

ASSESSMENT AND COMPARISON OF TIME REALIGNMENT METHODS FOR SUPERVISED HEART BEAT CLASSIFICATION

G. de Lannoy¹, M. Verleysen¹ and J. Delbeke²

¹*Machine Learning Group, Université catholique de Louvain, pl. du Levant 3, 1348 Louvain-la-Neuve, Belgium*

²*Departement of physiology and pharmacology, Université catholique de Louvain
av. Hippocrate 54, 1200 Bruxelles, Belgium*

Keywords: Heart beat classification, Time realignment, Dynamic time warping, Trace segmentation, Wavelet transform, Nearest neighbor classifier.

Abstract: A reliable diagnosis of cardiac diseases can sometimes only be obtained by observing the heart of a patient for a long time period where every single heart beat is of importance. Computer-aided classification of heart beats is therefore of great help. The classification of the complete heart beat has many advantages compared to a classification of the QRS complex only or feature extraction methods. Nevertheless, the task is challenging because of the time-varying property of the heart beats. In this work, four time-alignment methods are evaluated and compared in the context of supervised heart beat classification. Among the four methods are three time series resampling methods by linear interpolation, cubic splines interpolation and trace segmentation. The fourth method is a realignment algorithm by dynamic time warping. The multiple sources of artifacts are filtered by discrete wavelet transform. As it only relies on a dissimilarity measure, the k -nearest neighbor classifier is a suitable choice for supervised classification of time series like ECG signals in multiple classes. Two different experiments corresponding to inter-patient and intra-patient classification are conducted on representative dataset built from the standard public MIT-BIH arrhythmia database.

1 INTRODUCTION

The importance of the electrocardiogram (ECG) signal for the diagnosis of cardiac diseases is widely known (Clifford et al., 2006). However, a reliable diagnosis can sometimes only be obtained by observing the ECG of a patient for a long-term period, comprising hundreds to thousands of heart beats (Chudacek et al., 2007; Cuesta-Frau et al., 2003; Jekova et al., 2008). Just a very few number of these beats may actually reveal a pathology and the complete set of beats must therefore be taken into account for the diagnosis.

Such long-term ECG signals are often recorded using the very popular Holter recorders. These systems are ambulatory heart activity recordings, with a signal storing time ranging from 24 to 48 hours. They are used in the clinical diagnosis of some disease conditions and by pharmaceutical groups for the evaluation of new drugs during phase-one studies.

Due to the high number of beats to evaluate, analysis is performed off-line by cardiologists, keeping in mind that the diagnosis may rely on just a few tran-

sient patterns. The duration of the process makes this task very expensive. Reliable visual inspection is difficult and computer-aided detection of the pathological beats is of major importance. However, this is a difficult task in real situations.

First, several sources of noise typically pollute the ECG signal. Among these, power line interferences, muscular artifacts, poor electrode contacts and baseline wandering due to respiration can sometimes be identified. These artifacts can largely degrade the quality of the signal and therefore complicate the beat identification. Second, another hurdle is that the heart rhythm can be quite unstable and variable in normal conditions.

The classification of the complete heart beat is a different problem than the classification of the QRS complex only. In the latter case, beats are defined by cutting a fixed-size window around the R spikes. The size of the window is of similar length for all beats, but the duration of QRS complexes varies with time and pathologies. For this reason, a fixed window size may lead to incorrect extraction of the QRS complex. Furthermore, some pathologies can only be

identified by using more information than this QRS complex only. A complete heart beat is defined as the activity between the start of a P wave and the start of the next one. The time-varying property of such data challenges the automatic computer-aided classification methods.

Indeed, standard classification algorithms work on data arrays of identical dimension. A pre-processing step is therefore required to deal with time series of different lengths. This pre-processing step can be the extraction of the same number of typical ECG features from each beat. However, the complete morphology of the beat can hardly be summarized in a set of features, and the discriminative information may be missing in the chosen features. A more sophisticated pre-processing step is to resample and/or to realign the signal in order to have the same number of samples in each beat. The advantage is that the complete heart beat can then be used as input to any classifier without loss of information. Moreover, advanced resampling methods can correct for time shifts and realign observations.

