Journal of Medical Statistics and Informatics
ISSN 2053-7662 | Volume 4 | Article 8

HOAJ
Herbert Open Access Journals

Original Research Open Access

Categorizing atrial fibrillation via Symbolic Pattern Recognition

Oguz Akbilgic[1,2*], John A. Howe[3,4] and Robert L. Davis[1]

*Correspondence: oakbilg1@uthsc.edu

[1]UTHSC-ORNL Center for Biomedical Informatics, Memphis, TN, USA.
[2]Department of Preventive Medicine, University of Tennessee Health Science Center, USA.
[3]King Abdullah Petroleum Studies and Research Center, Riyadh, Saudi Arabia.
[4]Risk Dynamics Consultancy, Istanbul, Turkey.

Abstract

In this research article, we apply clustering within the Symbolic Pattern Recognition (SPR) framework to problems related to classifying different clinical categories of atrial fibrillation by modeling the changes in electrical activity of the heart. SPR characterizes a sequential dataset by modeling the transition behavior exhibited by patterns of symbols; clearly, this technique requires continuous data to be discretized into a set of defined symbols. With SPR, we were able to find hidden patterns in electrocardiograms (ECG) recorded during normal sinus rhythm that allowed us to classify patients as having paroxysmal atrial fibrillation (PAF) vs. those that did not. Even without extensive tuning of the model, our correct classification rate of 80% is inline with other published models. Additionally, we were able to identify normal sinus rhythm ECGs of PAF patients when a PAF episode was imminent. Finally, we used SPR clustering to distinguish between episodes of atrial fibrillation which would end within one minute (spontaneously-terminating) vs. those which needed intervention to stabilize (sustained). These are very important considerations for clinical practitioners for several reasons. The ability to screen for, and diagnose, PAF even with no known history or ongoing episode would be invaluable. This is especially true as related to elderly patients whom are at greater risk from atrial fibrillation, many of whom undergo regular ECG screenings anyway. Secondly, early warning that a PAF episode is imminent can give caregivers the chance to prepare an appropriate intervention in advance. For certain patients, this could mean the difference between life and death. Lastly, it is recognized that intervention to stabilize atrial fibrillation is not always in the best interest in the patient. One consideration is how long the episode is expected to last; in many cases, it may be better to allow an episode to spontaneously terminate.

Keywords: Symbolic Pattern Recognition, clustering of sequential data, ECG, atrial fibrillation, paroxysmal atrial fibrillation, cardiac arrhythmia, time series modeling

Background

For the human heart to pump blood efficiently, the muscular layer (the myocardium) must be electrically stimulated and respond in a manner conducive to a rhythmic pattern of activation and deactivation; this is called normal sinus rhythm. However, there are many conditions that can disrupt normal cardiac rhythm. The general term for abnormal pumping speed and/or rhythm is cardiac arrhythmia, and there are many types and causes of arrhythmia. A necessary condition for the optimal rhythmic pumping is that the electrical signals need to propagate from the sinoatrial node throughout the myocardium in a periodic, systematic fashion.

One irregular pattern and/or rate of signal propagation and myocardium stimulation is called atrial fibrillation (AF). Atrial fibrillation is, in fact, the most prevalent abnormal heart rhythm condition with serious consequences [14]. In 2013, atrial fibrillation and atrial flutter together resulted in over 110,000 confirmed deaths [9]. As of 2014, about 2%-3% of the population were afflicted with AF [24].

While many people exhibit no symptoms from atrial fibril-

doi: 10.7243/2053-7662-4-8

lation, it is a potentially serious condition; it can increase the risk of stroke, and even lead to heart failure, dementia, and death [15,18].

There are two broad approaches to treating atrial fibrillation: slowing the heart rate to normal (rate control) or correcting the rhythm to a normal sinus rhythm (rhythm control) [7]. Of course, there are often side effects and unwanted consequences of any medical intervention. Hence, one problem related to AF is accurately distinguishing between AF that will stop by itself shortly, called spontaneously-terminating AF, and sustained AF. In many cases of the former, medical intervention may not be necessary, or indeed may be harmful. Atrial fibrillation can be further classified into four types based on duration and frequency [6]:

- *First Detected*: first diagnosed episode
- *Paroxysmal*: recurrent episodes which cease without intervention before 7 days
- *Persistent*: recurrent episodes which last longer than 7 days
- *Permanent*: on-going episode

Neglecting newly detected AF, among patients with paroxysmal, persistent or permanent, about half of patients who experience atrial fibrillation end up being classified as permanent. The remaining patients are classified evenly between the persistent class, and as paroxysmal atrial fibrillation (PAF) [24]. Relatively longer episodes of PAF are associated with increasing risk of ischemic stroke [1,20,21]. Approximately 18% of PAF cases evolve into permanent AF within 4 years [4]. Unfortunately, accurate diagnosis of PAF can be difficult - especially when it occurs with short periods of time. Therefore, it is important to develop methods that can help identify PAF during normal sinus rhythm.

Quantitative diagnostic evaluation of atrial fibrillation can use several techniques, including surface electrocardiogram (ECG), transthoracic ECG, transesophageal ECG, exercise stress test, and correlation of heart rate response to exercise. The arrhythmia can be easy to visually identify in an ECG recording. During normal sinus rhythm, the P wave, *QRS complex* (the triplet of deflections in the spike), and T wave should be clearly visible, as seen in the bottom of **Figure 1**.

