Big-Data Analytics for Arrhythmia Classification using Data Compression and Kernel Methods

J.M. Lillo-Castellano[1], I. Mora-Jiménez[1], R. Moreno-González[2], M. Montserrat-García-de-Pablo[2], A. García-Alberola[3] and J.L. Rojo-Álvarez[1]

[1] Department of Signal Theory and Communications, Telematics and Computing, Rey Juan Carlos University, Madrid, Spain
[2] Hospital Solutions, Medtronic Ibérica® S.A., Madrid, Spain.
[3] Arrhythmias Unit, University Hospital Virgen de la Arrixaca, Murcia, Spain

Abstract

Big data analytics is broadly used today in multiple research fields to discover and analyze hidden patterns and other useful information in large databases. Although Cardiac Arrhythmia Classification (CAC) has been studied in depth to date, new CAC methods need to be still designed. In this work, we propose a new big data analytics method for automatic CAC of intracardiac Electrograms (EGMs) stored in Implantable Cardioverter Defibrillators (ICDs). The proposed method combines the effectiveness of a measure based on data compression concepts (Jaccard dictionary similarity), which exploits the information among EGMs, and the classification power of kernel methods. It also requires minimal EGM preprocessing and allows us to deal with EGMs of different duration. A database of 6848 EGMs extracted from a national scientific big data service for ICDs, named SCOOP platform, were used in our experiments. Performance for two classifiers (k-Nearest Neighbors or k-NN, and Support Vector Machines or SVM) were compared in two CAC scenarios using four different input spaces. Results showed that k-NN worked better than SVM when previous episodes from the same patient were available in the classifier design, and vice-versa. For the best cases, k-NN and SVM yielded accuracies near to 95% and 85%, respectively. These results suggest that the proposed method can be used as a high-quality big data service for CAC, providing a support to cardiologists for improving the knowledge on patient diagnosis.

1. Introduction

Patients at high risk of suffering a severe arrhythmia are increasingly treated with an Implantable Cardioverter Defibrillator (ICD) [1]. An ICD is a device with limited memory and computational resources, able to record intracardiac electrical signals given by Electrograms (EGMs) and detect arrhythmic episodes for performing cardioversion, defibrillation or pacing [1]. Regardless of the extensive development to date in ICD arrhythmia detection algorithms, Cardiac Arrhythmia Classification (CAC) remains an active research in the ICD field [2], specially in the improvement of the most appropriate shock therapy according to the arrhythmia type. In this setting, criteria such as cardiac cycle length, RR intervals, width of QRS complex or morphological methods have been broadly used to define new CAC methodologies. However, these approaches present limitations and often require an intensive EGM preprocessing, not always validated in large databases [2].

Currently, a great part of the cardiac arrhythmia research is oriented towards big data analytics, and scientific big data repository systems as SCOOP platform (developed by Medtronic Ibérica® S.A. in 2011) are allowing the development of new CAC algorithms. In this new context, we propose a new big data analytics method for CAC. This method is based on an effective combination of information theory concepts (similarity based on data compression) and kernel methods. It is computationally very fast, requires minimal EGM preprocessing and allows us to deal with episodes of different duration.

The remaining of the paper is organized as follows. Section 2 presents the basis of similarity based on data compression and classification using kernel methods. SCOOP platform and the set of arrhythmia episode EGMs used in this work are shown in Section 3. Results and discussion are presented in Sections 4 and 5, respectively.

2. Data compression and kernel-based classifiers

A measure based on data compression approximates the *Kolmogorov Complexity* (KC) to exploit the amount of information shared by two elements [3]. The KC is defined for a bit string $\boldsymbol{x}$, and it corresponds with the ulti-

mate compressed version of $\boldsymbol{x}$ from which $\boldsymbol{x}$ can be exactly recovered by a decompression program. Measures as *Information Distance* and *Normalized Information Distance* [3] were defined to express a similarity between two bit strings $\boldsymbol{x}$ and $\boldsymbol{y}$ using the KC. Unfortunately, the KC can not be computed in practice [3], therefore data compressors (as LZW, zip, Gzip, LZMA, or PPMZ) are used to approximate it to a quantitative magnitude. Thus, new Compression-based Similarity Measures (CSMs) to be used in real computers have been defined from measures based on KC. The most common CSM in the literature is the Normalized Compression Distance (NCD) [3].

