Time Series Classification using Motifs and Characteristics Extraction: A Case Study on ECG Databases

André G. Maletzke[1] Huei D. Lee[1] Gustavo E.A.P.A. Batista[2] Solange O. Rezende[2] Renato B. Machado[1,6]

Richardson F. Voltolini[3] Joylan N. Maciel[4] Fabiano Silva[5] Leandro B. dos Santos[1] Feng. C. Wu[1,6]

[1]State University of West Parana (UNIOESTE), Foz do Iguaçu (PR), Brazil
[2]University of São Paulo (USP), São Carlos (SP), Brazil
[3]Center for Higher Education of Foz do Iguassu (CESUFOZ), Foz do Iguaçu (PR), Brazil
[4]Federal University of Latin-American Integration (UNILA), Foz do Iguaçu (PR), Brazil
[5]Federal University of Parana (UFPR), Curitiba (PR), Brazil
[6]State University of Campinas (UNICAMP), Campinas (SP), Brazil
andregustavom@gmail.com, gbatista@icmc.usp.br, rfvoltolini@gmail.com, joylan@gmail.com, fabiano@inf.ufpr.br, wufengchung@gmail.com

Abstract

In the last decade, the interest for temporal data analysis methods has increased significantly in many application areas. One of these areas is the medical field, in which temporal data is in the core of innumerous diagnosis exams. However, only a small portion of all gathered medical data is properly analyzed, in part, due to the lack of appropriate temporal methods and tools. This work presents an alternative approach, based on global characteristics and motifs, to mine medical time series databases using machine learning algorithms. Characteristics are data statistics that present a global summary of the data. Motifs are frequently recurrent subsequences that usually represent interesting local patterns. We use a combination of global characteristics and local motifs to describe the data and feed machine learning algorithms. A case study is performed on three databases of Electrocardiogram exams. Our results show the superior performance of our approach in comparison to the naïve method that provides raw temporal data directly to the learning algorithms. We demonstrate that our approach is more accurate and provides more interpretable models than the method that does not extract features.

Keywords: morphological pattern, attribute extraction, decision trees.

1. Introduction

Time series data consist of an ordered set of observations about a determined phenomenon, measured along a time period. Knowledge extraction from this kind of data has attracted the attention of researchers and experts in several application areas, which include stock market; amino acid sequences data; monitoring of chemical, physical and biological variables that describe a patient clinical state, among many others.

In Medicine, characteristics related to the clinical state of a patient can be monitored by equipments, which capture information on physical, chemical and biological variables. One example is the Electrocardiogram exam (ECG), which consists in monitoring the changes in electrical potential generated by heart activity over time and is essential for the diagnosis of many heart diseases and other disorders [1].

Under the World Health Organization, cardiovascular diseases are the leading cause of death worldwide, accounting for about 29% of the cases. According to the Ministry of Health, cardiovascular diseases are also a major cause of death in Brazil, totalizing 32% of deaths, reaching younger population more pronouncedly when compared to other countries such as the United States, Japan and Western European countries. Only in 2005, 22% of patient care expenditures (except deliveries) were related to cardiovascular diseases followed by chronic respiratory diseases (15%) and neoplasms (11%) [2]. These statistics highlight the need for more efficient public policies to promote the prevention and diagnosis of this disease.

The temporal characteristic is of main interest for the Data Mining process when the data under investigation presents this feature. This process aims at extracting relevant and interesting knowledge from large data sets, such that knowledge can be used to support the decision-making process.

Machine Learning (ML) is among the many areas that give support to Data Mining. Nevertheless, most of the ML methods do not deal directly with the temporal characteristic, as they assume that the data are independent and identically distributed (i.i.d.). However, since time series are time-oriented data, the occurrence of an observation in a certain instant of time usually depends of previously observed values.

In this work, we present an approach that unifies two strategies [4, 5, 6]: the extraction of global characteristics and the discovery of motifs, both directly from the time series. The first strategy is a traditional method used in time series research, in which global data characteristics, for instance, descriptive statistics, are extracted. Notwithstanding, these global characteristics may not represent some important details and a more detailed analysis may be necessary. The second strategy considers motifs discovery and aims to evidence local views of the temporal data.

The proposed approach is applied on a case study of three natural datasets of patient ECG signals from UCR Time Series Classification/Clustering datasets [3].