In 2001, a challenge was issued jointly by PhysioNet and Computers in Cardiology on the subject of applying data mining to the problem of predicting PAF. There were two events described in this challenge. In event 1, called PAF screening, the challenge was to correctly classify subjects into PAF and non-PAF groups. Separately, in event 2, called PAF prediction, the challenge was to determine for which ECG sets of PAF subjects a PAF episode was imminent of PAF. A detailed description of the challenge and summary of entrants and their classification/prediction performances can be found in [13]. The team of Schreier, Kastner, and Marko [19] obtained the best results for the first event, correctly classifying 82% of patients, and predicting 41/50 of the PAF episodes. In the second event, their performance was 71% and 20/28, respectively. Their technique used standard preprocessing techniques and statistical hypothesis testing procedures. More recently, in his thesis [10], Gilani used various feature selection/extraction/engineering and supervised classification techniques to develop models to classify atrial fibrillation and screen for PAF.

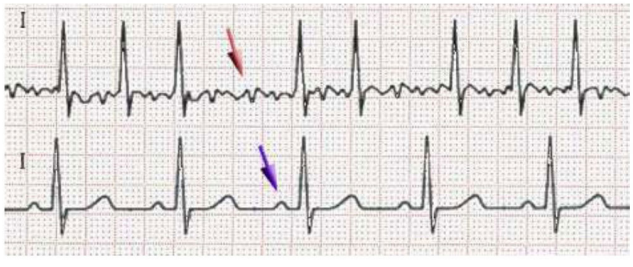

Figure 1. ECG During Atrial Fibrillation (Top) and Normal Sinus Rhythm (Bottom). The Purple Arrow Indicates a P Wave, Which is Lost in Atrial Fibrillation.†

Several authors have analyzed the time interval between successive R-R peaks in the QRS complex in order to model AF. Krstacic et al. [12] used inductive machine learning by logic minimization to identify ratios of the interval which tended to precede onset of AF. Sun and Wang [23] applied a multilayer perceptron neural network and fuzzy support vector machine to several engineered features of the R-R interval data to distinguish between sustained and spontaneously-terminating AF, with an accuracy rate over 90%. This work was an extension of their previous work, which used other structural characteristics of the QRS complex [22]. Alcaraz and Rieta [5] and Nilsson et al. [16] both analyzed the organization properties of the main atrial wave from ECGs to classify atrial fibrillation as sustained vs. spontaneously-terminating; both efforts resulted in approximately 90% correct classification of known cases in their respective testing sets, which came from the 2004 PhysioNet / Computers in Cardiology challenge [8,11].

Objectives

In this work, we implement a novel method called Symbolic Pattern Recognition (SPR) [2] to identify underlying patterns and signals of AF in normal sinus rhythm ECG recordings. The SPR method is designed to model the pattern transition behaviour of symbolic series. Therefore, in using SPR for clustering of raw ECG data, we begin by applying a symbolic discretization to the continuous ECG series into an alphabet of discrete symbols. The discretized series is then composed into pattern transition matrices that define the predicted behavior of the series, given the observed patterns. SPR defines a procedure whereas unsupervised clustering of series can be performed, based on a similarity measure of these transition matrices.

In this research, we first apply Symbolic Pattern Recognition to cluster normal sinus rhythm ECG data and distinguish between people with and without PAF. Secondly, we use SPR clustering of R-R interval data to classify PAF episodes as sustained vs spontaneously-terminating, as well as to identify normal sinus rhythm ECGs for which an episode of PAF

doi: 10.7243/2053-7662-4-8

is imminent. The first example was originally presented as a poster at the Southern Regional Council of Statistics Summer Research Conference [3].

Methods

The Symbolic Pattern Recognition (SPR) framework of [2] is based on learning pattern transition behaviors in sequential data. Based on these probabilistic transition behaviors, the researcher can characterize and cluster series. For completeness, we've repeated some of [2] in this section.

Learning pattern transition behaviour

SPR aims to learn and model the pattern transition behavior in a discrete series represented with ns unique symbols. To do this, SPR looks for joint occurrences of observed patterns of length n_p followed by a single symbol. For example, a two-symbols pattern ba followed by d or a three-symbols pattern acd followed by c. This maximum pattern length n_p is an important parameter of SPR; especially when very long sequences of data are modeled.

By observing the frequency with which these patterns and transitions occur, we can infer the transition probabilities governing how the series evolves. As an example, consider a sample series $S=\{aabcabccbabcabcbaabc\}$, defined over the alphabet $\{a, b, c\}$. We see the pattern ab occurring five times, always followed by a c: $\{a|abc|abc|cb|abc|abc|ba|abc|\}$. **Table 1** shows the i-symbols pattern transition frequencies (PTF_i) and probabilities (PTP_i) for S, with patterns of length $i = n_p = 2$, in lexicographical order. We see that the bc pattern is observed in S four times (5th column). Furthermore, we observe bca twice, and bcb/bcc each once. According to PTP2, if we later observe either aa, ca, or cb we can reasonably expect the next symbol will be b. We acknowledge that the uncertainty associated with probabilities for rarely observed patterns may be quite high. However, the impact of such rarely observed patterns can be reduced by using frequencies as weights.

For an alphabet of length ns, there are ns^{np} possible n_p-length patterns. Hence, PTF_i is a matrix of at most $ns^i \times (ns+1)$ size,

Table 1. PTF_2 and PTP_2 for S.