In this paper, four temporal realignment methods are evaluated in the context of complete heart beat classification. These methods have proven successful in various disciplines such as speech processing, spectral data analysis and signal filtering. The remaining of this paper is structured as follows. After this introduction, Section 2 gives a short review of the literature about the state of the art in ECG beat classification. Section 3 sets the theoretical background for the methods used in this work. Section 4 introduces the methodology followed in this work. Finally, Section 5 shows the experiments on a real database and the corresponding results.

2 STATE OF THE ART

A large body of literature about the ECG beat classification has arisen in recent years. Computer-aided algorithms for the automatic classification of the heart beats can be separated in two groups: supervised or unsupervised (Clifford et al., 2006). In the first case, it is necessary to have a set of manually labeled heart beats. In the second case, because manually labeled reference beats are not available, automatic diagnosis is impossible.

Clustering techniques are among the most used unsupervised processing techniques (Clifford et al., 2006). A comparison between clustering algorithms, realignment algorithms and feature extraction methods for unsupervised beat classification has been conducted in (Cuesta-Frau et al., 2003). More recently,

the results of the previous work have been improved by introducing the J-means heuristic for heart beat clustering (Rodriguez-Sotela et al., 2007). Self-organizing maps (SOM) have also been investigated with very promising results (Clifford et al., 2006). The disadvantage of unsupervised methods, beside the fact that an automatic diagnosis cannot be obtained, is the empirical choice of the number of clusters (or of SOM prototypes). In the case of heart beats, this choice is hard to make a priori and it can significantly affect the results.

The vast majority of supervised beat classification algorithms reported in the literature work either on the QRS complex only, either on features extracted from the signal, or a combination of both (Clifford et al., 2006). In the first case, only the QRS complex is extracted from the ECG signal by defining a fixed-size window around the R spike. The beat time-varying duration property is therefore avoided. In the second case, common time-domain features include the time interval between characteristic ECG patterns such as R-R and Q-T intervals, the Hermite basis function expansion of the QRS complex and order statistics of the QRS sequence. Frequency domain descriptors such as Fourier or wavelet transform coefficients are alternative interesting features. Classification methods such as artificial neural networks (ANNs), support vector machines (SVMs) and combined methods have been applied with success for QRS classification (Clifford et al., 2006). A recent comprehensive review of supervised classification methods of the QRS complex can be found in (Jekova et al., 2008).

These methods relying on the classification of the QRS complex or of selected features may miss potentially useful information for discrimination. Some pathologies can indeed be more accurately identified by using more information than the QRS complex alone, such as Q-T intervals. Furthermore, using a fixed-size window around the R spike may lead to wrong QRS extraction because the duration of the QRS complexes is not stationary. Also, the choice of the features is of great importance. Summarizing the complete morphology of the beats into a set of features is a difficult and application dependent task (Clifford et al., 2006; Jekova et al., 2008).

In order to circumvent these issues, realignment methods have recently been applied with success to unsupervised beat clustering (Cuesta-Frau et al., 2003). The complete beat can then be considered by the classification procedure. However, contrarily to clustering, supervised classification algorithms rely more easily on a set of features extracted from the time series rather than on the raw time series themselves. For this reason, despite the advantages of

these methods, very few studies have been reported on supervised full beat classification after realignment. The k -nearest neighbor (KNN) classifier is an exception to this. In a similar way to clustering methods, it only relies on a dissimilarity metric between observations. Such dissimilarity metrics can naturally be computed between time series of similar lengths. In this work, four time series realignment methods are evaluated in the context of complete heart beat classification by k -nearest neighbor classifier.

3 THEORETICAL BACKGROUND

This section provides a brief account of the theoretical background for the signal processing methods used in this work. The time-frequency filtering by discrete wavelet transform is briefly described first. Thereafter, the dynamic time warping realignment algorithm is introduced. Finally, a brief summary to the three resampling algorithms is provided. In the remainder of this work, these methods are globally referred to as the realignment methods.

