COMP5331 Presentation Script

DING Mu Cong

November 21, 2017

1 Slide 1: Opening

Hello everyone, we are group 13 and I am Ding Mucong, she is Guo Wenshuo and he is Cheung Tsz Him. Our presentation today is on the implementations and applications of Toeplitz Inverse Covariance-Based Clustering, which is abbreviated as TICC.

2 Slide 2: Outline

In the following 15 minutes, we are going to have a brief introduction on our work done on implementation, model comparison and hyper parameter tunning. Apart from that, we will give a brief introduction to the TICC paper and the algorithm, and have a final discussion and conclusion.

3 Slide 3: Time Series Clustering

The problem we are looking into is the subsequence clustering problem, which could be defined as discovering the repeated patterns in temporal data. Compared to the classical clustering problem, subsequence clustering on time series has to respect the temporal structure of a sequence, so that it cannot treat each time unit independently, instead, it requires simultaneous segmentation and clustering of the whole series. The time-series given are usually multi-variate, which can be understand as, multiple time-series sharing the same latent patterns. And our goal is to discover this latent pattern, by means of unsupervised learning. The real world applications of this problem are abundant. For example, we can analyze the sensor data recorded by wearable devices like smart watch, phone, and even shoes, and identify the state of motion of the people in real time.

4 Slide 4: Challenges and Related Work

However, this subsequence clustering problem is challenging, as it performs clustering and segmentation at the same time. To design an appropriate algorithm

for this problem, we first need to consider which metric we are basing on to perform this clustering. The most classic approach is distance bashed, like KMeans, however it treats the data at different time independently, each as a new vector, and does not consider the temporal structure of a series at all. To overcome this, some people pointed out the shaped-based metric, which measures the similarity between two temporal sequences even if they varying in speed. However, with algorithm based on shape similarity like Dynamic Time Wrapping only, we can only clustering the complete sequence but not split it into subsequence simultaneously.

5 Slide 5: Introducing TICC

To achieve the subsequence clustering, we probably need a brand new metric that describing the similarity of sub-sequences of arbitrary length, instead of the the full sequence of a single time unit. The David Hallac et al of Stanford university investigated on this difficulty in depth and found a new algorithm called Toeplitz inverse covariance-based clustering. They define each cluster by a correlation network, or more specifically a Markov random field, which characterizes the interdependencies between different variables in a typical pattern, and defined the metric by the similarity of that correlation networks. The solve the optimization problem in this setup by alternating minimization, as a variation of the EM algorithm. The TICC algorithm can perform simultaneous segmentation and clustering, and works well on multivariate time series. Following is an illustration of TICC's mechanism where we can see that subsequences in the same cluster have similar covariance relations.

Our project is to discover the applications and improvements of TICC algorithm. So let me pass my turn to Wenshuo, as she will introduce the work we did on model comparison.

6 Slide 15: RNN using LSTM

Ok, my turn again. As the last part of model comparison, we want to compare the TICC algorithm, as a unsupervised learning model, to the sequence classification algorithm like recurrent neural network with long short term memory cells. This comparison is unfair, as the latter one is supervised learning and has the access to training labels. However, when we put TICC into real applications, often, the data we are facing usually has labels. Like the case of wearable sensor data, we can easily record the activity states. But however, it is hard to use the Toeplitz inverse covariances metric to classification models. In a word, TICC has to beat some powerful sequence classification algorithm to prove its real world value in subsequence clustering problem. The sequence classification model we choosed it

7 Slide 16: Comparison on Gesture Data

8 Slide 17: Comparison on Human Activity Data