	PTF_2				PTP_2		
	a	***b***	***c***	**Total**	***a***	***b***	***c***
aa	0	2	0	2	0.00	1.00	0.00
ab	0	0	5	5	0.00	0.00	1.00
ba	1	1	0	2	0.50	0.50	0.00
bc	2	1	1	4	0.50	0.25	0.25
ca	0	2	0	2	0.00	1.00	0.00
cb	2	0	0	2	1.00	0.00	0.00
cc	0	1	0	1	0.00	1.00	0.00

and PTP is at most $ns^i \times ns$. If any of the possible patterns are unobserved, the matrices will be accordingly smaller; note how ac is never observed in S, so is not shown in **Table 1**.

For a given series of length n, there are up to n_p pattern transition matrices calculated, with n_p an integer ranging from 1 to n-1. Optimal determination of n_p is important to balance computational cost vs. information modeled. When n_p is small, the method will not truly learn the pattern transition behavior of the series. On the other hand, setting n_p too high will be computationally expensive, and result in a sparse transition matrix. To select the optimal value for n_p, we use a criterion called the Sparsity Index (SI_i), which is defined to be the ratio of the number of observed patterns to the number of possible n_p-length patterns. As an example, for the diagram patterns in S, we have observed 7 patterns (aa, ab, ba, bc, ca, cb, cc) in **Table 1**, while there are 9 possible such patterns (aa, ab, ac, ba, bb, bc, ca, cb, cc), so SI_2=7/9=0.778. The sparsity indices for S are shown in **Table 2**.

Table 2. Sparity Indices SI_i for S.

i	**Observed**	**Possible**	***SI*** $_i$
1	3	3	1.000
2	7	9	0.778
3	10	27	0.370
4	13	81	0.161
5	13	243	0.054
6	13	729	0.018
7	13	2187	0.006
8	12	6561	0.002
9	11	19683	0.001
≥10	--	--	<0.001

In the SPR algorithm, we sequentially compute PTP_i and SI_i, looping for $i=1, 2, \ldots min\ (ns, n_p)$, using a subjectively-set threshold tn_p to determine n_p.

1) set $i=1$
2) compute PTP_i and SI_i
3) if $SI_i \geq tn_p$, increment i and go to step 2, otherwise, set $n_p = i-1$ and exit loop

Using the sparsity indices for S in **Table 2**, we set $n_p=2$ when the threshold is 0.5, but if $tn_p=0.1$, we would prefer $n_p=4$. It is true that this has merely pushed the optimization back from determining n_p to determining tn_p. However, as the primary purpose of tn_p is as a stopping criterion that allows for sufficiently large n_p, we can set it to a low value, such as $1E\text{-}4$. Alternatively, for series that are not too long, we may instead compute all transition matrices for $i = 1, 2, \ldots, n\text{-}1$, then plot i vs. SI_i. As with interpretation of scree plots, we can then select n_p by identifying the largest deflection point, or when the curve becomes almost horizontal.

Clustering with SPR

By using the SPR framework to model pattern transition behavior, we are essentially reducing a discrete series into a matrix of pattern transition probabilities. Therefore, series with similar

doi: 10.7243/2053-7662-4-8

pattern transition behavior should exhibit similar *PTP*s. The framework specifies a pattern transition similarity (*PTS*) measure as a function of distance between a pair of *PTP*s. We would like two identical series to have *PTS*=0, indicating perfect similarity. To simplify the calculations, the *PTP*s should be the same size with any missing pattern transition probabilities zero-filled. After this zero-fill, we compute the *PTS* of two series D_i and D_j by aggregating the absolute distance between their respective matrices, PTP^i_k and PTP^j_k, for $k=1,2,\ldots,N_p$, with N_p indicating the largest n_p.
For the *kth* pair of *PTP* matrices, we compute the distance as

$$dist\left(PTP_k^i, PTP_k^j\right) = \sum_{r=1}^{N_p k} \sum_{c=1}^{N_s} \left| PTP_k^i(r,c) - PTP_k^j(r,c) \right| \quad (1)$$

Where *Npk* indicates the number of patterns in the *kth PTP* matrices, and *Ns* is the length of the longest alphabet. The distance $dist\left(PTP_k^i, PTP_k^j\right)$ is hence the sum of absolute differences in transition probabilities. Finally, we compute the similarity $PTS_{i,j}$ as

$$PTS_{i,j} = \sum_{k=1}^{N_p} dist\left(PTP_k^i, PTP_k^j\right) \quad (2)$$

A logical extension of simple comparison, we can leverage the *PTS* scores to perform unsupervised clustering of *m* discretized series by computing and comparing PTS_s for all pairs of series under consideration. This results in an *m*×*m* symmetric matrix with $PTS_{i,j}$ being the distance between the i^{th} and j^{th} series; clearly the diagonals must be 0.

We return to our series *S* once again to demonstrate measuring similarity and dissimilarity. We want to compare *S* to two series simulated by SPR (see [2, Section3.3]): S^*=*abcbaabcabccbabcabcb* and S^{**}=*bcbabcbaababcbababcc*. We simulated S^* using n_p=6, while S^{**} was simulated using n_p=2. Logically, S^* should capture the pattern transition behavior in *S* better than S^{**}, so should be more similar. The *PTS* matrix shown below indicates that S^* is indeed more similar to *S* than S^{**} is.

PTS=

	S	S*	S**
S	0.00	0.43	4.08
S*	0.43	0.00	3.99
S**	4.08	3.99	0.00

Visual inspection of the *PTS* matrix, and/or a visual representation of it can guide researchers in developing clusters. In addition to visual inspection, we can identify meaningful patterns with statistical analysis of the *PTS* matrix.