3.1 The Discrete Wavelet Transform

The continuous wavelet transform (CWT) is a time-frequency decomposition of a signal $x(t)$ by the convolution of this signal with a so-called *wavelet function* $\psi(t)$ (Mallat, 1999). From a wavelet function, one can obtain a family of time-scale waveforms by translation b and scaling a , with $a, b \in \mathbf{R}$:

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \psi\left(\frac{t-b}{a}\right) . \quad (1)$$

When $a = 1$ and $b = 0$, $\psi(t)$ is called the *mother wavelet*. The *wavelet transform* of a function $x(t) \in L^2(\mathbf{R})$ is a projection of this function on the wavelet basis $\{\Psi_{a,b}\}$:

$$T(a,b) = \int_{-\infty}^{+\infty} x(t) \Psi_{a,b}(t) dt . \quad (2)$$

The discrete wavelet transform (DWT) removes the redundancy of the CWT by using dyadic scales and discrete translations. The first DWT step produces two sets of coefficients: approximation coefficients (low frequency components) and detail coefficients (high frequency components), followed by a dyadic decimation (downsampling). This process is iterated $\log_2(N)$ times at most, where N is the size of the signal $x(t)$ (Mallat, 1999).

3.2 Dynamic Time Warping

Dynamic time warping (DTW) is an algorithm that finds an optimal alignment function between two sequences of different lengths (Myers and Rabiner, 1981). It has for example been used with great success in speech recognition and process control. The sequences are warped non-linearly in the time dimension to determine a measure of their similarity independent of certain non-linear variations in the time dimension.

Let \mathbf{x} and \mathbf{y} be two sequences of lengths n_x and n_y respectively. The objective is to find the best alignment between the two sequences, according to a cost function. The alignment procedure allows us to compare a value $x(i)$ of \mathbf{x} with a value $y(j)$ of \mathbf{y} . The whole set of possible comparisons can be represented as a matrix of size $n_x \times n_y$, that can be seen as a multi-stage graph. The objective is then to find a node path $(i_1, j_1), (i_2, j_2), \dots, (i_f, j_f)$ of length f along the graph such that the final cumulative cost D is minimal. The latter is defined as

$$D = \sum_{k=1}^f [d(i_k, j_k) | (i_{k-1}, j_{k-1})] , \quad (3)$$

where $d(i, j)$ is a cost function that allows us to compare a value $x(i)$ of \mathbf{x} with a value $y(j)$ of \mathbf{y} . A typical cost function is for example the squared Euclidean distance:

$$d(i, j) = (x(i) - y(j))^2 . \quad (4)$$

The optimization process is performed using dynamic programming. The cumulative cost of the optimal alignment path can then be used as a dissimilarity measure.

3.3 Trace Segmentation

Trace segmentation (TS) is a non-uniform sampling method used in speech recognition to normalize the duration of utterances. Standard vector-space norms can then be used to compare them. The objective is to retain only the samples of the signal where the main changes take place (Cuesta-Frau et al., 2003).

Given a sequence \mathbf{x} of length n , let us define the partial accumulated derivate Δ_j of \mathbf{x} :

$$\Delta_j = \sum_{i=2}^{j+1} |x(i) - x(i-1)| . \quad (5)$$

The accumulated derivative at the end of the sequence is

$$\Delta = \sum_{i=2}^n |x(i) - x(i-1)| . \quad (6)$$

If h is the desired number of samples after trace segmentation, the average interval amplitude value of Δ is given by

$$L = \frac{\Delta}{h} . \quad (7)$$

Let us now define the output sequence \mathbf{x}_{tr} of length h obtained by trace segmentation of \mathbf{x} . Each sample $x_{tr}(l)$ is taken from the sample $x(j)$ corresponding to the time when the accumulated derivate exceeds an integer multiple of L :

$$x_{tr}(l) = x(j) | j = \operatorname{argmin}_{1 < i < n-1} (l \times L < \Delta_i) , \quad (8)$$

with $x_{tr}(1) = x(1)$ and $x_{tr}(h) = x(n)$. The sequence \mathbf{x}_{tr} thereby includes the values of \mathbf{x} where only the main changes take place.

