Сатыбалдиев Д.

БӨЛҮК-СЫЗЫКТУУ ЖАКЫНДАШТЫРУУНУН ЖАРДАМЫ МЕНЕН СТАЦИОНАРДЫК ЭМЕС УБАКЫТ СЕРИЯЛАРЫНЫН ОКШОШТУГУН АНЫКТОО: ЭМПИРИКАЛЫК МИСАЛ

Сатыбалдиев Д.

ОЦЕНКА СХОДСТВА НЕСТАЦИОНАРНЫХ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ КУСОЧНО-ЛИНЕЙНОЙ АППРОКСИМАЦИИ: ЭМПИРИЧЕСКИЙ ПРИМЕР

D. Satybaldiev

NON STATIONARY TIME SERIES SIMILARITY MEASUREMENT BASED ON PIECEWISE LINEAR APPROXIMATION: EMPIRICAL EXAMPLE

УДК: 530.18/621.313

Акыркы жылдары, убакыт сериясы анализи боюнча пайыздык өсүп жатат. Илимий байкоолордон алынган дээрлик бардык маалыматтарды ырааттуу өлчөө убакыттын өтүшү менен жүзөгө ашырылат. Убакыт-катар анализине негизги максаты базасы баалуулуктарды алуу болуп саналат. Биз убакыт катар окшоштуктарын тез жана натыйжалуу баалоо боюнча жаңы ыкмасын иштеп чыктык. Биздин ыкма техниканын, бөлүк сызыктуу жакындоого негизделген. Биз ошондой эле бул ыкма кыйла убакыт серия окшоштугуна баа берүү жараянын тездетүүгө боло турган эксперименталдык натыйжаларды көрсөтөт.

Негизги сөздөр: *убакыт сериясы, бөлүк сызыктуу жакындоого, f-критерия, убакыт серияларынын окшоштугу.*

В последнее время интерес к анализу временных рядов растет. Почти все данные, получаемые от научных наблюдений, последовательные измерения выполняются в течение времени. Основная цель анализа временных рядов является извлечение значений баз данных. Мы разработали новый метод для быстрой и эффективной оценки сходства временных рядов. Наш метод основан на известной технике кусочно-линейного апроксимирования. Мы также показали экспериментальные результаты, что наш метод может заметно ускорить процесс оценки сходства временных рядов.

Ключевые слова: *временные ряды, кусочно-линейная аппроксимация, f-критерия, оценка сходства временных рядов.*

Recently, interest in time series data mining is exponentially increasing. Almost all the data obtained from the scientific observations, successive measurements are performed over a time. The main purpose of the time series data mining is to extract all meaningful knowledge from the shape of the data. Time series similarity measurement is one of the main parts of time series data mining which is used in variety of the fields such as forecasting, clustering and etc. We propose a new method for fast and efficient similarity search of time series. Our method based on well known Piecewise Linearity Approximation technique. We provide experimental results that our method can noticeably speed up the process of measuring similarity of time series.

Key words: *time series, piecewise linear approximation, f-criteria, similarity measurement.*

Introduction

Recently, one of the most important issues in data analysis has been extracting information from large data warehouses. This process of data discovery from data warehouses is called data mining and is successfully and commonly being used in many business and research applications. Research target can focus on variety of objects: texts, graphs, multimedia, web data etc. One of objects is time series. Why time series? Currently world supply of data is delivered substantially in the form of time series. Time series contain large volumes of vital information about various phenomena of our life from physics, economics, finance etc. So, the extraction of desired time series from a large information database is one of the main tasks of Time Series Data Mining (TSDM). For instance, having some dynamic stock and larger time series data set, from which the most similar time series graph should be depicted. Therefore proper algorithm is needed to be implemented in order to be able to discover the most similar time series among thousands. This type of example can be extended into much more complex statistics and other diverse fields.