Results

In this section, we apply SPR clustering to two symbolically discretized ECG datasets. The first is from the 2001 Physio Net/Computers in Cardiology challenge and the second is from their 2004 challenge.

Searching for finger prints of paroxysmal atrial fibrillation

The data for this example is the training set of the PAF prediction database from the 2001 PhysioNet/Computers in Cardiology challenge [11,17]. It consists of excerpts of two-channel long-term Holter monitor ECG recordings, measured at a frequency of 128Hz, from 75 individuals. Fifty of the patients, whose 30-minute ECGs we'll annotate as NN, have not been diagnosed with PAF; they are considered to be the control group. ECG excerpts from the remaining 25 patients are categorized into sets of three series:

PAF_E 5-minute ECG recording during an episode of PAF
PAF_P 30-minute ECG recording immediately preceding an episode of PAF
PAF_N 30-minute ECG recording that is at least 45 minutes distant from any PAF episode

PAF_P and PAF_N ECGs were recorded during normal sinus rhythm. The database is more thoroughly described in the challenge summary [13].

To apply Symbolic Pattern Recognition, we first discretized each continuous ECG recording using an alphabet of five symbols. Because of the skewed nature of the data, we applied an equal-frequency discretization, rather than equal-width. Cutoff points for a sample ECG recording is shown in **Table 3**, and demonstrated visually on one second of a normal sinus rhythm ECG in **Figure 2**, showing the entire QRS complex.

Table 3. Equal-Frequency Discretization Rules for Sample PAF ECG Recording.

Symbol	>	≤
a	-∞	-0.02
b	-0.02	0.03
c	0.03	0.11
d	0.11	0.32
E	0.32	∞

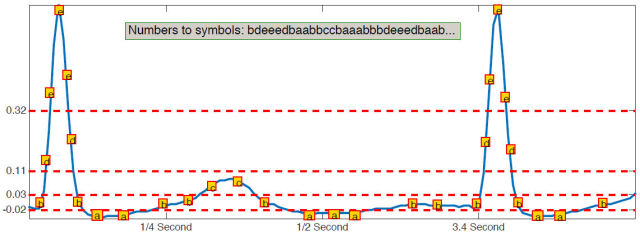

Figure 2. Equal-Frequency Discretization for Sample PAF ECG Recording.

After each ECG recording was individually discretized, we computed pattern transition matrices for each series up to PTP_7, and used this to build the partial pattern transition similarity matrix comparing the NN and PAF_N series to PAF_E. Overall computation time from loading data to creating the similarity matrix took 20 minutes on a machine with a 2.8 GHz Intel quad core i7 processor and 16 GB 1600 MHz DDR3 memory using 4 workers on Matlab 2015b parallel computing.

doi: 10.7243/2053-7662-4-8

This 75×25 matrix is too big to show here, but we show part of it in **Table 4**. Average distances across all 25 PAF_E series are shown in the last column of **Table 4**, and plotted in **Figure 4**.

Table 4. Partial Transition Similarity Matrix Comparing Normal Sinus Rhythm ECGs to PAF ECGs.

	PAF_{E1}	PAF_{E2}	…	PAF_{E25}
NN_1	2.75	2.89	…	3.35
NN_2	3.30	3.01	…	3.42
…	…	…	…	…
NN_{50}	3.30	2.70	…	2.96
PAF_{N1}	3.36	3.93	…	3.99
PAF_{N2}	2.16	1.93	…	2.98
…	…	…	…	…
PAF_{N25}	1.89	2.16	…	2.89

We began our analysis by testing for the equality of the mean of the NN and PAF_N average distances. The standard parametric t-test rejected the null hypothesis with a p-value <0.001. Furthermore, the nonparametric Mann-Whitney U test rejected the null hypothesis that the observations come from the same population with the same low p-value. We thus conclude that the discretization and SPR clustering results in the two being statistically different series with the ROC curve given in **Figure 3**.

By using a subjectively chosen cutoff point of 3.02, we can classify patients as either normal or with PAF. Applying the decision rule,

$$class = \begin{cases} Avg(dist) \geq 3.02 Normal \\ Avg(dist) < 3.02 PAF \end{cases}$$

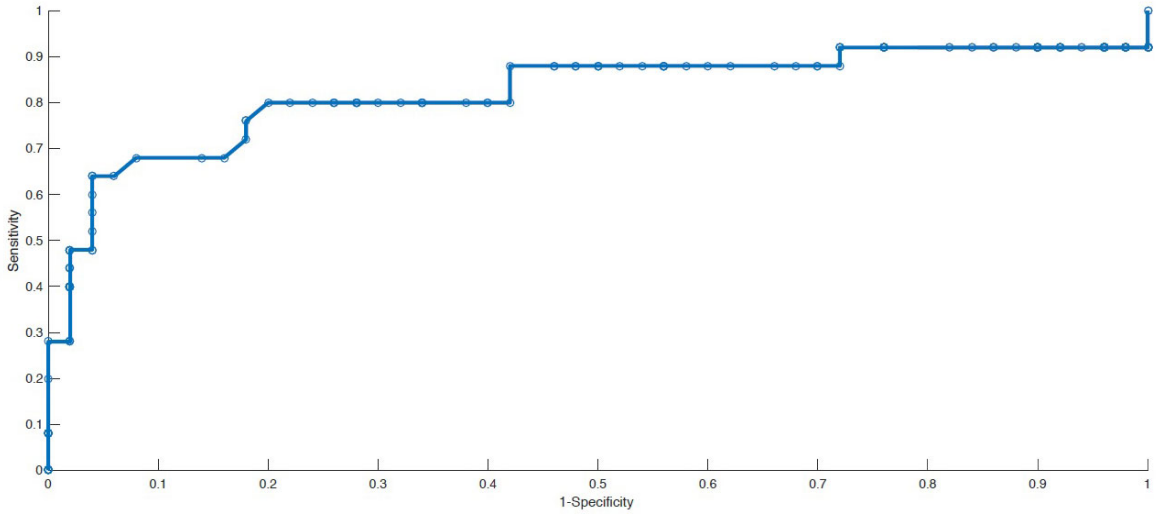

Figure 3. ROC curve for classification of ECGs into Normal and PAF.

