

Clustering and Visualization of Temperature Time Series

2006 Spring CS235 Project
Danhua Guo
Jianfeng Yang
University of California, Riverside
{dguo, jyang}@cs.ucr.edu

Abstract

In this project, we concentrate on the problem of temperature time series clustering with implementations of some relevant techniques to help us analyze the data. The goal is to find an easy and effective way to measure the similarity between different temperature time series. We use SAX to represent time series. The Hierarchical Clustering algorithm is applied with dendrogram and icon as a way of assisting visualization. We claim that hierarchical algorithm with SAX and Euclidean distance on symbols frequency can cluster time series easily and effectively and that dendrogram and icon can show the clustering result in a more straightforward way. We illustrate the power of our approach by achieving a “good” clustering on a real temperature time series dataset.

1. Introduction

With the increasing requirements for information inquiry, people are no longer satisfied with result that answers their question exactly. Instead, the more relevant information provided the better search engine it is believed to be. Such concern has been widely considered for online services. At Amazon or other online shopping website, customers will be recommended with relevant products based on the single piece she/he purchased.

Grouping cities by temperatures is valuable for people to choose a place for travel or job. For instance, if one person likes the temperature of Los Angeles, it is highly possible that he would like the temperature of San Diego.

In this project, we focus on finding an easy way to demonstrate the temperature difference between cities. We use time series to identify yearly day time temperatures in different US cities. Unlike previous classic clustering approaches, such as Euclidean distance, DTW and sequence alignment, we did our project with a more effective and clear way.

In this work, we do it in an easier way. First we use SAX to identify 16 features on the input temperature time series. Then we employ Euclidean Distance on the frequency of symbol, which eliminates the effect of phase shift. Finally, we use group average linkage to generate a dendrogram of all the cities. Meanwhile, we generate intelligent icons for each temperature time series. This icons show the temperature distributions in a more straightforward way. Our experiments results on a real dataset containing 157 cities temperature time series show that our approach is very effective.

The rest of the paper is organized as follows. In section 2, we briefly review the prior work in this field. In section 3, more details about the experiments is provided. We show the experiments results in section 4. And the conclusion is in section 5.

2. Related Work

Clustering is a traditional research field in data mining. Given the dataset which contains unlabeled data, a well-learned model should be able to tell the similarity of these data. In time series clustering problem, we are trying to find some “good” definition of similarity and using them to measure the distance among the data.

Many clustering criteria/methods have been proposed for time series clustering. Euclidean distance is the most popular similarity measure. But it is a bit brittle. DTW allows a more elastic shifting of the time axis, which can match similar sequences out of phase. Sequence alignment must be applied on discrete data.

While there have been a great variety of previous work, our work is meaningful in that we apply Euclidean measure on symbols frequency from SAX results and that we use icon to visualize dendrogram in a straightforward way.

3. Experiment

The experiment outline is shown in Fig.1.

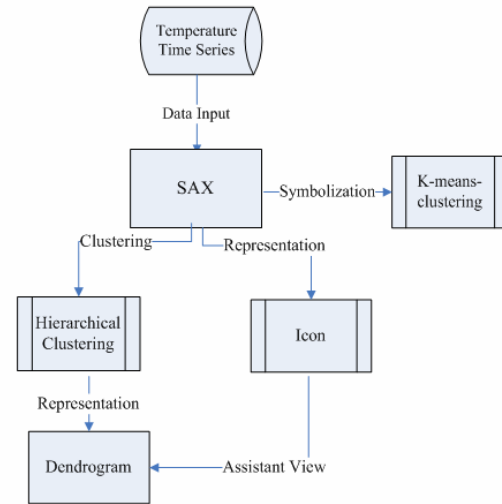


Figure 1 Experiment outline

3.1 Data Set

The data set (Fig. 2) is the highest temperature in 157 US cities, from 1995.1 to 2002.12. Even though we only have access to the current data set, it is still acceptable because most outdoor activities are carried during daytime. Hence the highest temperature affects people more than the lowest temperature or the average.