A variety of similarity measure techniques exist for measuring similarity between time series. Basically known methods are Euclidean Similarity Measure, Dynamic Time Warping, The Longest Common Subsequence Measure, and Piecewise Linear Representation of Time Series. These techniques and not mentioned ones have their advantages and disadvantages when compared with each other. A general overview of differences between the approaches will be presented. In this work, we paid our attention to the method of Piecewise Linear Representation of time series. New approach to PLR is being developed. The approach has new and interesting mathematical basis. Comparison criteria elaborated on the basis of this approach also will be new. Two types of algorithms were developed with help of MATLAB software, which take time series as input from data set and produce a piecewise representation for every single time series, as segmentation algorithms. That, in turn, will be used for the purpose of similarity measure. This technique can be

used in variety of the fields for solving such problems as provided in the article by Agrawal et al. 1993;

- Identify companies with similar pattern of growth.
- Determine products with similar selling patterns.
- Discover stocks with similar movement in stock prices.
- Finding is a musical score is similar to one of the copyrighted scores.

Piecewise Linear Representation for Fast Similarity Search

For purpose of measuring similarity between two time series sequence in case when they are within measurable distance all types of techniques such as Euclidean distance or dynamic time warping can be used. But in our case when we are working with long time series and large database of scientific data more general representation of the data is required. We need fast and efficient process of comparing time series. That is why we are not interested in actual values that occurring in the sequences. The actual object of observation and measurement is the approximated sequence which involves less data but contains all the significant features.

A piecewise linear approximation is one method of constructing a function $g(\mathrm{x})$ that fits a nonlinear objective function $f(\mathrm{x})$ by adding extra binary variables, continuous variables, and constraints to reformulate the original problem. The specific goal is to approximate a single valued function of one variable in terms of a sequence of linear segments. For the function $f(\mathrm{x})$, defined on the interval [a, b], a piecewise linear approximation will approximate a function $g(\mathrm{x})$ over the same interval. Where $g(\mathrm{x})$ is to be made up of a sequence of linear segments. Then $g(\mathrm{x})$ is in the form $g(\mathrm{x}) = \mathrm{c} + \mathrm{dx}$ for every x in [a,b] as shown in Cameron, 1966. A commonly known example of a piecewise function is $f(x) = |x|$ where

$$|x| = \begin{cases} x, \text{ if } x \geq 0 \\ -x, \text{ if } x \leq 0 \end{cases}$$

The purpose of doing a piecewise linear approximation is that the new linearity will allow the previously nonlinear problem to be solved by linear programming methods, which are much easier to employ than their nonlinear counterparts.

In figure 1, the exchange rate of European Euro to US Dollar is shown. Time series of exchange rate is represented by linear segment using Piecewise Linear Representation.

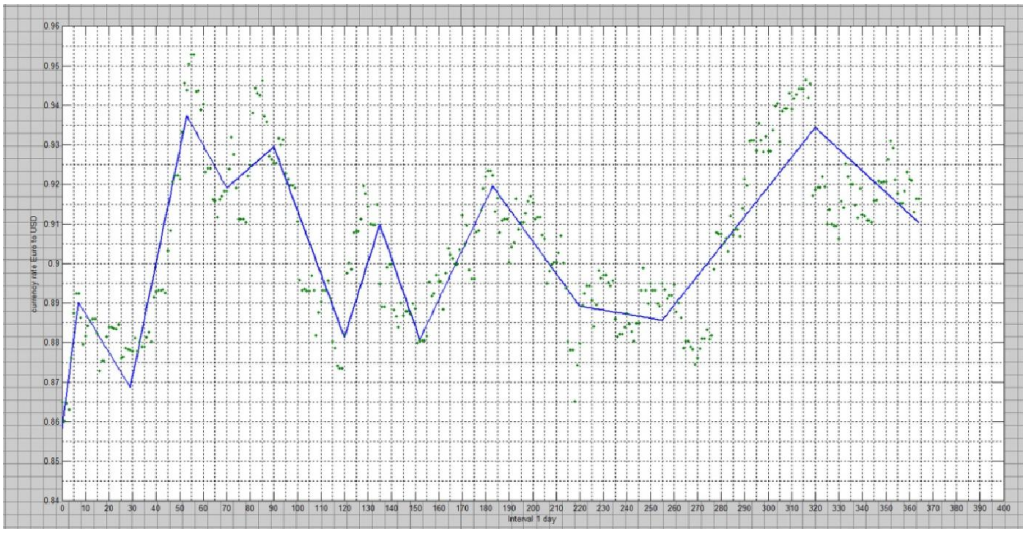

Figure 1. Currency rate of Euro to USD and its Piecewise Linear Representation.