The rest of this paper is organized as follows. In Section 2, we formally define the concepts related to the problem. Section 3 presents the related work. Section 4 introduces the proposed approach, which is experimentally evaluated in Section 5. The results are presented in Section 6. Discussion and conclusion are presented in Sections 7 and 8, respectively.

2. Background

In this section we present basic concepts and the terminology used in this paper.

2.1. Basic Concepts

A time series Z is an ordered collection of real-valued observations of length m, that is, $Z = (z_1, z_2, ..., z_m)$ with $z_t \in R$, for $1 \leq t \leq m$ [7].

In some domains, it may be necessary to transform the real-valued observations into symbolic values. In doing so, algorithms developed exclusively to work with symbolic sequences can be applied to time series [8].

A symbolic time series Z of length m' is defined as a collection of ordered values $Z = (z_1, z_2, ..., z_{m'})$ with $z_{t'} \in \Sigma$, for $1 \leq t' \leq m'$ where Σ is a finite alphabet of symbols.

Another important issue regards to the portion of the time series that is taken into account during the analysis process. Many methods analyze small portions of a time series. These are named subsequences. The objective is the identification of local characteristics or the reduction of the search space.

Given a time series Z of length m, a subsequence C of Z is a continuous sample of Z of length n, with $n << m$. Therefore, $C = (z_p , . . . , z_{p+n-1})$ for $1 \leq p \leq m-n+1$ [8]. A subsequence is extracted with a sliding window which consists of extracting all subsequences of length n of a time series Z of length m, resulting in subsequences $(z_1 , . . . , z_n)$, $(z_2 , . . . , z_{n+1})$, . . . , $(z_i , . . . , z_{n+i-1})$, for $1 \leq i \leq m-n+1$.

As mentioned before, the motifs discovery aims to evidence local views of the temporal data. For that, the match concept is necessary to determine the similarity between two subsequences. Consider a positive real number r and a time series Z containing a subsequence C beginning at position p and another M in position q, in which D is the distance between the two objects. If $D(C, M) \leq r$, then M is similar to C [8].

From the presented concepts, the idea of motifs can be formalized as follows: given a time series Z and a threshold r, the most significant motif, called Motif-1, located in Z is the C_1 subsequence which has the largest amount of non-trivial matches [4].

During the process of motif discovery, the best matches to a subsequence tend to be subsequences that begin just one or two points to the left or to the right of the subsequence in question. These subsequences are called trivial matches and should not be considered motifs [9].

Thus, the most significant k^{th} motif present in Z is the subsequence C_k having the k^{th} greatest amount of non-trivial matches and satisfying the condition $D(C_k, C_i) > 2r$, for all $1 \leq i < k$ [9].

3. Related Work

The development of methods to analyze temporal data and more specifically ECG data has been focus of several papers in a number of tasks such as prediction, description/summarization, classification and data visualization. Most of the classification problems are related to the extraction of relevant characteristics for the posterior induction of classifiers based mostly on artificial neural networks [10,1,11] and support vector machines [12].

In [13], the ECG exams are represented using coefficients obtained by the discrete wavelet transform, which are then used as input to an artificial neural network. Results of this work are promising, however, the intelligibility of the built models is quite complex and somewhat non intuitive, impairing the utilization of the embedded knowledge by the experts.

In [14], characteristics extracted based on statistical measures, as well as correlation dimension and entropy are given as input for induction of distinct classifiers such as artificial neural networks, decision trees and support vector machines. Although the use of decision trees and promising results, still the models have low intelligibility. Another strategy for inducing decision trees is presented in [15], which is also based on characteristic extraction considering specific events related to ECG.

The ensemble approach to combine multiple neuronal classifiers in an efficient system for classification of ECG exams was proposed in [16]. Nevertheless, in spite of the superior results of this approach when compared to the approach without ensemble, the problem of intelligibility still remains.

In general, most of the studies combine different characteristic extraction and machine learning algorithms to increase the rates of classification problems involving ECG data. In this sense, many papers have achieved important results relevant to the area. Nevertheless, little effort has been devoted to build intelligible classifiers that enable and support experts in the analysis and understand of cardiovascular disorders.