The well-known concept of Dictionary Matching (DM) is increasingly used to approximate KC and define new CSMs computationaly faster than NCD. The two most common CSMs based on DM are the Normalized Dictionary Distance (NDD) and the Fast Compression Distance (FCD) [4]. Applying concepts similar to those used in NDD and FCD, the notion of *Jaccard similarity* of sets [5] can be extended to the dictionary space. Thus, we define the Jaccard Dictionary Similarity (JDS) between two bit strings $\boldsymbol{x}$ and $\boldsymbol{y}$:

$$JDS_{\boldsymbol{xy}} = \frac{|W_{\boldsymbol{x}} \cap W_{\boldsymbol{y}}|}{|W_{\boldsymbol{x}} \cup W_{\boldsymbol{y}}|} \quad (1)$$

where $|\cdot|$, $\cup$ and $\cap$ denote the set cardinality, set union and set intersection operators, and $W_{\boldsymbol{p}}$ is the set of words from the dictionary obtained in the compression of bit string $\boldsymbol{p}$.

In this work, the proposed CAC approach corresponds to a kernel method where: (1) JDS is applied to arrhythmia episodes to construct a *kernel matrix* among episodes (we have defined this kernel as Jaccard Dictionary Kernel or JDK); and (2) each row/column of JDK is used by a machine learning classifier as input vector. An important point when a similarity matrix is considered as a kernel is that it fulfills the Mercers's condition, i.e., the matrix is positive semi-definite [6]. We experimentally checked that JDK is a Mercer's kernel because the constructed JDS matrices always yielded positive eigenvalues.

By means of JDK, two different machine learning schemes were used here to classify the arrhythmia episodes EGMs: k-Nearest Neighbors (k-NN) and Support Vector Machines (SVMs). First, k-NN classifies each instance (in this work instance refers to episode EGM) according to the most frequent class among those belonging to the most similar k instances [7]. Parameter k must be adequately chosen so that the classifier can offer good generalization properties. We have considered JDK as similarity measure. Second, the SVM classifier creates a decision hyperplane maximizing the *margin*, defined as the distance between the boundary and its closest instances, to subsequently classify each instance into a class [6]. A selection of the regularization parameter C (which controls the trade-off between error and *margin* size) was required.

Table 1. Classes of cardiac arrhythmia episodes determined by the expert cardiologist committee and their relative occurrences in our SCOOP database.

Label		# Episodes		Ocurrence	
8-class	**3-class**	**8-class**	**3-class**	**8-class**	**3-class**
ST	Atrial	956	2341	13.96%	34.19%
AF		803		11.73%	
SVT		307		4.48%	
UST		275		4.02%	
SMVT	Ventricular	4113	4387	60.06%	64.06%
VF		274		4.00%	
TWO	Other	82	120	1.20%	1.75%
NS		38		0.55%	

3. Arrhythmia database

The EGM database used in this work has been provided by Medtronic Ibérica® S.A. In 2011, this company developed a Spanish-level scientific big data platform named SCOOP. This platform supports cooperatively the knowledge generation in the ICD field by taking advantage of the automatic information transmission from the ICD to a remote server. To date, SCOOP has stored more than 20,000 two-channel-EGMs of arrhythmia episodes from more than 2,500 patients from 50 Spanish hospitals, with an average follow-up around 2.5 years. Each patient's ICD remote follow-up is within an observational framework research study, so-called UMBRELLA, which ensures the legal, normative, and scientific data exploitation, as well as privacy requirements [8]. Besides the large amount of ICD information stored in SCOOP, a systematic clinical evaluation and classification process is also carried out on each arrhythmia episode by a scientific committee consisting of 6 expert cardiologist, ensuring high quality of data. This committee defined 8 arrhythmia episode classes, namely: *Sinus Tachycardia* (ST), *Atrial Fibrillation* (AF), *Supraventricular Tachycardia or Flutter* (SVT), *Uncertain Supraventricular Tachycardia* (UST), *Sustained Monomorphic Ventricular Tachycardia* (SMVT), *Sustained Polymorphic or Ventricular Fibrillation* (VF), *T-wave Oversensing* (TWO), and *Noise* (NS).

