Research of SAX in Distance Measuring for Financial Time Series Data

Liu Wei / Liaoning Technical University
College of Science
Fuxin, China
lv8218218@126.com

Shao Liangshan/Liaoning Technical University
System Engineering Institute
Fuxin, China
lntushao@163.com

Abstract—An effective similarity measure approach on specific data sets is becoming the focus in time series data mining. To solve the problem that financial time series are lacking dynamic information of trend after they are deal with dimension reduction with SAX, in this work we propose a novel similarity measure function, Composite-Distance-Function which joins point-distance advantages and trend-distance advantages together. Through the experiments of SAX with different distance function, we prove that Composite-Distance-Function is a useful function which provides new ideas to reveal the interdependence between the financial data and helps to solve the problem of time series similarity.

Keywords—financial time series; similarity; SAX; Composite-Distance-Function

I. INTRODUCTION

Time series similarity is the basic theoretical issues in the field of time series data mining [1]. Data mining for time series is to study the correlation between sequences, that is, in some measure to represent the degree of similarity between the two sequences.

Time series similarity measure based on generally fall into two categories, one is point-distance, and the other is trend-distance. Time series is often considered higher-dimensional space of points, so we can use point-distance between points to reflect the similarity between sequences. The trend-distance comes from "trend", which is abstract into "rising", "falling" or "maintaining" from the subset of raw time series. Similarity between time series is measured by the proportion of the same trend.

Financial time series data has its own characteristics over other time series data. One of its special characteristics is that dynamic information is always necessary for long-term and short-term analyses. Second one is that financial time series data is continuous, we always need the information of trend starting point, that is, static information. There are many technical analytical methods for financial time series to identify patterns of market behavior. In those financial analytical methods, dynamic information and static information are both very important. So whether we use point-distance or trend-distance to measure similarity separately, we can’t get enough information.

In this work, we propose a new similarity measure approach based on SAX which is called Composite-Distance-Function. The rest of the paper is organized as follows. In Section 2 we review related work and background material, and Section 3 introduces the novel similarity measure function, Composite-Distance-Function. In Section 4 we perform a set of experiments on data mining problems. Finally, in Section 5 we offer conclusions and suggest directions for future work.

II. BACKGROUND

Time series similarity measure approach has attracted enormous attention in the last decade.

SAX (Symbolic Approximation Aggregation) is the first symbolic representation for time series that allows for dimensionality reduction and indexing with a lower-bounding distance measure [3, 4]. In classic data mining tasks such as clustering, classification, index, etc., SAX is as good as well-known representations such as Discrete Wavelet Transform (DWT) and Discrete Fourier Transform (DFT), while requiring less storage space [5]. There are some limitations encountered when we use lower-bounding distance as the similarity measure of financial time series. Lower-bounding distance based on point-distance, but point-distance is a kind of static measure, it can’t effectively reflect the time series dynamic characteristics (the dynamic characteristics of time series).

To reduce a loss of these dynamic characteristics, we propose a new similarity measure approach based on SAX, Composite-Distance-Function, which is especially for financial data analysis and mining tasks.

III. THE IMPROVEMENT OF SIMILARITY MEASUREMENT APPROACH --COMPOSITE-DISTANCE-FUNCTION

A. Novel Pick-up & Measure Method of Symbol List Mode Features--Trend StatisticsVector & Trend Statistics Distance

In order to pick-up the symbol list's ascend, keep, descend modes through use the SAX methods lower dimension well, reflect more dynamic information after lower dimension, and better description for the current statistics character, we enriches the included dynamic information of symbol list based on the literature [6] in this text, coming into being a new current statistics vector construct mode, the details of the process is:

a. calculate the percent of each symbol in the symbol list(unify relation), the size of these character data is ‘k’.
b. set the glide windows' length to '2', discrete the symbol list, get (N-1) sequence duality relations, calculate the percent of each trend of ascend(before<after), keep(before=after), descend (before>after) modes in all the sequence duality relation, these character data is 3.