Proposed Technique

Time series $x(t_i)$, where $(i = 1, 2, ..., n)$. This time series has local linear trends, and presented by n sample taken at equal time intervals. $\Delta t = \frac{T}{n}(t_i = t_{i-1} + \Delta t)$, where T is the time of observations. It is required to find a piecewise linear approximation for this time series. For this purpose more general case of piecewise polynomial approximation is needed. System of m polynomials of an independent variable $t \geq 0$.

$$P_i\left(\alpha_1^i, \alpha_2^i, ..., \alpha_{r_i}^i, \tau_i, t\right), (i = 1, 2, ..., \mathrm{m}) \qquad (1)$$

We simplify the notation of function (1) and write $P_i(t)$ implying that polynomial $P_i(t)$, in addition to the independent variable t, it also depends on r_i parameters and terms τ_i which determines the upper limit of its pivot points. Milnikov et al. 2015. As a result after simplification we can introduce new function (2)

$$F(t) = \sum_{i=1}^{m} P_i(t) \qquad (2)$$

The function $F(t)$ is finite on its pivot points $[0, T]$ and it is called approximation aggregate and polynomials (1) are this aggregates components.

Piecewise structure of the Aggregate is determined by the fact that certain components of polynomials are finite, and their initial and final pivot points $I_i = [0, \tau_i]$ form a nested sequence of non-decreasing intervals $I_i \subseteq I_{i+1}$. So, formally recorded function of Aggregate in the form of (2), can be represented as the function (3).

$$F(t) = \sum_{i=k}^{m} P_i(t), \; for \; t \geq \tau_{k-1} \qquad (3)$$

For Piecewise Linear Approximation of initial time series $x(t_i)$ the problem of construction of a system of splines, determined by means of polynomial components. The special case of constructing finite functions of approximating aggregate is piecewise linear approximation where all polynomial components represented by straight lines. Approximation Aggregate (4)

$$F(t) = \sum_{i=i_0}^{m} (d_i - a_i t) \qquad (4)$$

where i_0 is the value of index that is equal to index of the τ_i for which condition for which $\min_{\tau_i}((\tau_i - t) \geq 0)$ is true. a_i is a slope of the aggregate's straight line equation, that is a component of a first degree polynomial; d_i - intercept of polynomial components.

Let's take n observed pivot points of given time series $x(t_i), (i = 1, 2, ..., n)$. $\tau_i \, (i = 1, 2, ..., m)$ are the grid nodes at interval $(0, \tau_{\max})$. It is required to determine such values of d_i which minimize the i functional of local linear approximations of initial time series. System of linear equations was taken from Milnikov, Sayfulin 2012.

Experimental Results

Before proceeding to similarity measure of time series given data source should be approximated by Piecewise Linear approximation. We process time series data represented by the means of 366 observations. As an example we took 1 year exchange rate of Kyrgyz som to United States dollar, particularly from 1st December of 2014 to 1st December of 2015. In figure 2 we provide time series of this 1 year observation. For obtaining Piecewise Linear Approximation segments our new algorithm requires predefined pivot points. We defined pivot points for this example with the step 20 points between them.

