Motif Based Feature Vectors: Towards a Homogeneous Data Representation for Cardiovascular Diseases Classification

Hanadi Aldosari[1], Frans Coenen[1], Gregory Y. H. Lip[2], and Yalin Zheng[3]

[1] Department of Computer Science, University of Liverpool, Liverpool, L69 3BX, UK. {H.A.Aldosari,Coenen}@liverpool.ac.uk
[2] Liverpool Centre for Cardiovascular Science, University of Liverpool and Liverpool Heart Chest Hospital, Liverpool, L69 7TX, UK. Gregory.Lip@liverpool.ac.uk
[3] Department of Eye and Vision Science, University of Liverpool, Liverpool, L7 8TX, UK. yalin.zheng@liverpool.ac.uk

Abstract. A process for generating a unifying motif-based homogeneous feature vector representation is described and evaluated. The motivation was to determine the viability of this representation as a unifying representation for heterogeneous data classification. The focus for the work was cardiovascular disease classification. The reported evaluation indicates that the proposed unifying representation is a viable one, producing better classification results than when a Recurrent Neural Network (RNNs) was applied to just ECG time series data.

Keywords: Motifs · Feature Extraction and Selection · Cardiovascular Disease Classification.

1 Introduction

Time series classification is a common machine learning application domain [5–7]. Using the computer processing power that is now frequently available the size of the time series we wish to process has become less of a challenge. However, for many time series applications, the data is presented in a range of formats, not just time series. One example is Cardiovascular Disease (CVD) classification, where typically the data comprises electrocardiogram (ECGs), echocardiograms (Echo) and tabular patient data. Deep learners, such as RNNs, do not readily lend themselves to such heterogeneous data. One solution is to train a number of classifiers, one directed at each data format, and then combine the classification results. However, this assumes that the data sources are independent. The alternative is to adopt a unifying representation so that a single classification model can be generated. The most appropriate unifying representation, it is argued here, is a feature vector representation because this is compatible with a wide range of classification models. The challenge is then how to extract appropriate features from data so as to construct the desired feature vector representation.

This paper explores the idea of generating a Homogeneous Feature Vector Representation (HFVR) from heterogeneous data by considering how we might

extract features from time series that can be included in such a representation; such as ECG time series. The idea promoted in this paper is the use of motifs as time series features that can be coupled with other features in a HFVR. The challenge here is the computational complexity of finding exact motifs [4]. Recently the use of matrix profiles has been proposed [8]. The mechanism proposed in this paper is founded on matrix profiles, namely the Correct Matrix Profile (CMP) technique described in [1]. A further challenge, once a set of motifs has been identified, is to select a subset of these motifs to be included in the final HFVR. The criteria here is the effectiveness of the generated classification model, we want to choose motifs (features) that are good discriminators of class.

To evaluate the utility of the proposed unifying approach a cardiovascular diseases classification application was considered; more specifically, the binary classification of Atrial Fibrillation (AF), the most common cardiac rhythm disorder. For the evaluation the China Physiological Signal Challenge 2018 (CPSC2018) data was used [3]. A SVM classification model was then generated using the proposed unifying representation and the performance compared with that of a classification models built using only the original time series data (an RNN was used). The proposed HFVR approach, that allows the inclusion of features from heterogeneous data sources, outperformed the time series only approach even though only a small amount of additional information was incorporated into the HFVR. It is anticipated that when further features from other data sources, such as Echo data, are added the proposed approach will significantly outperform classifiers built using only a single data source.

