A new method of piecewise linear approximation of non-stationary time series

Alexander MILNIKOV*
Cihan MERT**
Daniyar SATYBALDIEV***

Abstract

Recently, in data science one of the most important issues has been discovering actionable information, interpretable patterns and relationships in large volumes of data. This process is called data mining and is commonly being used in science, engineering, business and security. One of the main methods of data mining is similarity search of time series. The approach that is discussed in this article is based on Piecewise Linear Representation of time series that imply two steps of measuring time series similarity. A new method of piecewise linear approximation of non-stationary time series is developed.

Keywords: Piecewise Polynomial Aggregate, support points, multiple nonlinear regressions

Introduction

Recently, one of the most important issues in data analysis has been extracting information from large data warehouses. This process is called data mining and is commonly being used in many business and research applications. One of the main methods of data mining is similarity search of time series. If a large database of time series is given to select the time series that is the most similar to predefined time series, then a proper algorithm is needed in order to discover the most similar time series among thousands. To clarify the problem let's assume that we have predefined time series X which will be taken as a pattern for similarity search. Database D is the set to be tested time series. In this case the main task is to detect the most similar time series to pattern X from database D. This type of example can be extended into much more complex statistics and other diverse fields. A variety of similarity search techniques have been implemented to solve this problem. Finding similarity between two time series basically can be performed with the simplest method called Euclidean distance (Faloutsos, et al., 1994). Dynamic Time Warping technique (Berndt & Clifford, 1994) mostly used in speech recognition field. Another frequently used similarity search technique is Longest Common Subsequence Measure (Das,, Gunopulos & Mannila, 1997). In the present paper new approach and new computational algorithm has been developed for the certain types of data series – the non-stationary data series contained Piecewise Linear profiles (Milnikov & Satybaldiev, 2014) This approach is based on Piecewise Linear Representation of time series that imply two steps of measuring time series similarity which are developing of piecewise linear approximation of time series and test of similarity criteria of measured time series with pattern time series. For this purpose, we developed new technique based on the work of Milnikov & Sayfulin (2012) for the easier representation of nonlinear function in the form of linear segments by approximating it with the help of the sum of linear dependencies.

Theoretical Foundations

Let us set the time-series $x(t_i)$ $(i=1,2,\ldots,n)$, which is characterized by the existence of local linear trends, and presented by n samples taken at equal intervals of time: $\Delta t=T/n$ $(t_i=t_{(i-1)}+\Delta t)$, where the T - is time of observation. It is required to find a piecewise linear approximation. We first consider more general case of a piecewise polynomial approximation.

Piecewise Polynomial Aggregate

Let us consider the system of m polynomials of an independent variable $t\geq 0$.

*Prof. Dr. Department of Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia
E-mail: amilnikov@ibsu.edu.ge

**Assoc. Prof. Dr., Faculty of Computer Technologies and Engineering, International Black Sea University, Tbilisi, Georgia
E-mail: cmert@ibsu.edu.ge

*** Department Electronics and Nano electronics, International Ataturk Alatoo University, Bishkek, Kyrgyzstan
E-mail: daniyar.satybaldiev@iaau.edu.kg

$$P\_i\ (\alpha_1^i,\ldots,\alpha_{ri}^i,\tau_i,t),\ (i=1,\ldots,m) \tag{1}$$

Where $\alpha_j^i\,(\mathrm{j}=1,\ldots,r_i)$- unknown coefficients of *ith* polynomial; r_i – order of ith polynomial.

Let us assume that each of these polynomials equals to zero at $t>\tau_i$, where τ_i- real numbers such:

$$\tau_1 \le \tau_2 \le \ldots \le \tau_{m-1} \le \tau_m = T.$$

We call the intervals Ii=[0,τ_i], as supports of *ith* polynomial (1), i.e., $P_i\left(\alpha_1^i,\ldots,\alpha_{r_i}^i,\tau_i,t\right)\neq 0$ at t$\in\left[0,\tau_i\right],$, and $P_i\left(\alpha_1^i,\ldots,\alpha_{r_i}^i,t_i,t\right)=0$ if not. Let us emphasize that $I_i\subseteq I_{i+1}$ .The intervals I$_i$ are called netsystem of polynomials (1), and points of τ_i – its nodes. 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 supports.

