An Algorithm for Time Series Data Mining Based on Clustering

Shaozhi Wu[1] Yue Wu[1]
1 School of Computer Science and Engineering,
University of Electronic Science and Technology
of China Chengdu, 610054, China
E-mail:wszfrank@126.com
ywu@uestc.edu.cn

Ying Wang[2] Yalan Ye[1]
2 Electric Engineering Department Chongqing University
Chongqing, 400044, China
E-mail:wywpc@126.com
yalanye@uestc.edu.cn

***Abstract*—This paper presents a new method for time series data mining. Discrete Fourier Transform (DFT) is used to transform the time series data from time domain to frequency domain. By taking the transformed amplitude of power spectrum as the feature samples of the time series data, time series data can be mapped into a frequency domain space. We use OPTICS (Ordering Points To Identify the Cluster Structure) algorithm to detect clusters in these data. Several simulations are given based on the price histories of California power market.**

I. INTRODUCTION

With the rapid development of communication and data storage technology, more and more data is stored in database or warehouse. It's a great challenge for us to explore new usage of the mass data on current data analysis and prediction, thus arousing the birth of data mining technology. Data mining is a process during which we analyze and extract knowledge with learning techniques within a database. Its purpose is to extract interesting, connotative and unknown knowledge, and its object is structured data, time series data, hypertext or multimedia data.

Many studies focus on strategies of complex structured data mining. Commonly, the data we want to analyze contains a time dimension. These datasets with time-dependent attributes are called Time series data, and it is quite important to discover useful knowledge and rules in time series data.

II. KNOWLEDGE CORRELATIVE

In the past few years, many algorithms have been proposed for the knowledge discovery within time series data, such as trend analysis, similarity search, sequence pattern mining etc. [1][2][3] .

Faloutsos [4] put forward the concept of time window and disassembled time series into a series of subsets of time series, then extracted features for the similarity searching from the subsets.

Clustering [5] is a widely used strategy specialized in building models from data without predefined classes. Its result is in the form of a set of clusters. Moreover, objects resemble each other in the same cluster, and differ much from those in the other ones. As a tool, Clustering can be used independently to find out the distribution instance of the datam, to observe the characteristics of each cluster and to make further analysis for some interesting clusters. In general, there are three major methods of clustering, i.e. partitioning method, hierarchical method and density-based method.

III. METHODS

In this paper, density-based clustering method is used in time series data of power market. In order to measure the distance between each object,we use DFT to transform the time series data from time domain to frequency domain. Distance in the mapped space is defined based on Euclid distance of different objects.

A. *Discrete Fourier Transform(DFT)*

DFT is used to transform sampling from time domain to frequency domain. According to the Theory of DFT, the maximum values of the harmonic wave in frequency domain are samples of the original time series. In order to discover the knowledge in history of power market, we disjoint the dataset of the power market into many subsets. Different rules can be applied to disjoint the time series. Subsets can be grouped in hours, days or months; furthermore, it can be classified by the rules "each Monday of every week", "The same date of each month" etc. We can disjoint the original series into many subsets by different rules. DFT is applied respectively to these subsets.

Given a time series$\{X[t]\}0\le t\le N$ N is the sample number of the dataset, and we get m subsets through the disjoint.X[1],X[2],…X[l],…X[m],$X[1]=\{x_{ij}| j=0,1,2,\ldots n\text{-}1\}$, N=nm, DFT was applied to these subset X[l]

$$x_f = \frac{1}{\sqrt{n}} \sum_{t=0}^{n-1} x_t \exp\left(-i \cdot 2\pi ft / n\right) \quad (1)$$

$f = 0, 1, 2, \ldots, n-1$ i is the plural symbol

x_f stands for the Fourier coefficient of the subset, it can be dissolved into amplitude and phase angle of f-harmonious wave in the frequency domain. These coefficients are samples of the Fourier transformed original time series. $\vec{X}_{iF}$ is a set of coefficient of Fourier transformed subset $X[i]$

$$\vec{X}_{iF} = \left\{ x_{if} \middle| f = 0, 1, 2, \ldots n-1 \right\} \quad (2)$$

Based on the Parseval theory, energy spectrum function of time domain is equal to that of frequency domain.