For this work, a set of 6848 EGMs from 629 patients recorded from January 2012 to December 2013 were extracted from SCOOP. Table 1 shows the relative occurrence for each arrhythmia class in our database, as well as the grouping of these classes according to the arrhythmia origin into three major sets (3-class), namely, atrial, ventricular and other. Arrhythmia episodes had a mean length of 24.42 $\pm$ 16.64 s, median 19.82 s, and interquartile range 13.23 s. The mean number of episodes per patient was 10.8 $\pm$ 22.9, median 4, and interquartile range 10. The diversity of ICD models monitored in SCOOP resulted in up to 10 possible lead configurations for far-field and near-field channels. The most usual configurations were *Can to HVB - Vtip to Vring* (3334 episodes) and *Atip To Aring - Vtip to Vring* (2899) (more detail of lead configu-

rations in [9]). All episodes were sampled at 128 samples/second in the ± 8 mV range, with an amplitude resolution of 0.063 mV.

Each episode consisted of two-channel-EGMs, denoted as $\boldsymbol{ch}_1$ and $\boldsymbol{ch}_2$, with the same number of values simultaneously recorded from different lead configurations. We considered four input spaces for the CAC design. Two spaces consider values from both channels as they are, and the other two perform a transformation of each episode into a complex signal $\boldsymbol{c} = \boldsymbol{ch}_1 + j \cdot \boldsymbol{ch}_2$ (where j is the imaginary unit) to work with magnitude and phase values (inspired on phase portraits [10]). Thus, we considered: (1) Concatenated Channels (CCcat); (2) Alternate values of each channel (ACcat); (3) Concatenated (CXcat) and (4) Alternate (AXcat) magnitude and phase values.

Two resampling strategies based on the *Leave One Out - Cross Validation* (LOOCV) [7] technique were used to evaluate the CAC generalization performance. First, the strategy named LOEOCV considers that each instance is associated to an episode. The strategy named LOPOCV considers the equivalence instance-patient such that each instance encompasses all episodes from the same patient.

4. Results

Two merit figures (accuracy rate, Cohen's kappa coefficient κ [11]) and the *baseline accuracy* (which express the occurrence of the majority class) were considered for performance evaluation. Table 2 summarizes the results for SVM and k-NN. We experimentally checked that parameter C does not affect the SVM performance in our dataset, so it was fixed at a value of 100. The k value providing the highest κ according to each validation strategy (LOEOCV and LOPOCV) was selected for k-NN. Thus, $k = 1$ was selected in all cases for LOEOCV, and around $k = 3$ and $k = 5$ for 3-class and 8-class, respectively, in LOPOCV. The best performance (in bold) was always for CCcat input space, except for LOEOCV-8-class scenario where the accuracy was slightly better for AXcat, although κ was the same. The worst performance was for the LOPOCV-8-class case, which was more pronounced for k-NN classifier with accuracies lower than (but close to) baseline accuracy (given by 60.06%). Related with this performance is the dramatically different accuracy between LOEOCV and LOPOCV, which was higher with k-NN classifier. This performance reduction in LOPOCV was not so dramatic for the 3-class with SVM, which outperformed the baseline accuracy (64.06%) in 21%.

