HW2 : Motif Discovery

Bhavika Tekwani (btekwani@gmu.edu)

1. Classification Accuracy

I am referencing the DTW results from HW1 but I must mention that these are not to be considered accurate or useful in terms of comparing my motif discovery and classification implementation because I had made the mistake of not stripping test labels in 1NN with euclidean distance and DTW which is what resulted in the high accuracies below.

Dataset	1 NN with euclidean distance	1NN with Dynamic Time Warping	KNN and Dynamic Time Warping with 20% warping window	KNN with LB_Keogh Dynamic Time Warping	**Motif Discovery and Classification using bag of patterns**
1	85.22	100	31.22	85.22	**32.77**
2	78.28				**13.72**
3	51.65				**13.63**
4	78.88	98.88	8.64		**10.54**
5	99.54				**10.78**

2. Approach

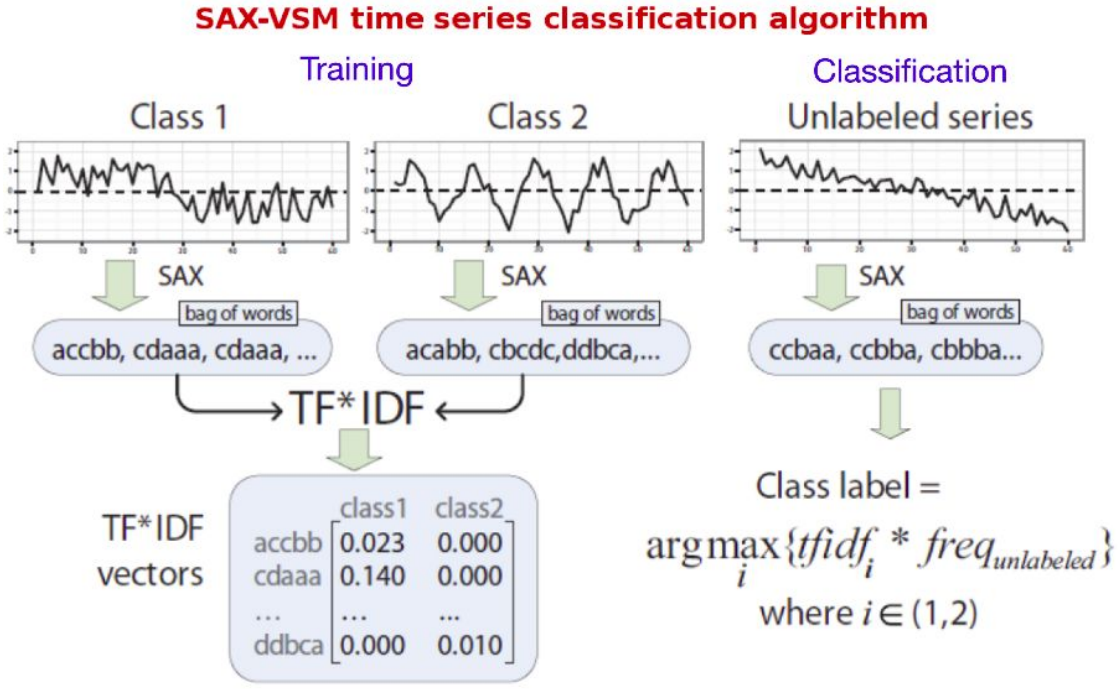
For this problem, I used two R libraries for motif discovery and eventually for classification. For motif discovery, I used TSMining [1] and for TFIDF with SAX representations, I used jMotif-R [2].

For motif discovery, TSMining implements the exact discovery algorithm outlined in Probabilistic Discovery of Time Series Motifs[3]. Based on this algorithm, we extract motifs from each train data point in each of our datasets separately. At this time, we have to make some decisions about the parameters to be used for motif discovery. I have set window size to 5, word size and alphabet size to 5. Also, global and local standardization is run on the time series and each subsequence. Overlap is set to 0 whereas epsilon or the threshold for variance is 0.01. We must note that changing the window size, epsilon and alphabet size will drastically change the number of motifs we discover.