c. set the glide windows' length to '3', discrete the symbol list, get (N-2) sequence ternary relation, calculate the percent of each trend of wave crest(before<middle, middle>after), sequence ascend(before<middle, middle<after), keep after ascend(before<middle, middle<after), trough(before<middle, middle>after), sequence descend(before<middle, middle<after), keep after descend(before<middle, middle<after), sequence keep (before=middle, middle=after), nine kinds of current in all the sequence ternary relation, these character data is 9. Make all the calculated data into a vector which included (k+12) hefts, we can call this "trend statistics vector":

\[ V = (t_1, \ldots, t_k, t_{k+1}, \ldots, t_{k+3}, t_{k+4}, \ldots, t_{k+12}) \]  

Trend statistics vector reflects the included dynamic information of symbol list to some degrees.

Definition 1: trend statistics distance, That is:

\[ \text{STATDIST} (V_i, V_j) = 1 - \frac{\|V_i - V_j\|}{\|V_i\| + \|V_j\|} \]  

B. Novel Similarity Measure Methods of Symbol List--Composite-Distance-Function

Due to the restriction of point-distance & model-distance single as novel similarity measure benchmarks; it can't reflect the similarity of the stock time series data well. At present, it is a valid method that combines the two distance methods, pool the static information and the dynamic information of the list, and express aggregately while the previous SAX methods have achieved to lower the dimension of the list. It also let us to use some more round and more detailed methods in allowable time complexity range, we put forward a similarity measure foundation of joins point-distance advantages and model-distance advantages together --Composite-Distance.

Set \(X_i, X_j\) is the originality time series, \(X_i, X_j\) is the symbol list which is exchanged by the SAX method, \(V_i, V_j\) is trend statistics vector of symbol list, so the Composite-Distance (Composite-Distance-Function) is:

\[ \text{COMPDIST} (X_i, X_j) = \lambda_1 \text{MINDIST} (X_i, X_j) + \lambda_2 \text{STATDIST} (V_i, V_j) \]  

MINDIST (\(X_i, X_j\)), STATDIST (\(V_i, V_j\)) are \(X_i, X_j\) MINDIST distance and trend statistics distance, measure the static information and the dynamic information of the list separately. We can use follow formula when use the MINDIST distance to measure the several rows time series similarity (for example n rows):

\[ \text{MINDIST} (X_i, X_j) = \frac{\min_m \{\text{MINDIST}(X_i, X_j)_{ij} \mid i, j = 1, 2, \ldots, n\}}{n} \]  

Make the mapping dispose into \([0, 1]\), so that we can do weighting compare. \(\lambda_1, \lambda_2\) is the weight, the value rang is \([0,1]\), it reflect the different status of point-distance advantages and model-distance, the initial condition of the two is equal, all equal to 0.5. If it is need to add more dynamic information when do the similarity measurement, add the value of \(\lambda_2\). The opposite: add the value of \(\lambda_1\) to reflect more spatial information between the lists. In this way, point-distance advantages and model-distance are combined effectively. The two is the limit case between \(\lambda_1=1, \lambda_2=0\) and \(\lambda_1=0, \lambda_2=1\)of the Composite-Distance. In practice, we can enact \(\lambda_1, \lambda_2\) bases on the specific data sets and different methods which are analyzed by the experts in the different data mining tasks, and also can exercise and tune up by the neural network. In the study of similarity of financial Time Series Data set, we can set \(\lambda_1=0.2, \lambda_2=0.8\) by the advice of negotiable securities commentator.