Let us introduce a new function

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

It is not difficult to see that function $F(t)$ is finite on its support [0,T], and continuous on it, however, its first derivative discontinues at net's node τ_i (i=1,…,m). Indeed, at

$$\tau_k \lim_{t\to\tau_k-}F(t)=\sum_{i=k}^{m}P_i(\tau_k) \tag{3}$$

Whereas at

$$\text{t>}\ \tau_k \lim_{t\to t_k+}F(t)=\sum_{i=k+1}^{m}P_i(\tau_k)$$

These limits are equal at point $t=t_k$, as, by definition, of $P_k\,(\tau_k)=0$. At the same time derivative of $P_{k'}\,(\tau_k)\neq 0$, therefore, it is clear that

$$\lim_{t\to t_k-}F'(t)\neq\lim_{t\to t_k+}F'(t) \tag{4}$$

at points $\tau_i\ \ (i=1,\ldots,n)$.

The function (2) is called Approximation aggregate, and polynomials (1) - Aggregate's components. Let us assume that the net's nodes are defined, such, that the points $\tau_i\ \ (i=1,\ldots,n)$ are known, then it is clear that the Aggregate is defined by parameters of

$$r=\sum_{i=1}^{m}r_i\ ,$$

which can be estimated via the least squares method by adding constraints, ensuring the continuation in the nodes.

The following should be noted. Piecewise structure of the Aggregate is determined by the fact that certain components of polynomials are finite, and their supports I$_i$=[0,τ$_i$] 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 follows:

$$F(t)=\sum_{i=k}^{m}P_i(t)\ \text{ for }\ t\geq\tau_{k-1} \tag{5}$$

where i0 –the value of index is equal to index of the τ$_i$, for which condition of

$$\min_{\tau_i}((\tau_i - t)\geq 0)$$

is hold, i.e.it is the one of the values of grid point intervals I_i nearest to the right of the current (wherein we calculate the value of the aggregate).

a_i - slope of the Aggregate's straight line equation, that is a component of a first degree polynomial;

d_i - intercept of polynomial component.

Taking into account that

$$a_i=\frac{d_i}{\tau_i}$$

one can rewrite (5) in the form of

$$F(t)=\sum_{i=i_0}^{m}d_i(1-\frac{t}{\tau_i}) \tag{6}$$

Note that the sum of lower limit is a variable, depending on t, and values of di are unknown. Now it is possible to pose an identification problem.

Given the n observed values of time series of $x(t_i)$ (i=1,2,…,n) and τ_i (i=1,…,m) grid nodes at interval (0, τmax) it is required to determine such values of di, which minimize i functionals of type (4), i.e. local linear approximations of initial time series of $x(t_i)$ (i=1,2,…,n)

$$\min_{\alpha_1^k,\ldots,\alpha_{r_i}^k} S^2=\sum_{i=\tau_{k-1}}^{\tau_m}(x(t_i)-\sum_{j=k}^{m}d_i(1-\frac{t}{\tau_j}))^2 \tag{7}$$

(k=1,2,…,m) (7)

Let us review the problem (7) in detail.

Let us have m intervals l_1 =(0, τ$_1$), l_i =(τ_i -1, τi),…,Ir=(τ$_{r-1}$, τ$_m$), given by m nodes, and let the values tj (j=1,..., n, n ≥m) distributed in such a way that at least one of them falls into one of the intervals Δi. Denote the numbers of points tj, falling into the i-th interval through ki. Obviously the numbers of ki must satisfy the constraint

$$\sum_{i=1}^{r} k_i = n$$

Hence we have m clusters of points t_j, randomly allocated all along the intervals l_i. Note that some of them can coincide with reference points. Consequently, there are m unknowns di, therefore, it is necessary to set n ≥ m values of ti, and thereby $\mathrm{x}(t_j)$ is leading to a system of linear equations with the matrix A_{ji}

$$\mathrm{x}(t_j) = A_{ji} d_i \text{ (i=1,…,m; j=1,…,n).} \tag{8}$$

which is rectangular matrix if n>m and square if n=m.