3.4 Interpolation

Given a sequence \mathbf{x} of length n , obtained by sampling or experiment, regression analysis tries to estimate a function which closely fits those data points. Interpolation is a specific case of curve fitting, in which the function must go exactly through the data points. One of the simplest interpolation methods is linear interpolation. In this case, a linear function is fit at each interval x_k, x_{k+1} .

Spline interpolation uses low-degree polynomials in each of the intervals, and chooses the polynomial pieces such that they fit smoothly together. The resulting function is called a spline (De Boor, 1978). Spline interpolation incurs a smaller error than linear interpolation, and the interpolant is smoother. However, the interpolant is easier to evaluate than the high-degree polynomials used in polynomial interpolation.

In both cases, the estimated function then allows us to resample (or *stretch*) the sequence by generating new data points within the range of the discrete set of known data points.

4 METHODOLOGY

In this work, supervised classification of complete heart beats is considered. Let us assume that a reference database has been obtained and annotated, with all pathologies of interest being represented. Given a new ECG signal, for example recorded using an Holter system, one wants to use the information contained in the reference database in order to predict the pathologies present in the new signal.

First, all the beats within the signal must be separated. Several computer-aided annotation algorithms have been reported in the literature in order to automatically detect the ECG characteristic points for beat extraction (Clifford et al., 2006).

Thereafter, a realignment of the beats must be achieved. Four sequence alignment methods are evaluated in this work. The first two are interpolation methods: linear interpolation (LI) and cubic splines interpolation (CSI). The third method is the trace segmentation method (TS) and the fourth is the dynamic time warping (DTW) realignment algorithm.

Next, noise and artifacts are filtered by time-frequency filtering using the discrete wavelet transform (DWT) of the heart beats. Moreover, the dimensionality of the observations is strongly reduced by the downsampling of the approximation coefficients at each step of the DWT. This reduces the computational cost of the classification algorithm.

Finally, the realigned and filtered heart beats are reduced to zero mean and unit variance. The beats are then given as input to a classifier. A natural choice for supervised classification of time series in multiple classes is the k -nearest neighbor (KNN) classifier. The k -nearest neighbor (KNN) algorithm is a supervised classifier where an observation (corresponding to one heart beat) is assigned to the class most common amongst its k nearest neighbors in the reference set. The algorithm is supervised, because the neighbors are taken from a reference set of observations for which the correct classification is known. This can be thought of as the training set for the algorithm, though no explicit training step is required.

In order to identify the closest neighbors, a dissimilarity metric must be defined. When using the three first realignment methods the KNN distance measure is the Euclidean distance between observations, computed on the aligned beats. In the case of the DTW warping algorithm, the DTW dissimilarity measure can directly be used in the KNN method, without effectively computing the realigned time series.

5 EXPERIMENTS AND RESULTS

The performances of the four realignment methods are evaluated on the public standard MIT-BIH arrhythmia database (Goldberger et al., 2000). It contains 48 half-hour recordings of annotated ECG with a sampling rate of 360Hz and 11-bit resolution over a 10-mV range. Except recordings 201 and 202, each recording comes from a different patient, so the database contains a total of 47 subjects. The original annotations are used. The method defined in (Rodríguez-Sotela et al., 2007) for extracting complete beats from the R spike annotations is used. The five main types of heart beats represented in the database are used in this study: (1) normal beats (N) - 74820 cases, (2) left bundle branch block beats (L) -

8050 cases, (3) right bundle branch block beats (R) - 7220 cases, (4) premature ventricular contractions (V) - 6970 cases, and (5) paced beats (P) - 7000 cases.