$$\left\| \vec{X}_i - \vec{Y}_i \right\|^2 = \left\| \vec{X}_{iF} - \vec{Y}_{iF} \right\|^2 \quad (3)$$

$$D\left(\vec{X}_i, \vec{Y}_i\right) = \left(E\left(\vec{X}_i - \vec{Y}_i\right)\right)^{1/2} = \left(E\left(\vec{X}_{iF} - \vec{Y}_{iF}\right)\right)^{1/2} = D\left(X_{iF}, Y_{iF}\right) \quad (4)$$

Even though this transformation had been made, the distance between objects still kept the same values.[5]. FFT (the Fast Fourier Transform) is a discrete Fourier transform algorithm and can reduce the times of computations for N elements from $2N^2$ to $2N\log_2 N$ when compared to Fourier transform algorithm, here the number of objects must be the power of 2, so the elements number of each subset must be the power of 2, zero can be added if the number doesn't satisfy this condition.

B. *Clustering within Time series mapped space*

The Ordering Points to Identify the Clustering Structure (OPTICS) algorithms promoted by Mihael Ankerst and Markus M. Breuning [6], is a density-based clustering algorithm. It can solve the problems faced by other density-based clustering algorithms: 1) difficulties to determine the input parameters; 2) extreme sensibility to the parameter values. This algorithm does not produce a clustering of a data set explicitly, but it creates an augmented ordering of the database representing its density-based clustering structure. This cluster-ordering contains information about every clustering level of the data set, and is very easy to analyze.

The mapped space forms a new data set involving k-dimension database. Clustering analysis can be processed within this database. Correlative definitions are illustrated as follows:

Definition 1. (Fourier transformed database) D_F

D_F refers to the mapped dataset; the elements are the Fourier transformed discrete points in frequency domain. $\{x[t]\}0 \le t \le N$ refers to the original time series dataset. If it is disjointed into m subsets, the element number of the D_F is m.

Definition 2. (Distance in D_F *) Dist-F(p,o)*

$p, o \in D_F$, $p=(p_1, p_2, \ldots p_k)$ $o=(o_1, o_2, \ldots o_k)$,then *Dist-F(p,o)*=

$$\sqrt{(p_1 - o_1)^2 + (p_2 - o_2)^2 + \ldots + (p_k - o_k)^2} \quad (5)$$

Definition 3. (Core-distance of an object p)

Let p be an object of database D_F, let ε be a distance value, let $N_\varepsilon(p)$ be the ε-neighborhood of p , let *MinPts* be a natural number and *MinPts-Dist-F(p)* be the distance from p to its *MinPts'* neighbor. Then, the *coer-distance* of p is defined as

$$core\text{-}distance_{\varepsilon, MinPts(p)} = \begin{cases} UNDEFINED, \ if \ Card(N_\varepsilon(p)) < MinPts \\ MinPts\text{-}Dist\text{-}F(p), \ otherwise \end{cases} \quad (6)$$

Card(A) is a function returning numbers satisfied condition A

Definition 4. (Reachability-distance of an object p w.r.t. object o)

Let p and o be objects of D_F , let $N_\varepsilon(o)$ be the ε-neighborhood of o ,and let *MinPts* be a natural number. Then the reachability-distance of p with respect to o is defined as

$$reachability\text{-}distance_{\varepsilon, MinPtspp}(p, o) = \begin{cases} \text{UNDEFINED}, & \text{if } Card\left(N_\varepsilon(o)\right) < \text{MinPts} \\ \text{MinPts-Dist-F(p)}, & \text{otherwise} \end{cases} \quad (7)$$

The OPTICS algorithm creates an ordering of the database. In addition, it stores core-distance and reachability-distance for each object. This information can help to extract all density-based clusters.

IV. SIMULATION

We simulate the algorithm based on the power price histories of California power market in 2000. The time-price curve is shown in figure 1. We disjoint it into subsets according to date. That is, each subset represents 24 hours of a day. DFT is used on each subset to get the Fourier coefficients, because the last 12 coefficients are conjugate to the first 12 coefficients, we pick the first 12

coefficients to represent the time series data. Plural Euclid distance is the measurement of distance between two objects. We find several core points with respect to ε and MinPt. Objects which are directly density-reachable from these points are extracted from the database and can be inserted into the seeds list for further processions. Points in the dataset with defined reachability-distance can be inserted into a queue sorted by their reachability-distance. The generation distance ε determines the number of clustering-levels. Fig.2 and Fig.3 illustrate the effect of ε in two simulations.

These two figures show that the generating distance ε influences the number of clustering-levels. The smaller we chose the value of ε, the more objects may have undefined reachability-distances. Consequently, we may not see clusters of lower density.