IV. EXPERIMENTAL

In order to validate the validity of composite-Distance in the similarity measurement, we choose the stock time series data set which has the apparent trend character. Take out the SSE (Shanghai Stock Exchange) Composite Index, Shenzhen Component Index, 9 stocks’ closing price of Stock Market at Home from Jan of 2007 to Nov of 2008 which is 450 sequence market days as the data of the experiment, the data cover many different field such as big-cap stocks, small-cap stocks, steel, bank, tour, public utility, communication, Real Estate, etc. Their standardization data pictures as follow:

![Figure 1. 9 stocks' closing price of Stock Market.](image)

There are many detailed fluctuation in the data, we still can class them into four categories bases on the whole trend:

C1: The trend of SSE (Shanghai Stock Exchange) Composite Index(1) and Shenzhen Component Index(2) is similarity, not only the time and range of the ascend but also the time and range of the descend. There are much tiny difference on the trend of MaGang stock (6), BaoGang stock (7), China Bank (3). But the whole trend and the fluctuations time are similar to the trend of SSE (Shanghai
Stock Exchange) Composite Index and Shenzhen Component Index.

C2: It was much similar between the trend of Beijing tour (4) and the trend of China Youth Travel Service (5), but they are significantly different to other stocks.

C3: ZhongChuangXinCe(8).

C4: YiYang communication (9). They are all different from other stocks.

In order to validate the validity of Composite-Distance in the similarity measurement, we designed four kinds experiment schemes based on the literature [6]:

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Data Set</th>
<th>Method of Dimensionality Reduction</th>
<th>Measure Methods (Distance Function)</th>
</tr>
</thead>
<tbody>
<tr>
<td>One</td>
<td>One stock data</td>
<td>null</td>
<td>Euclidean Distance</td>
</tr>
<tr>
<td>Two</td>
<td>Two stock data</td>
<td>SAX method</td>
<td>Lower Bounding Distance</td>
</tr>
<tr>
<td>Three</td>
<td>Three stock data</td>
<td>SAX method</td>
<td>Trend Statistics Distance</td>
</tr>
<tr>
<td>Four</td>
<td>Four stock data</td>
<td>SAX method</td>
<td>Composite-Distance</td>
</tr>
</tbody>
</table>

Use the hierarchical clustering arithmetic to analyze the four experiment results, we can get different clustering results, compare them to C1 C2 C3 C4 which is based on the whole current and classed by the standards, we can define the veracity of different distance function for the similarity measurement. Because there is no uniform method for the preferences in the symbol set approximate process, we make certain that: the originality time list data’s length is \( n=450 \), after lower dimension, the length is \( n=90 \), the symbol number is \( \text{alphabet size}=10 \), compressibility is \( \text{compression ratio}=5 \). It needs to use the formulas:

\[
x(i,j) = \frac{x(i,j) - \min(x)}{\max(x) - \min(x)}
\]

[5]

to do unitary dispose to the calculated result matrices of point-distance. The weight of point-distance advantages and model-distance are \( \lambda_1=0.2, \lambda_2=0.8 \) in the Composite-Distance. The four experiment clustering results as following:

From the clustering results, we can see that trend statistics distance method achieved to plot out the list from the trend exchange angle, but the spatial location measurement is not very good. Composite-Distance-Function do the further amend at the basic of trend statistics distance, achieved the similarity measurement of the stock time list data. Although its time are a little more complexity than Euclidean distance and the MINDIST, its precision and impact is better than Euclidean distance and the shortest boundary distance from the clustering results.

V. CONCLUSIONS AND SUGGESTIONS FOR FUTURE WORK

We bring a new symbol list similarity measurement methods--Composite-Distance-Function, it can use the symbol list which is lowered dimension by SAX method with the static information and the dynamic information to make aggregate measurement, exchanged the single character problems of the old similarity measurement. Emulate experiment shows the trend statistics vector which reflect the trend character of the symbol list; Composite-Distance joins point-distance advantages and model-distance advantages together very well, aggregate measurement from the spatial location and exchange trend angle, and well performance, it is a effective similarity measurement method. We will focus on the selection of different spot distance style in the future and the addition of dynamic information in trend statistics vector and turn up the weight of the two distances, to make further veracity of similarity measurement. In addition, it is one of the main research directions that how to make the effective measurement when there is more similar exchange or more different exchange between two time series.

REFERENCES