From the above it is not difficult to conclude that the matrix A of the system (8) for n> r is as follows (Milnikov & Sayffulin, 2012)

$$A = \begin{vmatrix} (1-\frac{t_1}{\tau_1}) & \dots & (1-\frac{t_1}{\tau_i}) & \dots & (1-\frac{t_1}{\tau_r}) \\ \dots & \dots & \dots & \dots & \dots \\ (1-\frac{t_{k_1}}{\tau_1}) & \dots & (1-\frac{t_{k_1}}{\tau_i}) & \dots & (1-\frac{t_{k_1}}{\tau_r}) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & (1-\frac{t_{k_{i-1}+1}}{\tau_i}) & \dots & (1-\frac{t_{k_{i-1}+1}}{\tau_r}) \\ 0 & \dots & (1-\frac{t_{k_i}}{\tau_i}) & \dots & (1-\frac{t_{k_i}}{\tau_i}) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & (1-\frac{t_{k_{r-1}+1}}{\tau_r}) \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \dots & (1-\frac{t_{k_r}}{\tau_r}) \end{vmatrix} \tag{9}$$

while for n=r

$$A = \begin{vmatrix} (1-\frac{t_1}{\tau_1}) & \dots & (1-\frac{t_1}{\tau_{r-1}}) & (1-\frac{t_1}{\tau_r}) \\ 0 & (1-\frac{t_2}{\tau_2}) & \dots & (1-\frac{t_2}{\tau_r}) \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & (1-\frac{t_r}{\tau_r}) \end{vmatrix} \tag{10}$$

In the first case, we have over determined the system of n × r (number of equations is greater than the number of unknowns), and to determine the di it is necessary to use at least square method, and in the second case - the system of linear equations r × r.

It is not difficult to solve the second problem since it leads to the solution of system (8) by matrix (10), which is triangular in our case. Therefore, the solution of the problem can be written even in explicit form without resorting the numerical methods. Unlike to the latter, there is an overdetermined system (the number of equations is greater than the number of unknowns) in the first case. Therefore, it is unnecessary to use the least square method.

It is necessary to note the following. We approximate the time series, which is dealing with the function of one variable . However, the system (8) presents this dependence as a function of m variables: t_i(i=1,…,m), moreover, in (8) there are not used ti themselves, but the functions of the following variables -

$$(1-\frac{t_i}{\tau_j}),$$

which was created by using a language regression analysis, a matrix of observation A of size n × r, shown in (9) and (10). The values of d_i (i=1,…,r) are considered as estimated parameters, i.e., the intercepts of linear components of the aggregate. Accordingly, the problem of piecewise linear approximation of time series is represented as a linear regression problem of m variables. As a result of identification of aggregate components, obtained in such a way, that approximating Aggregate (refer to the proof above in 2. 1) is continuous. In addition, proposed approximation method has advantages compared to the conventional method of constructing the linear splines, as it does not require components matching in nodes of approximation and enables application of standard procedure of multivariate linear regression with no additional restrictions. To reiterate, in this case the columns of matrix (9) must be used as independent variables.

Realization of the method

We represent efficiency of the developed method with the example of time series represented by means of its 463 observations (fig.1 and fig.2, green points).

The approximation procedure should start with forming m x n matrix of m independent variables of price, where m is the number of anticipated elementary demands, defined by backup-prices and observed data. We have created special function in MATLAB programming language. Input parameters of the function are: n - number of observations of demand; m – number of nodes (Milnikov & Sayffulin, 2012). The output of the function is Am×n matrix (9) of m variables of time corresponding to nodes points. Values of each variable corresponded to the certain node τ_i (i = 1,2,…,m) within the interval [0,τ_i] which are not equal to zero, and they all are zeros in the interval $(0,t_n]$. Thus there are m independent variables constructed on the base of the independent variable (time). This approach is very similar to the methodology of regression of dummy variables introduced by Suits (Draper & Smith,, 1998).