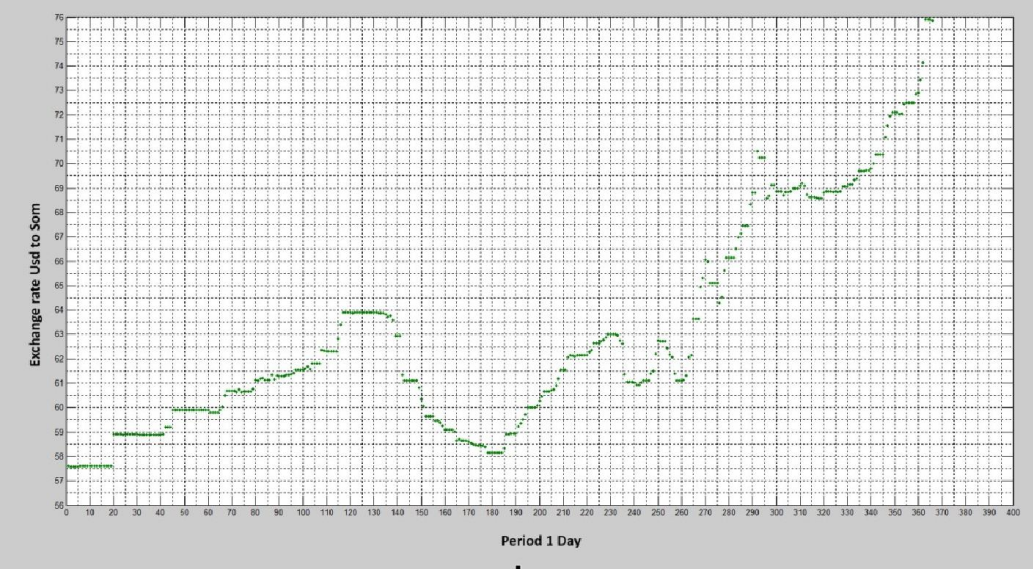

Figure 2. Exchange rate of Kyrgyzstan som to United States Dollar before piecewise linear approximation. Exchange rate was taken from www.nbkr.kg for 1 year from 1st December of 2014 to 1st December of 2015.

```
z = x./xm;
d = [20 40 60 70 80 100 120 130 140 160 180 200 220 240 260 280 300 320 340 xm];
zd=d'./xm;
```

It should be taken to the note that, visually these points can be defined more precise seeing real pivot points of time series. But we are trying to make realistic algorithm of how machines work for the similarity testing of various time series. Machines take vague points and step by step eliminating unnecessary points.

After processing given time series by our program we obtain its piecewise linear approximation. Figure 3 illustrates Piecewise Linear Representation of one year exchange rate of Kyrgyz som to United States dollar with all predefined pivot points before elimination.

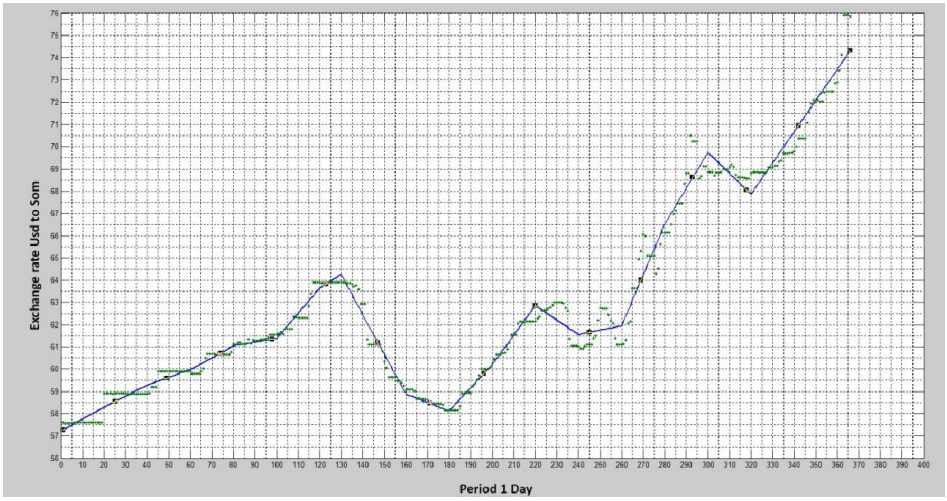

Figure 3. Exchange rate of Kyrgyzstan som to United States Dollar after piecewise linear approximation. Linear segments were taken depending on predefined pivot points.

Apparently can be seen that there are a lot of unnecessary points that can be eliminated. We translate visual definition of unnecessary points to machine by the help of Fisher's criteria.