4. Proposed Approach

As mentioned before, the proposed approach jointly applies two strategies to construct an attribute-value representation for the time series characteristics[1]. On one hand, part of the extracted characteristics is typically related to descriptive statistics such as average, standard deviation and maximums and minimums, which supply information on the global behavior of a time series. On the other hand, information about the existence of local behaviors is provided by motifs, which can be understood as a frequent subsequence present in a time series that have morphological similarity. Afterwards, this attribute-value representation will serve as input for ML algorithms.

The proposed approach consists of three main phases as shown in Figure 1.

4.1. Phase 1 (P1) – Time Series Preprocessing

In the first phase, time series are preprocessed aiming to solve some common problems found on temporal data, such as differences in scale and time intervals; data with noise; and presence of missing values.

4.2. Phase 2 (P2) – Characteristics Extraction and Motif Discovery

In this phase, features are identified using global character and local descriptions from time series data. Two inde compose this phase.

Stage 1 (S1-P2): Characteristics Extraction

In this stage, extracted characteristics usually include s mean, maximum and minimum values, and variance as dependent characteristics obtained, for instance, by i experts.

Figure 2 presents a schematic representation of this stag characteristics (*Ca1*, *Ca2*, *Ca3*, *Ca4*, *Ca5*) are used to rep series (T_1, T_2) in an attribute-value table. Vca_{ij}, with $i=1..2$ the characteristics values measured over the two time serie

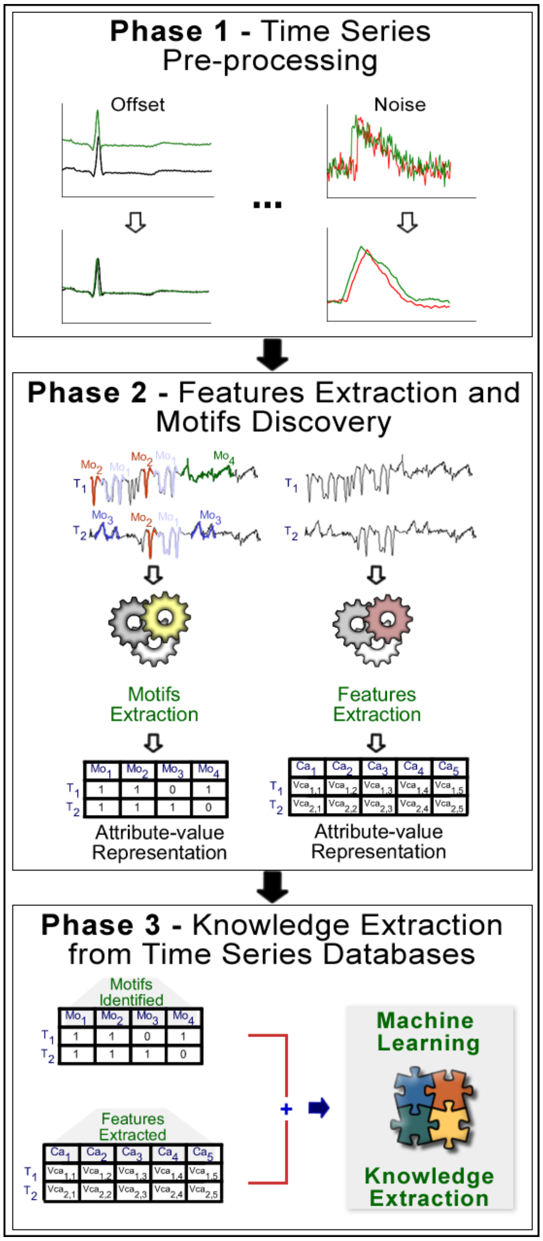

Phase 1 - Time Series Pre-processing
Offset
Noise
...
Phase 2 - Features Extraction and Motifs Discovery
Mo2 Mo1 Mo2 Mo1 Mo4
T1
Mo3 Mo2 Mo1 Mo3
T2
T1
T2
Motifs Extraction
Features Extraction
Mo1 Mo2 Mo3 Mo4
T1 1 1 0 1
T2 1 1 1 0
Ca1 Ca2 Ca3 Ca4 Ca5
T1 Vca1,1 Vca1,2 Vca1,3 Vca1,4 Vca1,5
T2 Vca2,1 Vca2,2 Vca2,3 Vca2,4 Vca2,5
Attribute-value Representation
Attribute-value Representation
Phase 3 - Knowledge Extraction from Time Series Databases
Motifs Identified
Mo1 Mo2 Mo3 Mo4
T1 1 1 0 1
T2 1 1 1 0
Features Extracted
Ca1 Ca2 Ca3 Ca4 Ca5
T1 Vca1,1 Vca1,2 Vca1,3 Vca1,4 Vca1,5
T2 Vca2,1 Vca2,2 Vca2,3 Vca2,4 Vca2,5
+
Machine Learning
Knowledge Extraction

Fig. 1: The three main phases of the proposed approach.