Kappa coefficient allows to interpret the reliability of the CAC schemes and to know quantitatively how random (κ=0) is the classification. We conclude that the accuracy when using LOEOCV is not random because κ >0.6 in all scenarios, yielding high values with k-NN. From a clinical point of view, this result is in accordance with a patient having similar individual physiopathological mechanisms characterizing his/her EGM episodes (specially when the patient suffers from an arrhythmic storm). Thus, 1-NN is the best classifier when episodes from the same patient were considered in the classifier design. On the opposite side are κ values obtained when considering the LOPOCV strategy. With k-NN classifier, κ is very low and accuracies close to the baseline (60.05%) were reached, suggesting k-NN is not the best scheme for this purpose. Note that accuracy and kappa increase significantly when SVM is used, reaching κ=0.47 and κ=0.7 in the best cases for 8 and 3-class scenarios, respectively.

Table 2. Accuracy rate (first value, in %) and κ coefficient (second value) in 32 different CAC scenarios.

		8-class		3-class	
		LOEOCV	LOPOCV	LOEOCV	LOPOCV
k-NN	CCcat	89.78 0.83	60.31 0.27	**95.43** **0.90**	77.56 0.49
	CXcat	89.72 0.81	58.78 0.23	95.33 0.90	73.10 0.37
	ACcat	88.39 0.83	49.72 0.16	94.17 0.88	68.84 0.29
	AXcat	**89.97** **0.83**	49.04 0.17	95.24 0.90	65.76 0.28
SVM	CCcat	84.01 0.72	**69.41** **0.47**	92.41 0.84	**85.81** **0.70**
	CXcat	83.17 0.71	65.51 0.39	91.40 0.82	80.91 0.60
	ACcat	82.49 0.69	62.94 0.35	90.73 0.80	80.53 0.59
	AXcat	84.06 0.73	61.14 0.33	91.95 0.83	78.72 0.55

The SVM implementation used in this work was the multi-class one [12], which also provides with an estimation of the corresponding posterior probability (PP). Figure 1 shows the PP histograms for correctly and wrongly classified arrhythmia episodes in the 3-class-CCcat-LOPOCV scenario (best case when episodes from the same patient are not available in the classifier design). Note that PP is high (in most cases near 100%) with pronounced skewness towards high values. However, PP histograms for the wrongly classified episodes follow almost uniform distributions, evidencing a lack of security in the SVM classification. Previous work in [13] highlighted the need to improve accuracy when classifying episodes of a patient not considered in the classifier design (i.e, LOPOCV scenario), which has been improved here about 7 percentage points.

5. Discussion

In this work, we have presented a new big data analytics method for automatic CAC of EGMs stored by ICDs. This new approach effectively combines concepts based on data compression with the power of kernel methods, allowing to classify EGMs with different duration while avoiding large preprocessing stages.

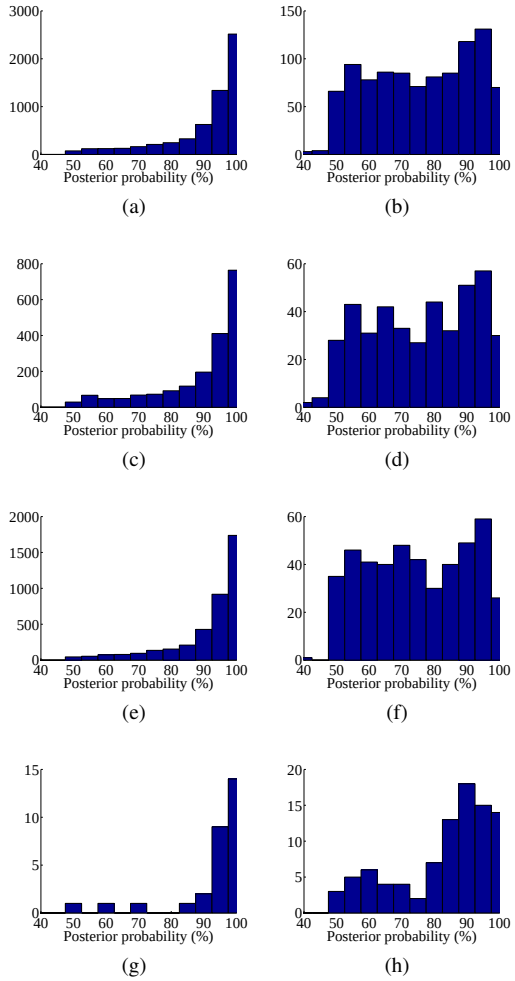

Figure 1. PP histograms provided by JDK-SVM for the 3-class-CCcat-LOPOCV scenario. Correctly-wrongly classified cardiac arrhythmia episodes taking into account all (a)-(b), atrial (c)-(d), ventricular (e)-(f), and other (g)-(h) classes.

