributions. The optimal partition was constituted by five clusters. We visualized the SOM clustering for every feature, as illustrated in Fig. 3. Some conclusions can be drawn:

- Variables TE, TEL and TEH scatter uniformly in all clusters out of VT, which indicates that does not contain discriminant information except for VT separation.
- Variable TSNZ and QTL trace smooth border lines for a cluster mainly formed with VT and VF. This suggests that these variables contain high discrimination power of those pathologies.
- Variable TEL reveals excellent prognosis capabilities since itself determines cluster ♯5, which contains the highest number of VT samples. This issue was not captured by any previous statistical method.
- Variables AREA and MAXFREC define specific clusters which are highly correlated with tachycardia episodes (figure not shown).

We additionally analyzed the quantization error to assess the identified clusters. The quantization error is a measure of how good the data vectors from the source data set are matched by a specific node. It is computed by the average of the squared distance of all data records associated with a node. Averaging over the quantization errors of all nodes yields the quantization error of the map. The map is well adapted if the quantization errors are very small and equally distributed over the map. Table 2 shows results on this test.

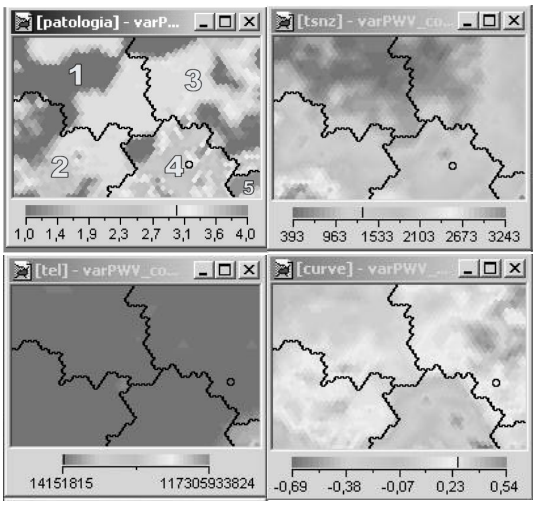

Figure 3. SOM-Ward clustering for **top:** pathology (left) and TSNZ (right); **bottom:** TEL (left) and CURVE (right).

We can observe that clusters ♯1-♯4 contain similar and lower variability w.r.t. cluster ♯5. This may be due to the presence of badly-fitted VT samples which are mainly associated with ♯5. However, despite of being a weak border in terms of SOM theory, it helped us to identify features associated with VT patterns.

Table 2. Analysis of the quantization error for each cluster.

Cluster	Mean	STD DEVIATION	Max
♯1	0.073	0.051	0.338
♯2	0.050	0.044	0.268
♯3	0.138	0.067	0.437
♯4	0.106	0.051	0.307
♯5	0.374	1.139	5.607

4.3. CART analysis

Analysis of surrogate and main splits in CART yields valuable information on the relevance of input features. Table 3 shows ranking of variables according to these measurements and the classical Wilks' lambda statistical analysis. Similar results are obtained in the parameter relevance list. Confidence on this analysis can be ensured since classification rates of the best CART achieved recognition rates higher than 83% in all classes, suggesting that the underlying differences between classes has been captured.

Table 3. Analysis of surrogate and main splits in CART and statistical analysis through Wilks' lambda†.

Variable	Main Splits Score	Variable	Surrogate Splits Score	Variable	Wilks' lambda
TSNZL*	100.00	TSNZL*	100.00	TSNZL*	.671
LTMP*	89.29	LFREC*	63.52	CT8*	.710
CT8	79.93	TE*	43.75	TMY*	.724
MAXFREC	79.80	TEL	32.03	QTL*	.747
TEL	78.18	CURVE*	31.03	VDL8*	.805
TE*	77.69	QTEL	30.21	QTEL*	.807
LFREC*	72.78	MAXIMFRE*	17.47	LTMP*	.807
TMY*	62.33	MAXFREC	17.34	LFREC*	.819
TSNZ	61.63	TEH	16.45	QTEH*	.822
MAXIMFRE*	54.83	QTL*	13.37	MAXFREC*	.834
QTEH	54.25	TMY*	6.53	QTH	.842
QTEL	49.38	PMXFREC*	4.58	CURVE*	.849
QTL*	46.43	MDL8	4.38	PMXFREC	.863
PMXFREC*	39.40	QTEH	4.10	MAXIMFRE	.875
QTH*	36.15	AREA	2.96	TSNZ	.911
TEH	33.80	TSNZ	2.59	DISPERSI	.924
CURVE*	32.49	LTMP*	2.47	MDL8	.937
TSNZH	23.58	QTH*	1.90	NAREAS	.947
MDL8	21.44	VDL8	1.44	MINFREC	.964
MINFREC	21.03	TSNZH	1.37	AREA	.966
NAREAS	18.72	DISPERSI	0.81	TSNZH	.986
VDL8	18.11	MINIMFRE	0.78	MINIMFRE	.988
AREA	7.44	MINFREC	0.71	TEL	.997
MINIMFRE	3.79	NAREAS	0.25	TE*	.998
DISPERSI	1.65	CT8	0.00	TEH	.999