It is important to notice that the relevance of the characteristics to be used as features in the classification problem is data dependent. Therefore, the task of choosing the right characteristics requires data and domain knowledge.

Step 1 of S2-P2 – Subsequence matrix (SM) building: this process consists of extracting all subsequences of length *n* from the time series, using a sliding window. Each subsequence is then transformed in a string using the Symbolic Aggregate aproXimation (SAX) method [19]. This method considers an alphabet, with the size provided by the user. In Figure 3, we present an example with an alphabet of three symbols (*a*, *b*, *c*).

[1] In this paper, the words characteristics, attributes and features are used indistinctly.

The slide window extracts subsequences of length 16 which are discretized using SAX giving origin to strings of length $n_{sax} = 4$.

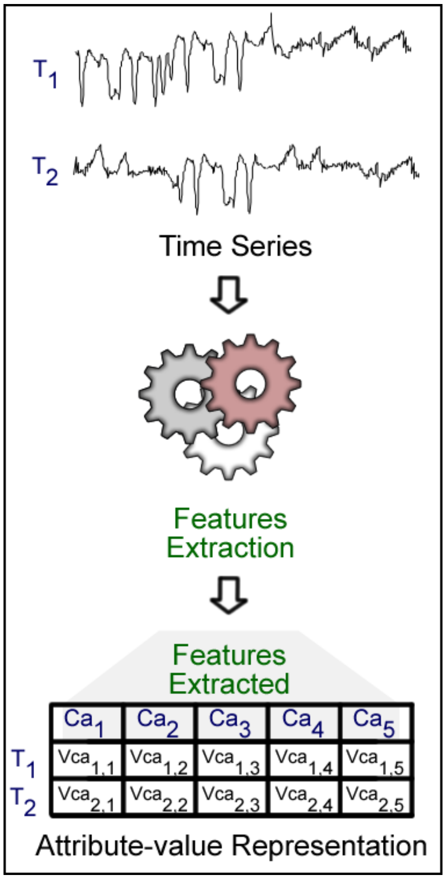

T_1
T_2
Time Series
Features Extraction
Features Extracted
Ca_1 Ca_2 Ca_3 Ca_4 Ca_5
T_1 $Vca_{1,1}$ $Vca_{1,2}$ $Vca_{1,3}$ $Vca_{1,4}$ $Vca_{1,5}$
T_2 $Vca_{2,1}$ $Vca_{2,2}$ $Vca_{2,3}$ $Vca_{2,4}$ $Vca_{2,5}$
Attribute-value Representation

Fig. 2: Schematic representation of the characteristic extraction stage, adapted from [4].

Step 2 of S2-P2 – Collision matrix (CM) building: the CM is used to identify subsequences that are likely to be motifs. The number of rows and columns are equal to the number of rows of the SM matrix. CM, initially null, is iteratively populated. A randomly chosen mask, different from the previous iteration, is used to indicate which columns of SM are currently active.

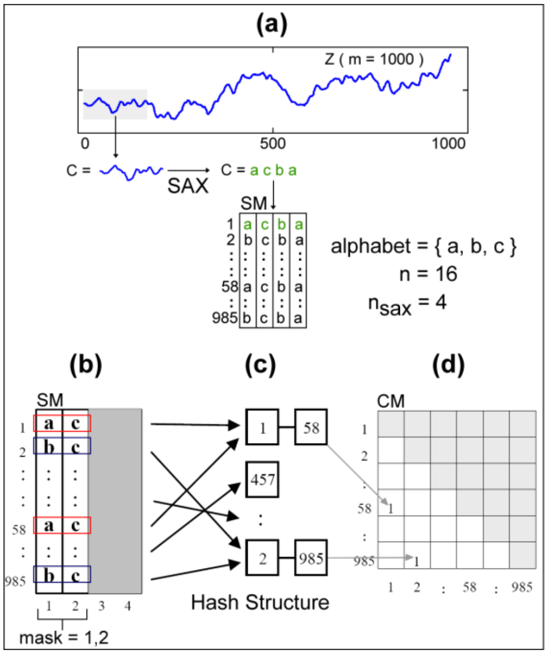


Fig. 3: Representation of the motif identification process, adapted from [7].