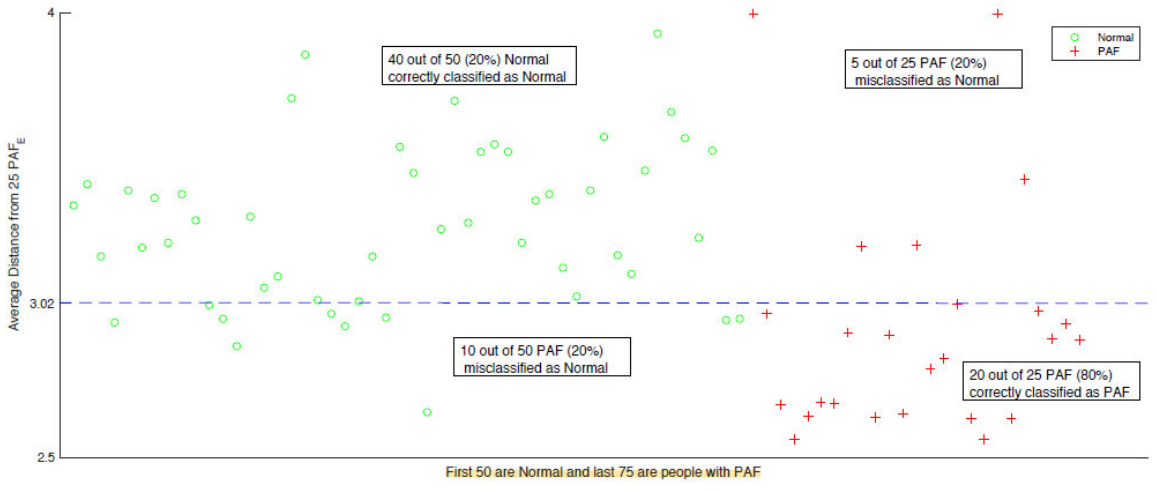

Figure 4. Average PTP Distances Showing the Decision Rule. Lower Distances on the y-axis Indicate that the Original ECG is more similar to the PAFE ECGs. Misclassified Patients are Highlighted.

doi: 10.7243/2053-7662-4-8

results in a classification accuracy of 80%, which is comparable to the best results from the PhysioNet challenge. The confusion matrix in **Table 5** shows that most of the misclassified observations are false positives, with only 5 (6.67%) false negatives.

Table 5. Confusion Matrix Using the Decision Rule.

		Predicted		
		Normal	**PAF**	**Total**
Actual	Normal	40	10	50 (Specificity = 80%)
	PAF	5	20	25 (Sensitivity = 80%)
	Total	45	30	75 (Accuracy = 80%)

We would have preferred to use the PhysioNet testing dataset mentioned in the challenge summary [**11**] to evaluate our decision rule. However, the autoscorer webpage is no longer available.

Nevertheless, our analysis with this small dataset suggests that clustering with SPR can create a decision rule to detect the underlying fingerprints of PAF even in ECGs recorded during normal sinus rhythm. In fact, recall that our SPR similarity matrix only included the PAF_E, PAF_N, and NN ECGs- the PAF_P data from immediately preceding a PAF episode was not used. This suggests the model is picking up on a subtle long-term difference in the patterns of electrical stimulation of the myocardium, implying success as a model to diagnose PAF would not rely on an imminent episode.

In addition to the need to diagnose PAF from a normal sinus rhythm ECG, it's important to be able to detect when a PAF episode is imminent. In terms of this dataset, that means distinguishing between PAF_P and PAF_N. However, when we computed the pattern transition similarities between NN, PAF_N, and PAF_P, we were unable to distinguish between PAF_P and PAF_N. Visual inspection of **Figure 5** shows that the mean of the average distances are very similar, and there is a tremendous amount of overlap in their 95% intervals.

In the next example, which focuses on a slightly different diagnostic problem, we see that we can distinguish between these two types of ECGs by applying a data transformation before a substantially different symbolic discretization.

Differentiating between sustained and spontaneously-terminating atrial fibrillation

The data for this example is from the PhysioNet / Computers in Cardiology 2004 challenge [**8**,**11**]. It is composed of 1-minute excerpts from ECGs - recorded at 128 Hz - of 30 patients during AF episodes.