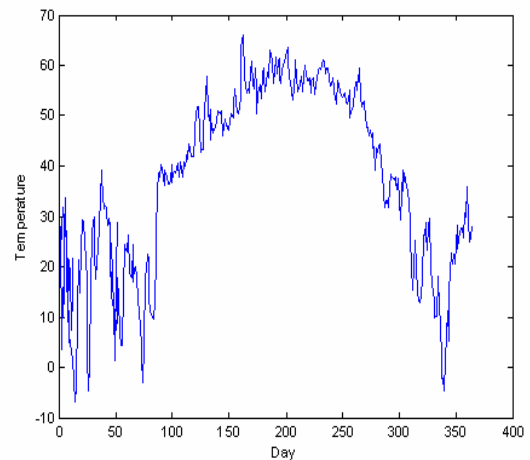


Figure 2 One example (1 out of 157*8) from the data set

3.2 SAX

We use k-means to get the “best” K quantization intervals for SAX (K = 16, K is selected for better SAX and Icon representation). Using k-means, the temperature range with higher frequency will be clustered into more groups and therefore get more accurate results (Fig. 3).

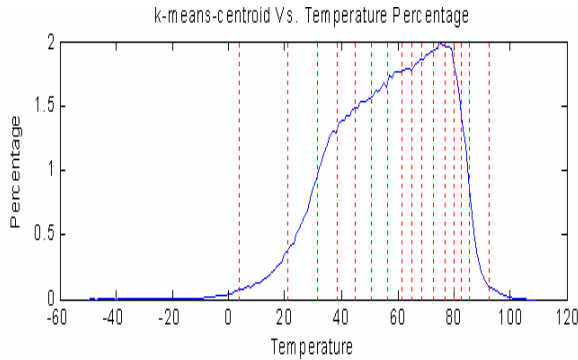


Figure 3 Temperature frequency and quantization intervals of SAX

After the step of SAX, each time series is converted into a symbol string (Fig. 4) with up to 16 different letters, each representing one of the 16 features of the time series. In our work, we use “ab...op” as the symbol set.

```
dfffffgfgmkjkmggdmdggffddmegff
mpeggffffggifgedgdmmpmdeem
gfgmepffdmpkkjpkkjppjkeem
mekpmkjllcjeepklljjkpeppepeee
epkpkllckkcbbcncpkjlcncclljjcnc
cbhnlkjllccccccnccncllccjlnnnb
hahhhnbbchcnhcnhchhhncnbnh
nhhhbhbbbbhbhbhbhhhhncllcnc
jjlcncncclljeepkjjllcckljjppjjllkppjj
jpepkpmpkeepkjddgdeddemgdggd
```

Figure 4 Symbols string

3.3 Icon generation

We calculate the frequency of different symbols (Fig. 5).

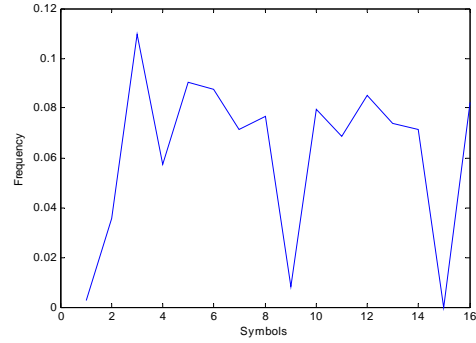


Figure 5 Symbols frequency

Then we generate icons according to the frequency distribution. Fig. 6 is the mapping from frequency to color. The frequency in Fig. 6 is from low (left) to high (right). And Fig. 7 is one icon example (generated from Fig. 5).

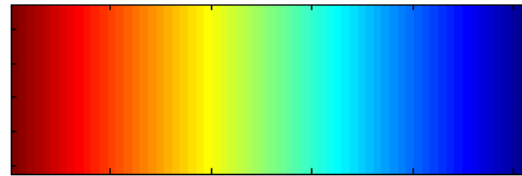


Figure 6 Mapping from frequency to color

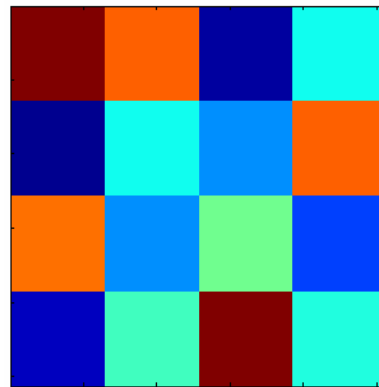


Fig. 7 One icon example

3.4 Hierarchical clustering

In Hierarchical Clustering, we use Group average linkage and Euclidean Distance for Distance Metric. In Fig. 8, we show one example dendrogram of 5 cities. In it, the number of x axle demonstrates

different temperature time series. 1 to 8 means the temperature time series of 8 year from city 1. 9 to 16 represent city 2 and so on.