A hash key is formed by the active columns, and the hash table receives the subsequences according to those keys – Figure 3 (c). For example, in Figure 3 (b) the current mask is (1, 2); consequently, considering only the values of columns 1 and 2, the subsequences in the rows (1 and 58) and (2 and 985) of SM will collide since they have the same hash key. At the end of each iteration, MC is updated by counting the number of subsequences that collided – Figure 3 (d). This process is repeated for a determined number of times. The hash structure is cleaned and completed again for each iteration.

Step 3 of S2-P2 – Collision matrix analysis: a large value in a CM position is likely to be indicating the existence of a motif, although it is not a guarantee. Thus, the CM matrix is analyzed in order to find the location of the subsequences with the largest number of collisions. Then, the distance between those subsequences is calculated regarding the original data (real-valued). Two subsequences are considered motifs in the case that they are within a radius *r*. Other subsequences that fall inside the same radius are also identified as motifs [7]. In general, a sequential search is performed using the subsequence defined as motif over the entire time series.

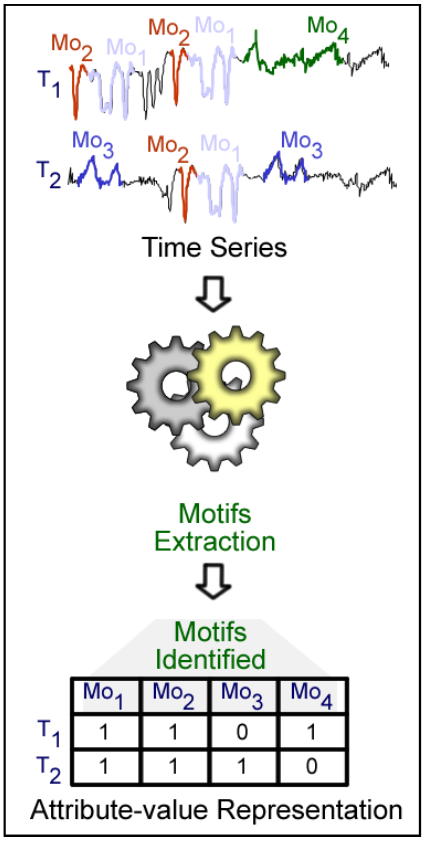
An empirical evaluation of this approach was performed in [18] and showed that it was significantly faster than the brute-force method (searches for the entire space of possibilities) and both methods identified the same motifs.

In Figure 4, we present a scheme of the described process with four motifs (Mo_1, Mo_2, Mo_3, Mo_4) that are identified in two time series (T_1, T_2). In the generated attribute-value table, the value "1" shows the presence of each motif in the time series and “0” otherwise. Notice that in this attribute-value table, instead of registering the presence or not of a motif, we may alternatively register the frequency or location of an identified motif.

Finally, the global characteristics, extracted at S1-P2, are combined with the identified motifs (S2-P2) in a final attribute-value table. In Figure 5, we show this process.

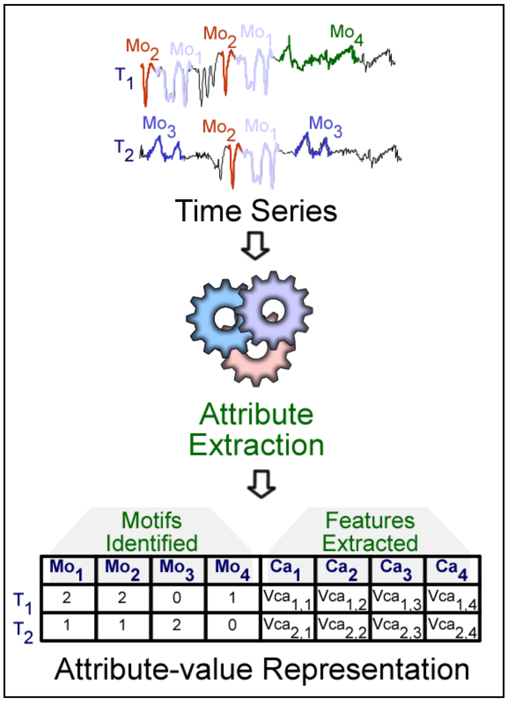
4.3. Phase 3 – Knowledge Extraction from Time Series Databases