† *: Selected Parameters.

5. Results

5.1. Parameter selection

Finally, a pool of all methods considered 10 parameters to be significant, readily: TSNZL, LTMP, TE, LFREC, TMY, MAXIMFRE, QTL, PMXFREC, QTH, CURVE. The decision on parameters kept is taken based on CART, PCA, SOM, Wilks' lambda and correlation analysis, which leads to eliminate some parameters which seem to have a good behaviour if only one analysis is taken.

5.2. Pathology discrimination

We performed discrimination with all features and with the ones selected in the previous process using a detection tree, previously presented in [1]. Patterns were randomly assigned to two sets; two thirds were used for training (38605 patterns) and the rest for validating (19302 patterns) the discrimination tool. The criterion used to choose the best model was the sum of sensitivity and specificity applied to the validation set in order to obtain well-balanced models. All discrimination models and PCA were developed in MATLAB® environment (Mathworks, Inc). SOM-Ward and CART processing were carried out using shareware implementations from `http://www.eudaptics.com/` and `http://www.salford-systems.com/`, respectively. Statistical analysis was carried out with SPSS.

Detection based on a detection tree using the reduced set of parameters (Sensitivity "VF-Flutter": 88.8%, Specificity "Normal -Other": 94.9%, Spec. "VT": 76.3%) shows better results compared to a detection tree using classical statistical parameter selection (Sens. "VF-Flutter": 85.3%, Spec. "Normal-Other": 95.3%, Spec. "VT": 73.9%). Using this approach, 10 parameters are used, while previous methods used 12 parameters and a more complex detection tree. We can conclude that proposed methods simplify the solution and improve discrimination scores.

6. Conclusions

This study has proposed the application of data mining techniques for the selection of parameters to accomplish effective discrimination of VF pathologies. We showed that exhaustive inspection of the input space with these techniques improved results of the posterior discrimination and reduced the computational cost involved, obtaining improved detection algorithms due to optimal parameter selection. Future work will consider to enhance classification scores by using Support Vector Machines and kernel-based methods in multi-classification schemes.

References

[1] Rosado A, Guerrero J, Serrano AJ, Soria E, Martínez M, Camps G. Ventricular fibrillation detection method using pseudo wigner-ville time-frequency representation. In Fifth Conference of the European Society for Engineering & Medicine. ESEM 99. Barcelona, Spain, 1999; .

[2] Rosado A, Guerrero J, Bataller M, Chorro FJ. A fast non-invasive ventricular fibrillation detection method using pseudo wigner-ville distribution. In Computers in Cardiology., number 28. Rotterdam, The Netherlands, Sep 2001; 249–252.

[3] Bradley PS, Fayyad UM, Mangasarian OL. Mathematical programming for data mining: formulations and challenges. Technical Report MSR-TR-98-04, 1998.

[4] Kohonen T. Self-Organizing Maps. 3rd extended edition. Springer Series in Information Sciences, Vol. 30, 2001.

[5] Michie D, Spiegelhalter DJ, Taylor CC. Machine Learning, Neural and Statistical Classification. D. Michie, D. J. Spiegelhalter, C. C. Taylor, 1994.

[6] Breiman L, Friedman J, Olshen R, Stone C. Classification and Regression Trees. 3rd edition. Chapman & Hall, New York, 1984.

[7] Duda RO, Hart PE, Stork DG. Pattern Classification and Scene Analysis: Part I Pattern Classification. 2nd edition. John Wiley & Sons, 1998.

Address for correspondence:

Alfredo Rosado-Muñoz
Grupo de Procesado Digital de Señales
Universitat de València
C/ Dr. Moliner, 50. 46100 Burjassot (València). Spain
tel./fax: +34 96 3160197/466
http://gpds.uv.es
e-mail: alfredo.rosado@uv.es