From each of the 47 recordings, three beats of each type available in the recording are randomly selected. Each patient is thus fairly represented in the dataset. Nevertheless, only a small subset of recordings contains L, R and P types, while almost every recording contains the N class. Therefore, in order to obtain the same number of representatives for each class, other beats are then randomly selected amongst the recordings and added to the experimental set until equally balanced classes are obtained. The final experimental dataset contains a total number of 600 heart beats, including an equal number of 120 beats per class where the 47 patients are fairly represented.

The level 3 approximation coefficients of the DWT with a biorthogonal mother wavelet are used as features. The classification performances are evaluated using a KNN classifier with $k = 3$ neighbors. Two different experiments are conducted. The first one is an *intra-patient* classification where the complete training set is used for the classification of each heart beat. A typical leave-one-out methodology is used for the evaluation of the performances. The second experiment is an *inter-patient* classification. A leave-one-out method is also used. However, when classifying a heart beat of a given patient, all beats coming from this patient are removed from the training set. The classification is therefore based only on the annotated beats of other patients, which is a much harder generalization task. Table 1 shows the intra-patient classification results and Table 2 holds the inter-patient classification results. The values are average rates computed over ten different random selections during the construction of the experimental dataset.

When considering the intra-patient experiment, very similar results are obtained by the two interpolation methods with an average of 90% correct classifications. Surprisingly, the dynamic time warping algorithm provides slightly worst results with an average of 88%. Amongst the four methods, trace segmentation obtains the worst results with an average of only 70%. The computational time of the two interpolation methods and the trace segmentation method are very similar, allowing real-time analysis. On the other hand, the running time of the DTW algorithm makes this method only suitable for off-line analysis.

The results obtained with the inter-patient experiment in Table 2 are unsatisfactory, especially for the L, R and N types. However, the low results for these classes are preliminary, since the MIT-BIH database only contains four patients with left bundle branch

block and five patients with right bundle branch block, which is quite insufficient for generalization. These results regarding generalization between patients of the MIT database are confirmed in other recent works (Jekova et al., 2008). Furthermore, the three L, R and N classes are often considered to be of the same clinical relevance. Indeed, the Association for the Advancement of Medical Instrumentation (AAMI) recommends to group these three classes together (Maier et al., 1999). Table 3 holds the results when the three L, R and N types are merged together. The results are comparable to intra-patient performances, and show that the generalization between patients is possible.

Although numerous studies dealing with heart beat classification have been reported in the literature, a comparison is difficult to make. Indeed, few studies address multi-class problems including more than the N and V heart beat classes. Also, only a subset including 5 to 10 recordings of the MIT database is usually used in these studies. Moreover, the training dataset is constructed in quite different ways, and portions of the recordings including noise are sometimes rejected during the extraction of beats. For example, (Ham and Han, 1996) obtained very different results with 44 recordings in comparison with (Moraes et al., 2002) using only 6 recordings to discriminate between V and N beat types.

In this work, the experimental dataset is created in such way that the number of heart beats is equally balanced between each class and that all 47 patients of the database are fairly represented. The noisy portions of the recordings are also included in the data. Furthermore, the inter-patient and intra-patient heart beat classifications are two very different objectives and are therefore separated in this study.

Table 1: Intra-patient correct classification rate.

	N	L	R	V	P
LI	91.1	84.5	89.7	92.4	95.8
CSI	91.3	84.4	89.8	92.4	95.8
DTW	90.3	82.1	85.8	91.2	95.0
TS	60.1	64.0	75.0	79.7	73.7

Table 2: Inter-patient correct classification rate.

	N	L	R	V	P
LI	32.0	15.9	38.2	71.1	85.0
CSI	32.1	15.9	38.4	71.1	85.2
DTW	43.3	4.6	46.2	51.2	75.1
TS	27.9	11.7	8.5	53.7	56.8

Table 3: Inter-patient correct classification rate, when merging the L R and N types.