In this phase, first the ML algorithms that will be used to build a prediction model or to explore and understand the data are defined. The choice for one algorithm or another algorithm needs to take into account the final objective of the extracted patterns. Then, the chosen algorithms are applied and the generated models can be evaluated using objective measures as well as qualitative ones.



	Mo_1	Mo_2	Mo_3	Mo_4
T_1	1	1	0	1
T_2	1	1	1	0

Fig. 4: Schematic representation of motif identification [4].



	Mo_1	Mo_2	Mo_3	Mo_4	Ca_1	Ca_2	Ca_3	Ca_4
T_1	2	2	0	1	$Vca_{1,1}$	$Vca_{1,2}$	$Vca_{1,3}$	$Vca_{1,4}$
T_2	1	1	2	0	$Vca_{2,1}$	$Vca_{2,2}$	$Vca_{2,3}$	$Vca_{2,4}$

Fig. 5: Attribute-value representation obtained with characteristics and motifs.

5. Experimental Evaluation

The proposed approach was evaluated on three ECG datasets (Table 1):

1. ECG (DB1): time series from different subjects related to supraventricular and non-supraventricular tachycardia;
2. ECGFiveDays (DB2): time series of the same subject recorded at interval of five days;
3. TwoLeadECG (DB3): time series from MIT-BIH Long-Term ECG Database about two-lead ECG records.

Table 1: Datasets summary description

	DB1	DB2	DB3
Number of Examples	200	884	1162
TS length	96	136	82
Classes	1 – 66,5% 2 – 33,5%	1 – 50,0% 2 – 50,0%	1 – 50,0% 2 – 50,0%
Majority Error	33,5 %	50,0%	50,0%

For each dataset, motifs with different sizes, as well as descriptive statistical characteristics were extracted with the following setup:

- **Motif size:** intervals of 1% to 10%, with increments of 1%, were considered to extract motifs; this percentage is related to the total number of each exams observations. This parameter was defined based on results presented in a previous work [6];

- **Statistical characteristics:** the mean, the global maximum and minimum for each time series were extracted;
- **Similarity threshold:** a 5% threshold was considered, i.e., the subsequences selected as motifs differ at most in 5% among themselves combining the Euclidean distance and the area of each subsequence;
- **Dimensionality reduction window:** no dimensionality reduction method was applied in this work;
- **Mask size:** the mask size considered in this work was two, corroborating with the work of [7];
- **Alphabet size:** in this work, we considered an alphabet composed by six symbols;
- **Number of iterations:** iterations performed corresponded to 50% of all possible combinations of masks.

The final attribute-value table contains the frequency of each identified motif as well as the statistical characteristics extracted for the respective time series (Figure 5).

In order to evaluate the capacity of description of the proposed approach, using motifs of different sizes in conjunction with the statistical characteristics, classifiers were constructed using each one of the three datasets mentioned before and their classification errors were measured using the 10 fold cross-validation sampling method.

Classifiers were induced using the *J*48 algorithm for decision trees available at the WEKA tool and considering the default parameters values [20].

The complexity of the induced classifiers was also measured by counting the number of node leafs.

In order to compare the proposed approach with the naïve method, the same procedure, maintaining the same folds, was applied providing the time series directly as input to the *J*48 algorithm. Results were analyzed using the paired *t*-test [21] for paired Gaussian data.

6. Results

Table 2 presents the mean error and its respective standard deviation for the induced classifiers for each one of the datasets. The smaller error values are marked in bold letters and the ones that present Significative Statistical Difference (s.s.d.) are marked at the column identified with s.s.d.

Table 2: Mean error results for classification task

Datasets	**Error**		
	Proposed Approach	**Naïve Method**	**s.s.d.**
DB1	**9,07**(8,93)	26,04(9,17)	▲
DB2	**3,66**(2,27)	4,30(2,07)	
DB3	**3,44**(1,34)	4,99(2,40)	

Table 3 shows the complexity of the induced decision trees for each one of the three datasets using *J*48.