The result, summarized in Figure 4, is that the algorithm produces result of f criteria of observed pivot points. Here you can see f criteria of observed and tabular ones. With the blue circles marked the points that are less than corresponding tabular f criteria. One of the main ideas of our algorithm is to eliminate these unnecessary points. After elimination of unnecessary pivot points we obtain new piecewise linear approximation with less pivot point and linear segments respectively in Figure 5. Piecewise linear approximation of time series is shown. With red circles we show pivot points. Now our new line segments are ready to be processed to find out similarity.

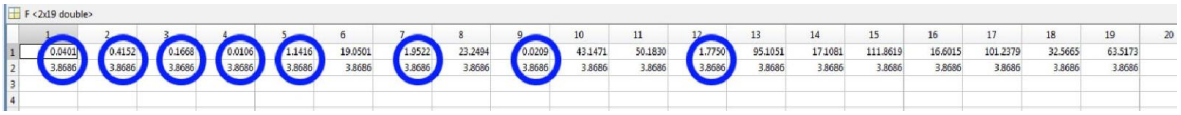

Figure 4. Fishers criteria. First row is the results given pivot points on time series. Second row is tabular f criteria.

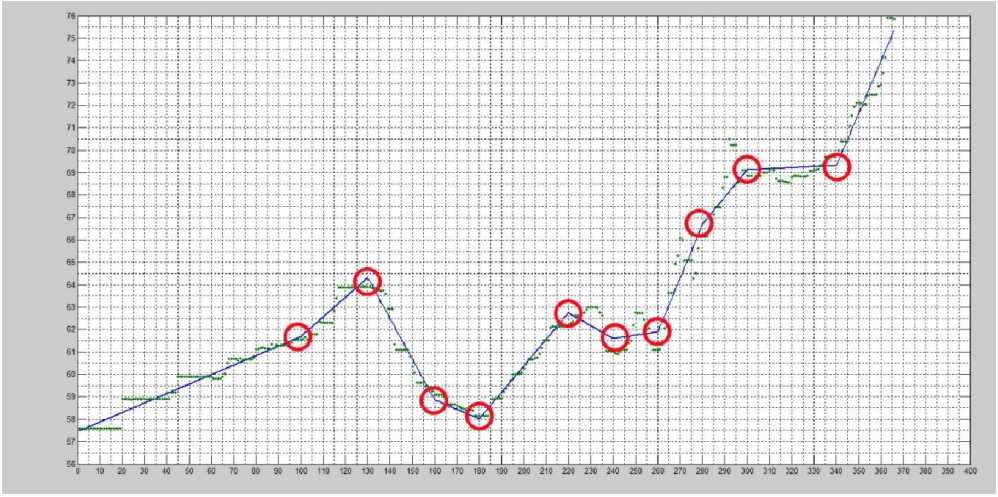

Figure 5. Exchange rate of Kyrgyzstan som to United States Dollar after piecewise linear approximation. Unnecessary pivot points have been eliminated.

Conclusion

We have reviewed most of similarity search methods for fast and efficient similarity measurement of large time series database. And we have shown our approach to this problem which is based on Piecewise Linear Approximation technique. In addition we have introduced the technique for elimination of unnecessary pivot points, which also speed up the process of similarity testing. In continuation of this work, comparison criteria are being elaborated. These criteria in its turn will define the similarity level of tested time series to pattern time series. This work can be extended to many directions such as testing the cyclicality of time series or voice recognition etc.

References:

1. Agrawal, Rakesh, Christos Faloutsos, and Arun Swami. *Efficient similarity search in sequence databases*. Springer Berlin Heidelberg, 1993.

2. Cameron, S. H. (1966). *Piece-wise linear approximations* (No. CSTN-106). IIT RESEARCH INST CHICAGO IL COMPUTER SCIENCES DIV.

3. Milnikov, A. Mert C. Satybaldiev D (2015). A New Method of Piecewise Linear Approximation of Non –Stationary Time Series (No. CSTN-106). IIT RESEARCH INST CHICAGO IL COMPUTER SCIENCES DIV.

4. Milnikov A., Sayffulin S,(2012) Principles of Analysis of Internal Structures of Aggregate Demands. IBSU Journal of Business, 1(1) pp.13-17.

5. www.nbkr.kg

Рецензент: к.э.н. Атабаев Н.У.