They are divided into three equally-sized groups:

AF_{NonT} AF that was not observed to have terminated for the duration of the long-term recording, for at least an hour following the excerpt

AF_{T1min} AF that terminated within one minute after the end of the recording

AF_{T1sec} AF that terminated within one second after the end of the recording

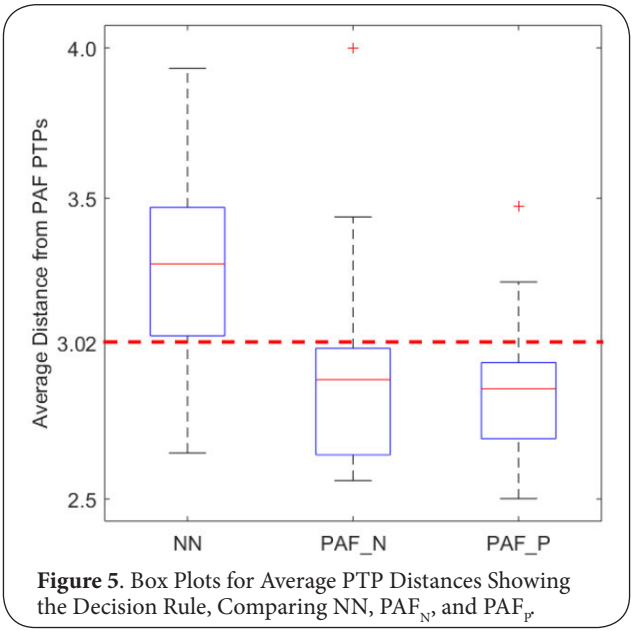

Figure 5. Box Plots for Average PTP Distances Showing the Decision Rule, Comparing NN, PAF_N, and PAF_P.

The data from PhysioNet included QRS annotations from an automatic detector, which we used to compute the R-R intervals. We then discretized the interval data using an alphabet of eight symbols according to the cutoff points in **Table 6**. Since R-R intervals associated with normal sinus rhythm typically vary between 0.6 and 1.2, we simply generated symmetric intervals around and within this range.

Table 6: Discretization Rules for AF R-R Interval Data. Symbols c Through f are Typically Associated with R-R Intervals Observed During Normal Sinus Rhythm.

Symbol	>	≤
a	0.00	0.30
b	0.30	0.60
c	0.60	0.75
d	0.75	0.90
e	0.90	1.05
f	1.05	1.20
g	1.20	1.50
h	1.50	∞

The same transformation and discretization was also applied to the PAF and NN data from the previous example, using only the final minute of each to be consistent with the AF data. Again, we computed the SPR pattern transition matrices up to PTP_8, then built a pattern transition similarity matrix comparing the PAF_N, PAF_P, AF_{NonT}, AFT_{1min}, and AFT_{1sec} data (rows) to the NN series (columns). Averaging the similarities for each series across all 50 patients in the control group resulted in the box plots shown in **Figure 6**.

The right-most three box plots are from the 30 ECGs related

doi: 10.7243/2053-7662-4-8

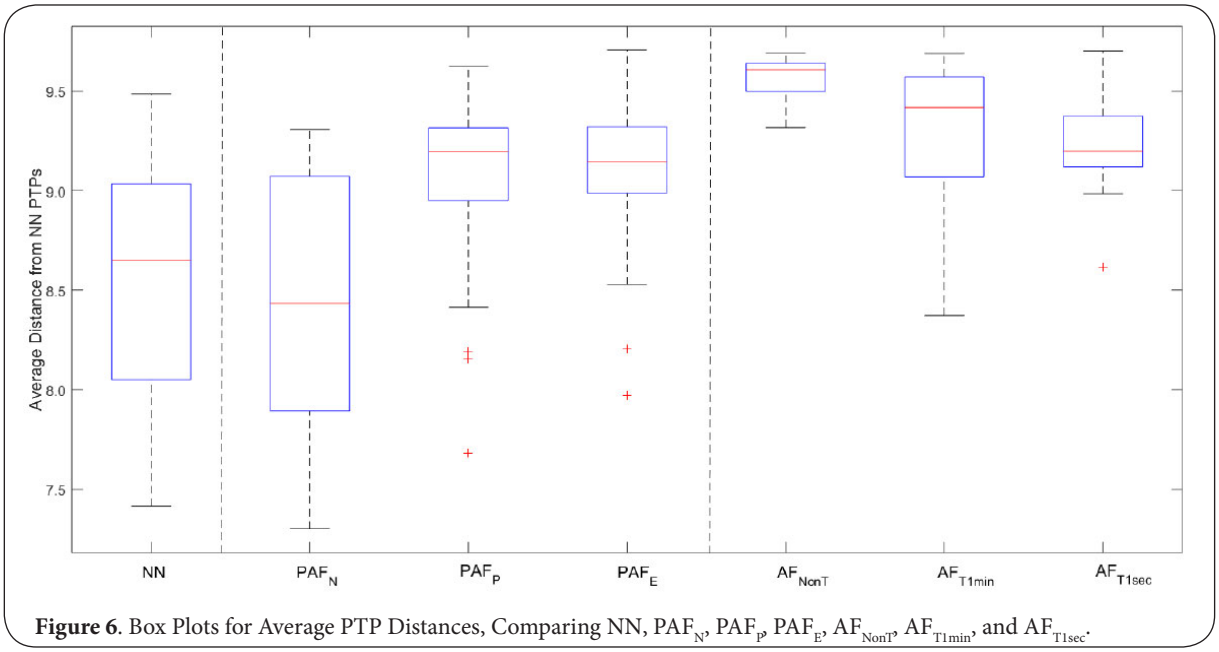

Figure 6. Box Plots for Average PTP Distances, Comparing NN, PAF_N, PAF_P, PAF_E, AF_{NonT}, AF_{T1min}, and AF_{T1sec}.

to AF termination. As compared to the pattern transition behavior in the control group, we see a clear difference between the PAF and AF data, in that AF ECGs are less similar to normal sinus rhythm ECGs. For all pairs of series, we computed the Mann-Whitney U nonparametric test for the null hypothesis of equal populations, with results shown in **Table 7**. We also applied the parametric t-test for equal means; the results are similar, so not shown here. We see that SPR clustering of the discretized R-R intervals separates all abnormal ECGs from the sinus rhythm ECGs. Additionally, we see that the PAF_P and PAF_N data are significantly different; this is also visible

Table 7. Mann-Whitney U-test for Equality of Populations for all Series. Red italic entries are statistically significant at α=0.05.