After forming the matrix of input variables conventional linear multiple regression method should be applied .

The result of the application of the Piecewise Linear Approximation is shown in the Figure1.

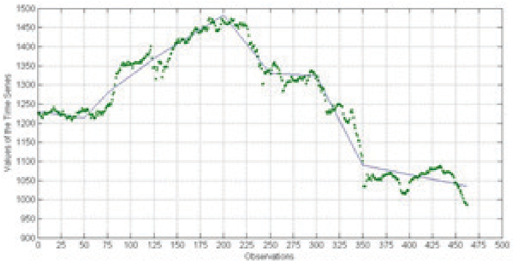

Figure 1. *Piecewise linear approximation of a time series: number of samples, n=463, Number of nodes m=7: 50, 75, 125, 200, 250, 300, 350*

Calculated value of F-criteria estimated for 7-fimensional regression model is $F_{7,362}$=33.61, whereas table value is $F_{7,362}^{tab}$=2.03

We also represent the same time series, but approximated by means of 5 nodes, which is shown in Figure 2.

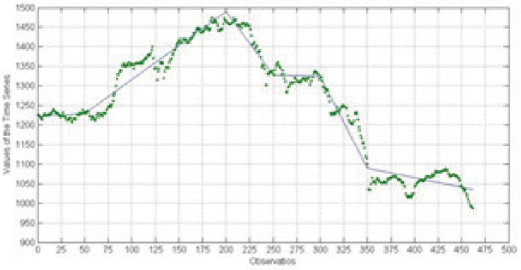

Figure 2. *Piecewise linear approximation of a time series: number of samples, n=463, Number of nodes m=5: 50 200 250 300 350*

We also represent the same time series, but approximated by means of 5 nodes, which is shown in figure 2. Calculated value of F-criteria estimated for 5-dimensional regression model is $F_{5,362}$=71.9, whereas table value is $F_{5,362}^{tab}$=2.24.

It is clear that in both cases approximations should be considered as adequate, but no doubts that taking few nodes is better, as it simplifies model without losing its accuracy. It raises the question: how to determine the minimal number of nodes which gives the best approximation of peculiarities of the Time series under interest? This problem is considered in details in the work of Milnikov & Satybaldiev (2014).

Conclusion

This article has presented the new method of piecewise linear approximation of nonlinear single variable functions with the relevant type of profile.

We would like to underline the benefits of the proposed approximation method compared to the conventional subject of building the linear splines:

1. It uses standard n-dimensional linear regression analysis procedure, which makes it easy to use;
2. As a result of the latter, it does not require usage of restrictions in approximation nodes, thereby simplifying the process of approximation design for certain Time series.

References

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

Norman R. Draper, Smith, H. (1998). Applied Regression Analysis (Wiley Series in Probability and Statistics).

Faloutsos, C., Ranganathan, M., & Manolopoulos, Y. (1994). Fast subsequence matching in time series databases (Vol. 23, No. 2, pp. 419-429).

Berndt, D. J., & Clifford, J. (1994, July). Using Dynamic Time Warping to Find Patterns in Time Series. In KDD workshop (Vol. 10, No. 16, pp. 359-370).

Das, G., Gunopulos, D., & Mannila, H. (1997).Finding similar time series. In Principles of Data Mining and Knowledge Discovery (pp. 88-100).Springer Berlin Heidelberg.

Milnikov A. Satybaldiev, D. (2014) Design Optimal Integral Aggregate Structure Journal of Technical Science and Technologies Vol3, No2 стр. 21-24

Ahlberg, J. H., Nilson, E. N., & Walsh, J. L. (1967).The theory of splines and theirapplications. Mathematics in Science and Engineering, New York: Academic Press, 1967, 1.

Draper, N.R & Smith, H. (1998) Fitting a Straight Line by Least Squares, in Applied Regression Analysis, Third Edition, John Wiley & Sons, Inc., Hoboken,NJ, 10.1002/9781118625590.ch1.