Table 3: Average complexity of the inducted trees

Datasets	**Number of leaf node**		
	Proposed Approach	**Naïve Method**	**s.s.d.**
DB1	**7,80**(4,37)	11,10(1,60)	
DB2	20,00(3,71)	**17,40**(1,43)	
DB3	**12,70**(2,71)	23,70(1,70)	▲

7. Discussion

In this work, the proposed approach presented smaller mean errors than the ones measured using the naïve method for all the three datasets. Nevertheless, in only one case it was possible to verify s.s.d.

Considering the model complexity, the proposed approach presented smaller complexity in two of the three datasets with one case of s.s.d.

These results indicate the superior classification performance of the proposed approach. On the top of that, the intelligibility of the models generated by the naïve method is not as good as the intelligibility of the proposed approach, since in the naïve method only specific time instants are considered and not the global and local behaviors as the proposed approach does.

Most of the previous work applied non-symbolic algorithms due to the temporal nature of the data. Our experiments show we can obtain competitive classification performance even when symbolic algorithms are used. In addition, our representation using motifs and characteristics usually lead to very simple models that can be easily interpretable by computational non-experts.

In this sense, the use of motifs permits the analysis of specific behaviors, embedded in small portions of the time series, which may help with the identification of anomalies, possibly in initial stages, contributing with the early identification of determined diseases.

It is important to notice that determining the size of the motif is a complex task. In [6] a previous study was performed considering other ECG datasets demonstrating that motifs with better description potential are the ones with smaller sizes. This characteristic was considered when performing the present work.

Another parameter that needs adjustment is the similarity measure. Although many studies, such as the one presented in [22], point that the Euclidean distance, in average, performs well, there are other metrics or combination of them that may be considered to the comparison of subsequences [23].

The rest of the parameters are not directly related with the acceptance of subsequences as motifs.

Another important issue to be considered is that results may vary according to the algorithm used to induce the classifiers. In this work, as we are interested also in the intelligibility of the models, we choose the induction of decision trees which are easier to be interpreted by domain specialists.

8. Conclusion

In this work, we presented an alternative approach in order to mine time series using machine learning algorithms. The proposed method rendered competitive results, superior in some cases when compared to the naïve method. In addition, it provides a more intelligible way of representing the time series based on global aspects and morphological characteristics. Thus, it is possible to induct more intelligible models as well. This important feature may contribute with the identification, for example, of anomalies in early stages, which may be represented by some of these patterns.

It is important to emphasize the contribution of the approach using motifs and global characteristics for the induction of symbolic models, especially in the medical field. In these cases, the analysis and the verification of the generated conclusions are fundamental issues to the research, diagnosis, treatment and prevention of diseases.

9. References

[1] D. Thanapatay, C. Suwansaroj and C. Thanawattano, "Ecg beat classification method for ecg printout with principle components analysis and support vector machines," *Proc. Of the International Conference on Electronics and Information Engineering*, pp. 72-75, 2010.

[2] Ministério da Saúde, "Elsa brasil: maior estudo epidemiológico da américa latina," *Revista de Saúde Pública*, vol 43, no 1, pp. 40, Feb. 2011.

[3] E. Keogh, Q. Zhu, B. Hu, B., Hao Y., X. Xi, L. Wei, and C. A. Ratanamahatana. (2013, April 2013). The UCR Time Series Classification/Clustering Homepage[Online]. Available: http://www.cs.ucr.edu/~eamonn/time_series_data/

[4] A. G. Maletzke, "Uma metodologia para extração de conhecimento em séries temporais por meio da identificação de motifs e da extração de características," M.S thesis, Universidade São Paulo, São Carlos, SP, Brasil, 2009.

[5] A. G. Maletzke, G. E. Batista, H. D. Lee, and F. C. Wu, "Mineração de séries temporais por meio da extração de características e da identificação de motifs". *Proc. of the VII Encontro Nacional de Inteligência Artificial* (ENIA) at the *XXIX Congresso da Sociedade Brasileira de Computação* (CSBC), pages 1–10, Bento Gonçalves, RS, Brasil, 2009.