2 Application Domain and Formalism

Atrial Fibrillation (AF) can be identified from range of tests, but ECG analysis is considered to be the most reliable [2].An ECG indicates the electrical activity of the heart in terms of a summation wave that can be visualised and hence interpreted. A set of ECG records is of the form $\{R_1, R_2, \ldots\}$ where each record R_i comprises a set of time series $\{T_1, T_2, \ldots\}$ associated with a patient. The number of time series is usually 6 or 12 depending on whether six-lead or twelve-lead ECG has been used. For model training and evaluation we turn this data into an training/test data of the form $\mathbf{D} = \{\langle T_1, c_1\rangle, \langle T_2, c_2\rangle, \ldots \langle t_n, c_n\rangle\}$ where c is a class label drawn from a set of classes C. For the evaluation presented later in this paper a binary classification scenario is presented, thus $C = \{true, false\}$. Each time series T_j comprises a sequence of data values $[t_1, t_2, \ldots, t_n]$. A motif m is then a sub-sequence of a time series t_j. M is set of motifs extracted from $\mathbf{T}$, $M = \{m_1, m_2, ...\}$. Not all the motifs in M will be good discriminators of class so we prune M to give M' and then M'', the set of attributes for our feature vector representation. The input set for the classification model generation thus consists of a set of vectors $\{V_1, V_2, \ldots\}$ where each vector V_i comprises a set of occurrence count values m$\{v_1, v_2, \ldots\}$, and a class label $c \in C$, such that there is a one-to-one correspondence between the values and M''.

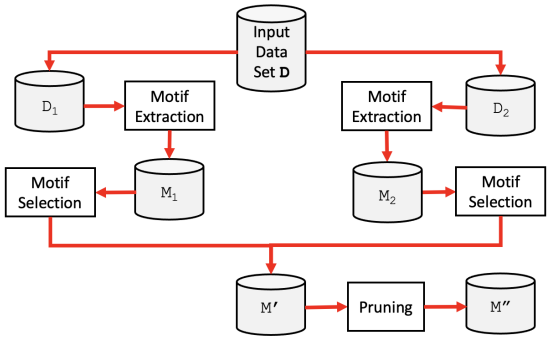

Fig. 1. Schematic of Motif-Based Feature Vector Generation Process.

3 Motif-Based Feature Vector Generation

This section presents the proposed motif feature extraction mechanism. A schematic describing the mechanism is given in Figure 1. The input is the set **D** which is divided into D_1 and D_2, where D_1 corresponds to class c_1 and D_2 corresponds to class c_2. These are then processed to extract motifs. The motif extraction will result in two sets of motifs M_1 and M_2 corresponding to D_1 and D_2 respectively. We wish to identify motifs that occur frequently throughout D_1 and D_2 as these are assumed to be good indicators of class. The next stage is thus motif selection. This produces two refined sets of motifs $M_1' \subset M_1$ and $M_2' \subset M_2$ which are combined to form M'. We wish to identify motifs that are good discriminators of class thus we want to remove motifs from M' that are associated with both c_1 and c_2. The result is a pruned set of motifs, M'' which is then used to generate our HFVR.

The adopted approach for motif discovery was founded on the Guided Motif Search (GMS) algorithm [1]. The basic idea behind GMS is to discover meaningful motifs that represent the user's expected outcomes. That is achieved by combining the Matrix Profile (MP) with an Annotation Vector (AV) to produce a Correct Matrix Profile (CMP). The AV consists of real-valued numbers between 0 and 1 and has the same length as MP; there is a one-to-one correspondence between AV and MP. A high AV value indicates the sub-sequence at position i is a desirable motif. Each element cmp_i in the CMP is calculated using Equation 1 where: (i) MP_i is the value in MP at position i, (ii) AV_i is the value from AV at position i, and (iii) $max(MP)$ denotes the maximum value in MP.

$$cmp_i = MP_i + ((1 - AV_i) \times max(MP)) \tag{1}$$

Thus if the $AV_i = 1$ the cmp_i value will be the same as the MP_i value, otherwise the cmp_i value will be the MP_i value increased by the $max(MP)$ value. The cmp_i value thus indicates whether position i contains a meaningful motif (to be selected), or not.

The motif discovery process identifies a set of motifs M_1 (M_2). However, we wish to retain motifs that are frequent across D_1 (D_2). There are a variety of ways

that this can be achieved. We could compare each motif $m_i \in M_1$ ($m_j \in M_2$) with each time series in D_1 (D_2) and record the frequency with which each motif occurs. However this would be computationally expensive. Instead we compare each motif with every other motif and determine the frequency of occurrence of each motif. Euclidean distance similarity was used for this purpose with a similarity threshold σ. If the distance between two motifs was less than σ they were deemed to match. For a motif to be selected it had to be similar to at least k other motifs or more. The result was a motif set M' containing selected motifs from both M_1 and M_2.

