Time Series Taxonomy

Javier Moreno

University of Granada
jmorenov@correo.ugr.es

Abstract

Given a time serie, we want determinate wich classifier is the best for it. Can data complexity give us a hint?, the answer is yes and with complexity measures we can. We can do this by measures of complexity, a lot of time series and performing a clustering. This paper attempts to give an idea of how to do this, also i will try to introduce a list of complexity measures the biggest i can.

1. Introduction

This paper explains a approach to create a taxonomy of time series, using complexity measures implemented 4, a Time Series Database and the classification of these.

The paper is divided into three parts:

- How to generate the taxonomy 2.
- Time Series Classification/Clustering 3
- Complexity Measures 4

2. How to build the taxonomy

The main elements for construction are measures of complexity 4 and the Time Series Database, the more we have time series is better to give us a greater perspective when classifying.

With complexity measures implemented in our case R language, the first is applied to each time series all measures of complexity, and storing these results. These data can now make a better classification of time series, obtaining the taxonomy, for this a clustering is performed (not supervised or directed by expert) with all time series and these data. With this taxonomy we would be prepared to classify new series with great success rate.

The idea is to reduce the number of prediction methods applied to each time series methods to limit certain specific classes.

3. Time Series Classification/Clustering

Time series clustering is to partition time series data into groups based on similarity or distance, so that time series in the same cluster are similar. For time series clustering with R, the first step is to work out an appropriate distance/similarity metric, and then, at the second step, use existing clustering techniques, such as k-means, hierarchical clustering, density-based clustering or subspace clustering, to find clustering structures.
Time series classification is to build a classification model based on labelled time series and then use the model to predict the label of unlabelled time series. The way for time series classification with R is to extract and build features from time series data first, and then apply existing classification techniques, such as SVM, k-NN, neural networks, regression and decision trees, to the feature set. For more information see [3]

4. Complexity Measures

4.1. Kolmogorov complexity

The Kolmogorov complexity K(x) of an object x is the length, in bits, of the smallest program (in bits) that when run on a Universal Turing Machine (U) outputs K(x) and then stops with the execution. This measure was independently developed by Andrey N. Kolmogorov in the

late 1960s. On the basis of KolmogorovâĂŹs idea, Lempel and Ziv developed an algorithm (LZA), which is often used in assessing the randomness of finite sequences as a measure of its disorder.
The Kolmogorov complexity of a time series x_i, i=1,2,3,4...,N can be summarized as follows:
Step 1: Encode the sime series by constructing a sequence s consisting of the characters 0 and 1 written as s(i), i=1,2,3,4,...,N, according to the rule:

$$s(i) = \begin{cases} 0 & x_i < x_* \\ 1 & x_i \geq x_* \end{cases}$$

Where x_* is a threshold, the mean value of the time series has often been used as the threshold.
Step 2: Calculate the complexity counter C(N), which is defined as the minimum number of distinct patterns contained in a given character sequence.
More information in [1] [4] [2].

4.1.1. Lempel-Ziv Complexity

The Lempel-Ziv Complexity is based on Kolmogorov Complexity. First is calculated the Kolmogorov Complexity K(x) of an object x and then calculate the normalized complexity measure $C_k(N)$, which is defined as

$$C_k(N) = \frac{c(N)}{b(N)} = c(N)\frac{log_2 N}{N} \quad (1)$$

More information in [2].

4.2. Entropy

4.2.1. Aproximation Entropy

Aproximation Entropy (ApEn) is a complexity measure used to quantify the amount of regularity and the unpredictability of fluctuations over time-series data.
The presence of repetitive patterns of fluctuation in a time series renders it more predictable than a time series in which such patterns are absent. ApEn reflects the likelihood that similar patterns of observations will not be followed by additional similar observations. A time series containing many repetitive patterns has a relatively small ApEn; a less predictable process has a higher ApEn.

4.2.2. Sample Entropy

Sample entropy (SampEn) is a modification of approximate entropy, used extensively for assessing the complexity of a physiological time-series signal, thereby diagnosing diseased state. Like approximate entropy (ApEn), Sample entropy (SampEn) is a measure of complexity. But it does not include self-similar patterns as ApEn does.

4.2.3. Permutation Entropy

Input: Given a stationary time series xt
We have to determine length of sliding window n and slide the time series according to it, calculate permutation entropy h_n of order n, repeat for some n and finally calculate h_n, where:

$$H(n) = -\sum p(\pi) log p(\pi)$$

$$h_n = \frac{H(n)}{(n-1)}$$

References

[1] Nikolic-Doric D.T.Mihailovic, .Mimic and I.Arsenic. Novel measures based on the kolmogorov complexity for use in complex system behavior studies and time series analysis. `http://arxiv.org/pdf/1310.1304v1.pdf`. [Online; accessed 20-April-2015].

[2] Da-Guan Ke and Qin-Ye Tong. Easily adaptable complexity measure for finite time series. `http://arxiv.org/pdf/nlin/0505052.pdf`. [Online; accessed 9-March-2015].

[3] RDM. Time series clustering and classification. `http://www.rdatamining.com/examples/time-series-clustering-classification`. [Online; accessed 6-April-2015].

[4] Wikipedia. Kolmogorov complexity — wikipedia, the free encyclopedia. http://en.wikipedia.org/w/index.php?title=Kolmogorov_complexity&oldid=646881276, 2015. [Online; accessed 10-March-2015].