Here we tried two linkage criteria, average linkage and single linkage. They do not show much difference. But we prefer average linkage. Since the temperature may change much in some years, average linkage is more meaningful.

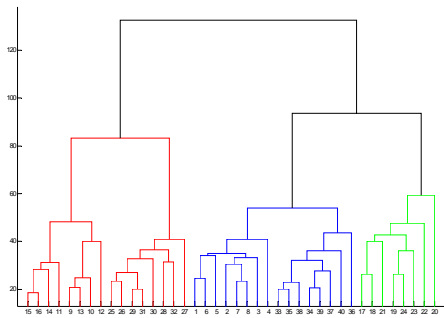


Figure 8 One dendrogram example of 5 cities

4. Result

Fig. 11 is the result of clusters we get for the complete data set of temperature time series. As stated above, we work on 8-year temperature data from 157 American cities. Since the result is too vague for such a huge number of input data, we select some part of it and shows the features in more detail.

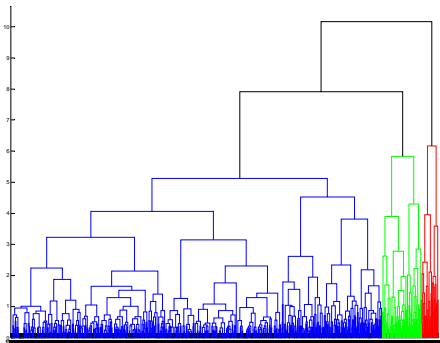


Figure 11 Dendrogram of the whole data set (157 cities, 8 time series for each city)

Fig. 9 is a sample of icon representations of temperature time series. People cannot always tell the difference between time series. Icon, in this case, is of great help. From Fig. 9, we can easily see that the low and high temperature frequency in the three time series are quite different, which can help us to make a decision without opening the files.

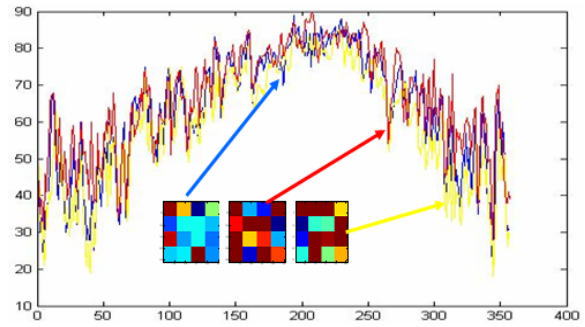


Figure 9 Icon representations of time series

Fig. 10 is a dendrogram of 3 far away cities. From it we can see that time series with more similarity also have more similarity in their icon representations. It has two-tier meanings 1) that icon can help us to easily cluster temperature time series and 2) that icon representation of time series is useful.

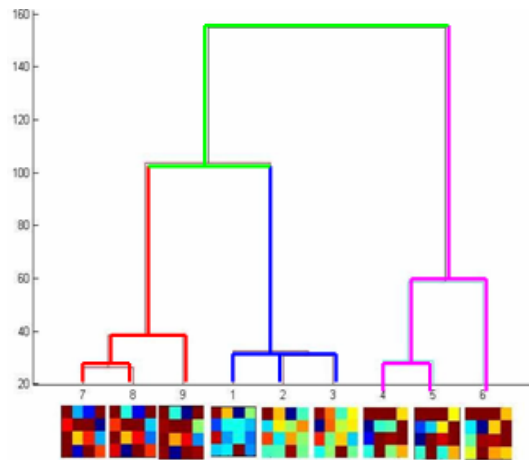


Figure 10 Dendrogram with icon assistance

5. Conclusion

SAX is an exceptional time series representation. With SAX, we can easily do time series clustering. Also, Icon can help visualize files without opening them.

In this specific temperature time series clustering problem, SAX and icon also works very well. When we put icon and dendrogram together, the similarity of the temperature of different cities are shown in a straightforward way. With our approach, people without data mining knowledge can see the temperature of which city they like.

Acknowledgments

Thanks to Professor Eamonn Keogh for the idea of SAX and Icon.