	N+L+R	V	P
LI	81.1	80.7	92.6
CSI	81.3	80.7	92.7
DTW	74.4	79.2	90.0
TS	60.0	72.6	72.6

6 CONCLUSIONS

In this work, four time-alignment methods are evaluated and compared in the context of supervised heart beat classification. Amongst these methods are three time series resampling algorithms by linear interpolation, cubic splines interpolation and trace segmentation. The fourth method is a realignment algorithm by dynamic time warping. The multiple sources of noise and artifacts are filtered by means of a time-frequency decomposition by discrete wavelet transform. The downsampling induced by DWT approximation coefficients also reduces the dimension of the observations. Since it only relies on a dissimilarity measure between observations, the KNN classifier is a natural choice for supervised classification of time series in multiple classes.

Experiments are conducted on a representative dataset built from the standard public MIT-BIH arrhythmia database. The five main types of heart beats are considered. The experimental dataset is created so as to include an equal number of beats per class with all patients being fairly represented. For intra-patient classification, very similar results are obtained by the two interpolation methods with an average correct classification rate of 90%. Surprisingly, the dynamic time warping algorithm provides slightly lower results with an average of 88%. Amongst the four methods, the trace segmentation obtains the worst results with an average of only 70%. On the other hand, the results obtained for inter-patient classification are unsatisfactory, especially for the L and R types. However, the low yield obtained for these classes are preliminary, since the number of patients with these pathologies in the MIT-BIH database is insufficient for generalization. When merging these three classes together as recommended by the AAMI, the results achieved are then comparable to intra-patient classification.

Very few other studies work on a reliable dataset with multi-class, inter and intra patient classification and comparisons to other works are difficult to obtain. Further works will include a comparison with the results of other QRS classification and feature extraction methods using the same dataset.

ACKNOWLEDGEMENTS

G. de Lannoy is funded by a Belgian F.R.I.A. grant. This work was partly supported by the Belgian "Région Wallonne" ADVENS 4994 project.

REFERENCES

- Chudacek, V., Petrik, M., Georgoulas, G., Cepek, M., Lhotska, L., and Stylios, C. (2007). Comparison of seven approaches for holter ecg clustering and classification. *Engineering in Medicine and Biology Society, 2007. EMBS 2007. 29th Annual International Conference of the IEEE*, pages 3844–3847.
- Clifford, G. D., Azuaje, F., and McSharry, P. (2006). *Advanced Methods And Tools for ECG Data Analysis*. Artech House, Inc., Norwood, MA, USA.
- Cuesta-Frau, D., Pres-Corts, J. C., and Andreu-Garcia, G. (2003). Clustering of electrocardiographic signals in computer-aided holter analysis. *Computer Methods and Programs in Biomedicine*, 72:179–196.
- De Boor, C. (1978). *A practical guide to splines*. Springer-Verlag, New York.
- Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., Mietus, J., Moody, G., Peng, C.-K., and Stanley, H. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220.
- Ham, F. and Han, S. (1996). Classification of cardiac arrhythmias using fuzzy artmap. *IEEE Transactions on Biomedical Engineering*, 43(4):425–429.
- Jekova, I., Botrolan, G., and Christov, I. (2008). Assessment and comparison of different methods for heartbeats classification. *Medical Engineering and Physics*, 30:248–257.
- Maier, C., Dickhaus, H., and Gittinger, J. (1999). Unsupervised morphological classification of qrs complexes. *Computers in Cardiology 1999*, pages 683–686.
- Mallat, S. (1999). *A Wavelet Tour of Signal Processing, Second Edition (Wavelet Analysis and Its Applications)*. IEEE press, San Diego. ISBN 978-0124666061.
- Moraes, J., Seixas, M., Vilani, F., and Costa, E. (2002). A real time qrs complex classification method using mahalabis distance. *Computers in Cardiology, 2002*, pages 201–204.
- Myers, C. S. and Rabiner, L. (1981). A comparative study of several dynamic time-warping algorithms for connected word recognition. *The Bell System Technical Journal*, 60:1389–1409.
- Rodriguez-Sotela, J. L., Cuesta-Frau, D., and Castellanos-Dominguez, G. (2007). An improved method for unsupervised analysis of ecg beats based on wt features and j-means clustering. *Computers in Cardiology*, 34:581–584.