We are asked the question, how do we use motifs as features? The approach I have taken involves converting motifs from a numeric representation to a SAX representation - the alphabet size and epsilon is used to define this SAX representation. Now, depending on the the time series, the number of motifs discovered may vary. Instead of setting a limit on how many motifs should be found, I find all possible motifs and then decide to weigh which ones are most likely for certain labels.

We do this using TF-IDF (Term Frequency Inverse Document Frequency) weighting. One can think of the SAX representation of time series as words. This leads us to the bag of patterns approach similar to the bag of words model used in text mining. We count the occurrences of each pattern (SAX representation) for specific classes in the train set and create a TF-IDF representation of the training set as a whole. This TF-IDF representation is created using the *bags_to_tfidf* method from jMotif[2]. Now, we convert each time series from the test set to a SAX representation using the same parameters we used for the train set (window size, alphabet size, epsilon, etc). Again, we use the exact discovery algorithm. Now, similar to the bag of words model, we try to find the similarity between the test data point's TF-IDF weight and each bag. We use cosine similarity. Note that we are using a similarity measure instead of a distance measure, which means the similarity has to be maximized to assign the label to the test time series.

The overall process is similar to the SAX-VSM classification shown in this illustration [6] below. In our implementation, we first find the motifs using TSMining and then convert them to a SAX representation before TF-IDF.



To address the questions asked in this assignment, a motif is discriminant based on the domain of the problem but by definition, it is of a certain length and does not overlap with other patterns in the time series. How many motifs we need can be determined by using cross-validation or a grid search over the parameters used for motif discovery (window size, alphabet size, etc). At the very least, I think the number of motifs found should be more than one and the motif we're matching it to should be within a certain distance *d* to serve as a basis for classification.

Since I have used an exact motif discovery algorithm from TSMining, no post-processing was required. We are using the bag-of-words classification technique but several other approaches use SVMs & Bayesian classifiers [4] and [5].

An example on SAX-VSM classification from [2] has been adapted and used in our R program in the classify() function.

References

[1] Cheng Fan , Mining Univariate and Multivariate Motifs in Time-Series Data, https://cran.r-project.org/web/packages/TSMining/TSMining.pdf

[2] Pavel Senin, jMotif-R, https://github.com/jMotif/jmotif-R

[3] Bill Chiu, Eamonn Keogh, and Stefano Lonardi. 2003. Probabilistic discovery of time series motifs. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* (KDD '03). ACM, New York, NY, USA, 493-498. DOI=http://dx.doi.org/10.1145/956750.956808

[4] Maletzke, André Gustavo, Huei Diana Lee, Gustavo Enrique, Almeida Prado Alves Batista, Cláudio Saddy Rodrigues Coy, João José Fagundes, and Wu Feng Chung. "Time Series Classification with Motifs and Characteristics." In *Soft Computing for Business Intelligence*, edited by Rafael Espin, Rafael Bello Pérez, Angel Cobo, Jorge Marx, and Ariel Racet Valdés, 125–38. Berlin, Heidelberg: Springer Berlin Heidelberg, 2014. http://dx.doi.org/10.1007/978-3-642-53737-0_8.

[5] Buza, Krisztian, and Lars Schmidt-Thieme. "Motif-Based Classification of Time Series with Bayesian Networks and Svms." In *Advances in Data Analysis, Data Handling and Business Intelligence*, 105–114. Springer, 2009. http://link.springer.com/chapter/10.1007/978-3-642-01044-6_9.

[6] Pavel Senin, SAX-VSM, https://github.com/jMotif/sax-vsm_classic

[7] Mueen, Abdullah, Eamonn Keogh, Qiang Zhu, Sydney Cash, and Brandon Westover. "Exact Discovery of Time Series Motifs." In *Proceedings of the 2009 SIAM International Conference on Data Mining*, 473–484. SIAM, 2009. http://epubs.siam.org/doi/abs/10.1137/1.9781611972795.41.