	NN	PAF_N	PAF_P	PAF_E	AF_{NonT}	AF_{T1min}	AF_{T1sec}
NN	-	0.862	*<0.001*	*<0.001*	*<0.001*	*<0.001*	*<0.001*
PAF_N	-	-	*0.001*	*<0.001*	*<0.001*	*<0.001*	*0.001*
PAF_P	-	-	-	<0.816	*<0.001*	<0.130	<0.571
PAF_E	-	-	-	-	*<0.001*	<0.149	<0.523
AF_{NonT}	-	-	-	-	-	<0.278	*0.032*
AF_{T1min}	-	-	-	-	-	-	<0.623

in **Figure 6**. Thus, we conclude that we can detect in normal sinus rhythm ECGs when a PAF episode is imminent, using the R-R intervals.

The mean of the average distances for the sustained AF records is approximately 9.57, while the averages for the spontaneously-terminating AF are 9.29 and 9.22 for the 1-minute and 1-second data, respectively. Merging the AF_{T1min} and AF_{T1sec} data, we then performed a t-test to evaluate the null hypothesis that the sustained and spontaneously-terminating AF series have the same underlying population mean; the p-value of 0.001 forces us to reject the null hypothesis. The Mann-Whitney U test for equal populations also rejected the null hypothesis, with a p-value of 0.017. We conclude that SPR clustering of the R-R interval data is able to distinguish between sustained and spontaneously terminating AF.

Discussion

In this research article, we have detailed how clustering within the Symbolic Pattern Recognition framework can classify and distinguish among different clinical categories of atrial fibrillation. In the first case, we analyzed the pattern transition behavior underlying discretized ECG recordings from patients during normal sinus rhythm, as well as during PAF episodes. Analysis with SPR was able to identify some underlying characteristics of PAF even in ECGs recorded during normal sinus rhythm that were only present in PAF patients. The pattern transitions during normal sinus rhythm were significantly more similar to those during PAF for patients that had been diagnosed with paroxysmal atrial fibrillation. We used the similarities to develop a simple classification rule that correctly classified 80% of patients in our dataset.

In the second case, we computed the sequence of intervals between R peaks in the QRS complex for the same data, plus another dataset composed of sustained vs. spontaneously-terminating atrial fibrillation. After symbolic discretization and analysis of the pattern transition behaviors, we were able to distinguish between patients not diagnosed with any form of AF, patients experiencing sustained AF vs spontaneously-terminating AF, and PAF patients about to experience an episode of PAF. A critical next step in our work is to validate the per-

doi: 10.7243/2053-7662-4-8

formance of our method on a larger dataset.

Conclusions

One key learning from this study is that, while SPR is a powerful method for analyzing sequential data series, selecting the right symbolic discretization is the key to successful application of SPR. This is likely especially true in the health sciences; expert opinions from the field on discretization of continuous data may improve the accuracy of SPR-based analysis. We believe that the SPR framework has potential for quantitative analysis in healthcare and the clinical environment. For the specific types of atrial fibrillation modeled here, our models can be used for screening and diagnosis, guiding intervention decisions, and early warning during monitoring.

In predictive modeling, predictive variables may be continuous, binary, categorical, or sequential data. It is easy to incorporate the first three (age, gender, income level etc...) into models because they are represented by a unique value for each subject. Conversely, representation of sequential data predictors (for example, systolic blood pressure records during surgery) as a unique value in predictive modeling is not a trivial task. Using simple descriptive statistics (average, standard deviation, quantiles, etc...) loses a substantial amount of information. Moreover, two series with same average values (and even variances) may be significantly different than each other in terms of the associated response outcomes. Instead of using basic descriptive statistics, we suggest that metrics from SPR - which incorporate pattern transition behavior of arbitrary length - could instead be integrated into these predictive models, and reduce information loss. In our examples, we obtained high classification accuracies, despite the fact that we did not use any patient characteristics (demographics, genetic risk factors, co-morbidities, etc...). We expect that the addition of such information into an AF classification model based on SPR would substantially increase the accuracy of our results.