The set of motifs M' will hold motifs associated with either class c_1 or class c_2, and motifs associated with both classes c_1 and c_2. The motifs associated with both classes will not be good discriminators of class, hence these should be pruned. Thus, we compared each motif in class c_1 with every motif in class c_2. Euclidean distance with a σ similarity threshold was again used. If the Euclidean distance similarity was less than σ the motifs being compared were deemed to be representative of both class c_1 and c_2 and should therefore be excluded. The result is a pruned set of motifs, M'' which contains unique motifs with respect to classes c_1 and c_2 to be used in the desired HFVR.

The last stage in the proposed process was the feature vector generation stage. During this stage the set $V = \{V_1, V_2, \ldots\}$ was generated using the identified set of motifs M'' and any additional features we might wish to add (forb example age and gender). Each vector $V_i = \{v_1, v_2, \ldots, c\} \in V$ represents an ECG time series $T_i \in \mathbf{T}$. Each value v_j is either the numeric occurrence count of a motif m_j in the time series T_i, or the value associated with some additional feature. When classifying previously unseen records there will be no element c in V_i as this is the class value we wish to predict.

4 Evaluation

For the evaluation, The China Physiological Signal Challenge 2018 (CPSC2018) data set was used. For the proposed HFVR to be of value it needs to produce a classification accuracy that outperforms mechanisms founded on a single data source, such as only ECG time series. A SVM classification model was used with respect to the proposed HFVR. An RNN was used with respect to the time series only classification. The metrics used for the evaluation were accuracy, precision, recall and F1. Ten fold cross-validation was used throughout. The objectives of the evaluation were: (i) to identify the most appropriate value for σ, (ii) to identify the most appropriate value for k, and (iii) to determine the effectiveness of the proposed motif-based feature vector approach in comparison with an a deep learner applied directly to the input ECG time series.

To determine the most appropriate value for σ, a sequence of experiments was conducted using a range of values for σ from 0.05 to 0.50 increasing in steps of 0.05. The value of k used was 150 because preliminary experiments indicated this to be an appropriate value. The best recorded F1 value was 73.33%, obtained

using = 0.15. This was thus the value for σ used for the further experiments reported on in the following sub-sections.

To determine the most appropriate value for k, a further sequence of experiments was conducted using a range of values for k from 50 to 250 increasing in steps of 20. The results obtained indicated that the best F1 value of 75.05% was obtained using $k = 90$. This was then the k used with respect to the additional experiments reported on below.

To ascertain whether the proposed HFVR operated in an effective manner a SVM classification model was generated and tested using the proposed representation and compared to a RNN generated from just the ECG time series data. Experiments were conducted using the SVM with fixed parameters and with GridSearch to identify best parameters. For the RNN the SimpleRNN algorithm was used (available as part of the Keras open-source software Python library), where the output for each time stamp layer is fed to next time stamp layer. SimpleRNN was applied directly to the time series data. With respect to the proposed representation experiments were conducted using: (i) just motifs, (ii) motifs + gender, (iii) motifs + age, and (iv) motifs + gender + age.

	Accuracy	Precision	Recall	F1
proposed approached (motifs)	76.68%	72.96%	78.93%	74.32%
proposed approached (motifs+gender)	77.61%	72.86%	80.58%	75.71%
proposed approached (motifs+age)	85.09%	86.22%	84.71%	85.16%
proposed approached (motifs+gender+age)	**85.28%**	**86.26%**	**84.82%**	**85.28%**
SimpleRNN model	81.17%	85.30%	75.33%	79.00%

Table 1. Comparison of proposed approach using SVM without GridSearch and RNN.