[6] A. G. Maletzke, H. D. Lee, W. Zalewski, J. T. Oliva, R. B. Machado, C. S. R. Coy, J. J. Fagundes, and F. C. Wu, "Estudo do parâmetro tamanho de motif para a classificação de séries temporais de ecg". *Proc. of the Congresso da Sociedade Brasileira de Computação in Workshop de Informática Médica*, pages 1-10, Natal, Rio Grande do Norte, 2011.

[7] B. Chiu, E. Keogh and S. Lonardi, "Probabilistic discovery of time series motifs," *Proc. Of International Conference on Knowledge Discovery and Data Mining*, pp. 493–498, 2003.

[8] M. Last, A. Kandel and H. Bunke, *Data Mining in Time Series Databases*. Denver: MA, World Scientific, 2004.

[9] J. Lin, E. Keogh, S. Lonardi and P. Patel "Finding motifs in time series," *Proc. Of the Second Workshop on Temporal Data Mining at the Eighth International Conference on Knowledge Discovery and Data Mining*, pp 53–68, 2002.

[10] V. E. Neagoe, I. F. Iatan and S. Grunwald, "A neuro-fuzzy approach to classification of ecg signals for ischemic heart disease diagnosis," *Proc. Of the Annual Symposium Proceedings Archive at the Americam Medical Informatics Association*, pp 494–498, 2003.

[11] T. Mar, S. Zaunseder, J. P. Martinéz, M. Llamedo and R. Poll. "Optimization of ecg classification by means of feature selection," *Biomedical Engineering*, vol. 58, no. 8, pp 2168-2177, Aug. 2011.

[12] L. Almazaideh, K. Elleithy, and M. Faezipour "Obstructive sleep apnea detection using SVM-based classification of ECG signal features", *Proc. of the Annual International Conference of the IEEE EMBS*, pp 4938-4941, 2012.

[13] S. N. Yu, and Y.H Chen, "Electrocardiogram beat classification based on wavelet transformation and probabilistic neural network." *Pattern Recognition Letters*, vol. 28, no. 10, pp 1142–1150, July 2007.

[14] A. Jovic and N. Bogunovic, "Ele-trocardiogram analysis using a combination of statistical, geometric, and nonlinear heart rate variability features." *Artificial Intelligence in Medicine*, vol. 51, no. 3, pp 175-186, March 2011.

[15] F. Charfi and A. Kraiem, "Comparative study of ECG classification performance using decision tree algorithms", *International Journal of E-Health and Medical Communications*, 3(4):102-120, doi:10.4018/jehmc.201 2100106.

[16] S. Osowski, K. Siwek and R. Siroic, "Neural system for heartbeats recognition using genetically integrated ensemble of classifiers", *Computers in Biology and Medicine*, vol 41, no. 3, pp. 173-180, March 2011.

[17] J. Buhler and M. Tompa, "Finding motifs using random projections," *Journal of Computational Biology*, vol. 9, no. 2, pp. 225–242, Feb. 2002.

[18] A. G. Maletzke, G. E. Batista, and H. D. Lee, "Uma avaliação sobre a identificação de motifs em séries temporais," *Ann. Congresso da Academia Trinacional de Ciências*, vol. 1, pp 1–10, 2008.

[19] J. Lin, E. Keogh, S. Lonardi and B. Chiu, "A Symbolic Representation of Time Series, with Implications for Streaming Algorithms," *Proc. Of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, pp 2-11, 2003.

[20] I. H. Witten and E. Frank, *Data mining: practical machine learning tools and techniques*. CA: Elsevier, 2005.

[21] H. Motulsky, *Intuitive Biostatistics*. NY: Oxford University Press, 1995.

[22] E. Keogh and S. Kasetty, “On the Need for Time Series Data Mining Benchmarks: a survey andempirical demonstration”, *Proc. of the 8th International Conference on Knowledge Discovery and Data Mining*, pp 102-110, 2002.

[23] J. Aikes Junior, H. D. Lee, C. A. Ferrero, and F. C. Wu, “Estudo da Influência de diversas Medidas de Similaridade na Previsão de Séries Temporais utilizando o Algoritmo kNN-TSP”, *Proc. Encontro Nacional de Inteligência Artificial in Series Brazilian Conference on Intelligent System,* pp 1-12, 2012.