The classification potential of this new method facilitates its use in multiple cardiac applications, whose main reference would be as a support tool in SCOOP platform. The actual scientific committee could receive support from the method in their labeling task, providing even a priority search of relevant episodes pending of cardiologist evaluation.

The use of elaborated machine learning techniques such as SVM has shown to improve the CAC performance in LOPOCV scenarios. Future work is oriented towards the use of new arrhythmia information provided by the ICD, and to new cardiac signal scenarios with different conditions on sampling rate and amplitude resolution.

Acknowledgment

Thanks to the team at Medtronic Ibérica® S.A. and the researchers/contributors in SCOOP platform for their professional support throughout all the project. Thanks to Dr. Ricardo Santiago-Mozos for his fruitful ideas and discussions. This work has been partly supported by TEC2010-19263 and TEC2013-48439-C4-1-R projects from Spanish Government. JMLC is supported by the Spanish FPU grant FPU13/03134.

References

[1] Stroobandt R, Barold S, Sinnaeve A. Implantable Cardioverter Defibrillators Step by Step: An Illustrated Guide. 1st edition. Wiley-Blackwell, 2009.

[2] Aliot E, Nitzschéb R, Ripartb A. Arrhythmia detection by dual-chamber implantable cardioverter defibrillators. a review of current algorithms. Europace 2004;6(4):273–286.

[3] Li M, Chen X, Li X, Ma B, Vitányi P. The similarity metric. IEEE Trans Inf Theory 2004;50(12):3250–3264.

[4] Cerra D, Datcu M. A fast compression-based similarity measure with applications to content-based image retrieval. J Vis Commun Image R 2012;23(2):293–302.

[5] Leskovec J, Rajaraman A, Ullman J. Mining of Massive Datasets. 2nd edition. Cambridge University Press, 2014.

[6] Vapnik V. The Nature of Statistical Learning Theory. 2nd edition. Springer Verlag, 1999.

[7] Duda R, Hart P, Stork D. Pattern Classification. 2nd edition. Wiley-Interscience, 2001.

[8] Fontenla A, López-Gil M, et al. Clinical profile and incidence of ventricular arrhythmia in patients undergoing defibrillator generator replacement in spain. Revista Espanola de Cardiologia English Edition 2014;67(12):986–992.

[9] Rojo-Álvarez J, Arenal-Maíz A, Artés-Rodríguez A. Discriminating between supraventricular and ventricular tachycardias from egm onset analysis. IEEE Eng Med Biol Mag 2002;21(1):16–26.

[10] Wegert E. Visual Complex Functions: An Introduction with Phase Portraits. 1st edition. Springer Science, 2012.

[11] Aggarwal C. Data Classification: Algorithms and Applications. First edition. Chapman & Hall, 2014.

[12] Fan R, Chen P, Lin C. Working set selection using second order information for training svm. J Mach Learn Res 2005; 6:1889–1918.

[13] Lillo-Castellano J, Mora-Jiménez I, Santiago-Mozos R, Chavarría-Asso F, A. CG, García-Alberola A, Rojo-Álvarez J. Symmetrical compression distance for arrhythmia discrimination in cloud-based big data services. IEEE J Biomed Health Inform 2015;PP.

Address for correspondence:

J.M. Lillo-Castellano

Departmental III, Rey Juan Carlos University, Camino del Molino s/n, 28943-Fuenlabrada, Spain.

josemaria.lillo@urjc.es