The results using SVM are given in Tables 1 and 2; best results highlighted in bold font. From the Table it can firstly be observed that motifs when used on their own do not perform as well as when an RNN is applied to the time series data directly. Secondly it can be observed that the inclusion of gender to the feature vector representation does not make a significant difference, while including age does make a significant difference. Using motifs and age, or motifs, age and gender, produces a performance better than the time series only RNN performance. Age is the most significant factor here; it is well known that AF is more prevalent in older age groups. Thirdly, using GridSearch to identify best parameters for SVM classification also serves to enhance performance. Fourthly that best results are obtained when using GridSearch and a feature vector that includes motifs + age. Finally, from the results, an argument can be made that gender acts as a confounder, in that when GridSearch is used with feature vectors made up of motifs + age + gender the results are not as good as when using feature vectors made up of motifs + age.

5 Conclusion

A mechanism for generating a motif-based feature vector representation for use with CVD classification has beern presented. The motivation was that effective

	Accuracy	Precision	Recall	F1
proposed approached (motifs)	79.57%	78.51%	79.45%	78.01%
proposed approached (motifs+gender)	80.17%	77.52%	82.0%	79.22%
proposed approached (motifs+age)	**86.41%**	**87.27%**	**86.18%**	**86.37%**
proposed approached (motifs+gender+age)	85.49%	86.22%	85.26%	85.47%
SimpleRNN model	81.17%	85.30%	75.33%	79.00%

Table 2. Comparison of proposed approach using SVM with GridSearch and RNN.

CVD classification requires input from a number of sources (typically ECG, Echo and tabular data) and that a feature vector representation would provide a unifying mechanism for representing such heterogeneous data. The concept was evaluated by comparing its operation, using motifs paired with age and/or gender data, in the context of CVD classification, with that when using a RNN applied on to the ECG time series alone. The most appropriate values for σ and k were found to be 0.15 and 90 respectively. The classification results produced indicated that using the proposed representation combing motifs + age, or motifs + age + gender, produced a better classification than when using the time series on their own; thus indicating the benefits of the proposed unifying representation.

References

1. Dau, H.A., Keogh, E.: Matrix profile V: A generic technique to incorporate domain knowledge into motif discovery. In: Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 125–134 (2017)
2. Lip, G., Fauchier, L., Freedman, S., Van Gelder, I., Natale, A., Gianni, C., Nattel, S., Potpara, T., Rienstra, M., Tse, H., Lane, D.: Atrial fibrillation. Nat Rev Dis Primers **31**, 16016 (2016)
3. Liu, F., Liu, C., Zhao, L., Zhang, X., Wu, X., Xu, X., Liu, Y., Ma, C., Wei, S., He, Z., et al.: An open access database for evaluating the algorithms of electrocardiogram rhythm and morphology abnormality detection. Journal of Medical Imaging and Health Informatics **8**(7), 1368–1373 (2018)
4. Mueen, A., Keogh, E., Zhu, Q., Sydney Cash, S., Westover, B.: Exact discovery of time series motifs. In: SIAM International Conference on Data Mining. p. 473–484. SIAM (2009)
5. Oh, S.L., Ng, E.Y., San Tan, R., Acharya, U.R.: Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. Computers in Biology and Medicine **102**, 278–287 (2018)
6. Pourbabaee, B., Roshtkhari, M.J., Khorasani, K.: Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. IEEE Transactions on Systems, Man, and Cybernetics: Systems **48**(12), 2095–2104 (2018)
7. Wang, G., Zhang, C., Liu, Y., Yang, H., Fu, D., Wang, H., Zhang, P.: A global and updatable ECG beat classification system based on recurrent neural networks and active learning. Information Sciences **501**, 523–542 (2019)
8. Yeh, C.C.M., Zhu, Y., Ulanova, L., Begum, N., Ding, Y., Dau, H.A., Silva, D.F., Mueen, A., Keogh, E.: Matrix profile I: all pairs similarity joins for time series: a unifying view that includes motifs, discords and shapelets. In: IEEE 16th International Conference on Data Mining (ICDM). pp. 1317–1322. IEEE (2016)