Finally, the resulting predictive models could be embedded into wearable devices currently used to collect physiological data, such as a Holter monitor. An obvious benefit being that a real-time monitor could provide early warnings of the onset of adverse events, rather than simply recording them.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

Authors' contributions	OA	JAH	RLD
Research concept and design	✓	--	--
Collection and/or assembly of data	✓	--	--
Data analysis and interpretation	✓	✓	--
Writing the article	✓	✓	--
Critical revision of the article	✓	✓	✓
Final approval of article	✓	✓	✓
Statistical analysis	✓	✓	--

Publication history

Editor: Feng Gao, Washington University School of Medicine, USA.
Received: 18-Oct-2016 Final Revised: 28-Nov-2016
Accepted: 16-Dec-2016 Published: 26-Dec-2016

References

1. Abdul-Rahim AH and Lees KR. **Paroxysmal atrial fibrillation after ischemic stroke: how should we hunt for it?** *Expert Rev Cardiovasc Ther*. 2013; **11**:485-94. | Article | PubMed
2. Akbilgic O and Howe J. **Symbolic Pattern Recognition for Sequential Data**. *Sequential Analysis*. 2016.
3. Akbilgic O and R.L.Davis. **Searching for fingerprints of paroxysmal atrial fibrillation: A symbolic pattern recognition approach**. *Presented in SRCOS Summer Research Conference*. 2016. | Pdf
4. Al-Khatib SM, Wilkinson WE, Sanders LL, McCarthy EA and Pritchett EL. **Observations on the transition from intermittent to permanent atrial fibrillation**. *Am Heart J*. 2000; **140**:142-5. | Article | PubMed
5. Alcaraz R and Rieta JJ. **Sample entropy of the main atrial wave predicts spontaneous termination of paroxysmal atrial fibrillation**. *Med Eng Phys*. 2009; **31**:917-22. | Article | PubMed
6. American College of Cardiology/American Heart Association Task Force on Practice Guidelines and the European Society of Cardiology Committee for Practice Guidelines. **ACC/AHA/ESC 2006 Guidelines for the Management of Patients With Atrial Fibrillation**. *Circulation*. 2006; **114**:e257-e354. | Article
7. Anumonwo JM and Kalifa J. **Risk factors and genetics of atrial fibrillation**. *Cardiol Clin*. 2014; **32**:485-94. | Article | PubMed
8. G.B. Moody. **Spontaneous Termination of Atrial Fibrillation: A Challenge from PhysioNet and Computers in Cardiology 2004**. 2004. | Pdf
9. GBD 2013 Mortality and Causes of Death Collaborators, 2015. Global, regional, and national age–sex specific all-cause and cause-specific mortality for 240 causes of death, 1990–2013: **a systematic analysis for the Global Burden of Disease Study 2013**. *The Lancet*. **385**:117-171. | Article
10. Gilani M. **Machine Learning Classifiers for Critical Cardiac Conditions**. *Master of applied science thesis, University of Ontario Institute of Technology*. 2016.
11. Goldberger AL, Amaral LA, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng CK and Stanley HE. **PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals**. *Circulation*. 2000; **101**:E215-20. | Article | PubMed
12. Krstacic G, Gamberger D, Smuc T and Krstacic A. **Some Important R-R Interval Based Paroxysmal Atrial Fibrillation Predictors**. *Computers in Cardiology*. 2001; **28**:409-412.
13. Moody G, Goldberger A, McClennen S and Swiryn S. **Predicting the Onset of Paroxysmal Atrial Fibrillation: The Computers in Cardiology Challenge 2001**. *Computers in Cardiology*. 2001; **28**:113-116. | Article
14. Munger T, Wu L and Shen W. **Atrial Fibrillation**. *Journal of Biomedical Research*. 2014; **28**:1-17.
15. National Institutes of Health. National Heart, Lung, and Blood Institute. **What Is Atrial Fibrillation?** 2016. | Website
16. Nilsson F, Stridh M, Bollmann A and Sornmo L. **Predicting spontaneous termination of atrial fibrillation using the surface ECG**. *Med Eng Phys*. 2006; **28**:802-8. | Article | PubMed
17. PhysioNet. **The PAF Prediction Challenge Database**. 2001. | Website
18. Rienstra M, Lubitz SA, Mahida S, Magnani JW, Fontes JD, Sinner MF, Van Gelder IC, Ellinor PT and Benjamin EJ. **Symptoms and functional status of patients with atrial fibrillation: state of the art and future research opportunities**. *Circulation*. 2012; **125**:2933-43. | Article | PubMed Abstract | PubMed FullText
19. Schreier G, Kastner P and Marko W. **An Automatic ECG Processing Algorithm to Identify Patients Prone to Paroxysmal Atrial Fibrillation**.

Computers in Cardiology. 2001; **28**:133-135. | Article

20. Seet RC, Friedman PA and Rabinstein AA. **Prolonged rhythm monitoring for the detection of occult paroxysmal atrial fibrillation in ischemic stroke of unknown cause**. *Circulation*. 2011; **124**:477-86. | Article | PubMed
21. Singer D, Ziegler P, Schmitt S, Chang Y, Fan J, Than C and Turakhia M. **Paroxysmal Atrial Fibrillation Poses a Large but Transient Increase in Ischemic Stroke Risk: a Case-Crossover Study**. *Journal of The American College of Cardiology*. 2015; **65**:A315.
22. Sun R and Wang Y. **Predicting termination of atrial fibrillation based on the structure and quantification of the recurrence plot**. *Med Eng Phys*. 2008; **30**:1105-11. | Article | PubMed
23. Sun RR and Wang YY. **Predicting spontaneous termination of atrial fibrillation based on the RR interval**. *Proc Inst Mech Eng H*. 2009; **223**:713-26. | Article | PubMed
24. Zoni-Berisso M, Lercari F, Carazza T and Domenicucci S. **Epidemiology of atrial fibrillation: European perspective**. *Clin Epidemiol*. 2014; **6**:213-20. | Article | PubMed Abstract | PubMed FullText

Citation:
Akbilgic O, Howe JA and Davis RL. **Categorizing atrial fibrillation via Symbolic Pattern Recognition**.
J Med Stat Inform. 2016; **4**:8.
http://dx.doi.org/10.7243/2053-7662-4-8