Further simulations are applied on these objects. Here we cluster the objects according to the cluster-order in Fig. 2. The clustering result is shown in Fig. 4.

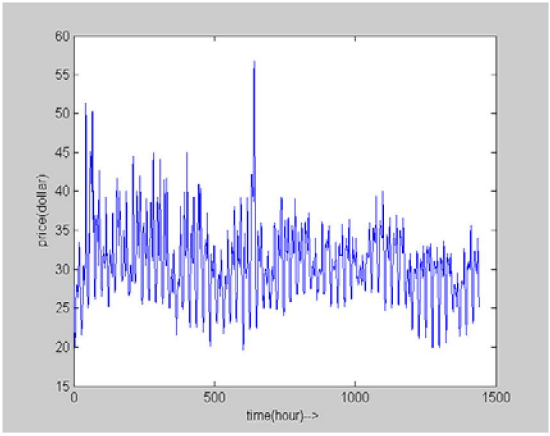

Fig. 1 time-price curve

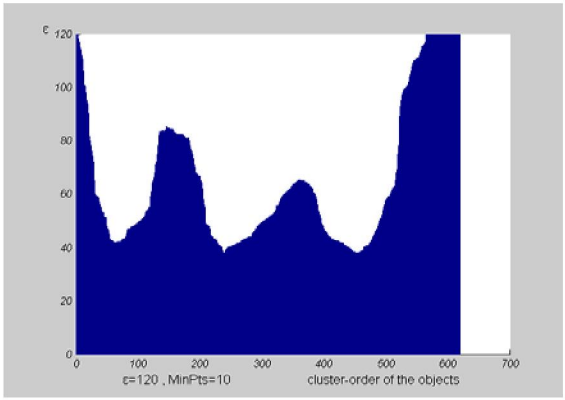

Fig. 2 cluster order of the objects with respect to ε=120 and MinPts=10

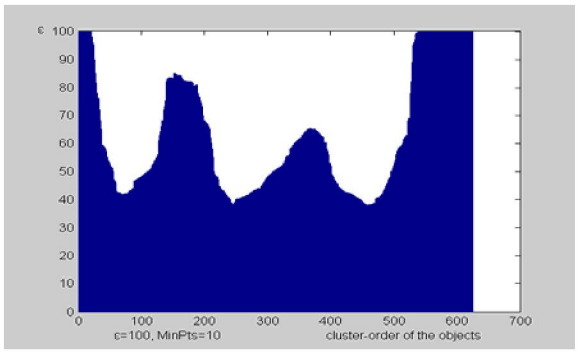

Fig. 3 cluster order of the objects with respect to ε=100 and *MinPts*=10

In Fig. 4, we find three high density clusters, and this clustering result indicates the cluster-order in a perfect way. Those pits in Fig. 2 and Fig. 3 represent cluster center, and most of the objects in the dataset gather around the three centers. We can get common rules of power market by analyzing the result, that is, in certain period of certain date, power price is similar with each other. We can predict price accurately and earn more in market competition by taking full advantages of these rules.

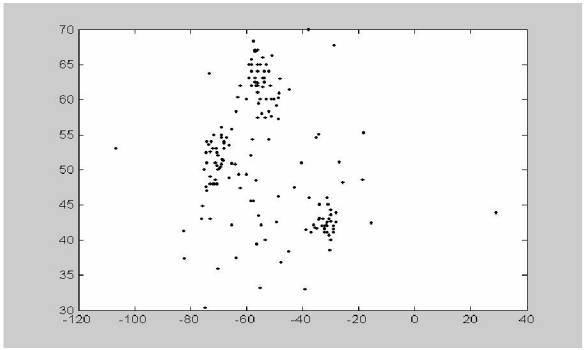

Fig 4 cluster result

V. CONCLUSION

In this paper, we proposed a clustering analysis within time series data. In order to make better analysis of time series data, we use DFT to transform the data from time domain to frequency domain. By taking the coefficients of DFT as the coordinate of a space, we create a mapped database in plural space. Then we use density-based clustering algorithm-OPTICS to detect clusters. These clusters can help us in knowledge discovery in the time series date. It can be used to predict the trend and to detect the anomaly in the power market. Simulations are given to demonstrate the feasibility of the method.

ACKNOWLEDGMENTS

I'd like to thank Mr. Zhen Liu and other co-workers for their help and support in making this work possible.

REFERENCES

[1] Keogh E and Kasetty S. On the need for time series data mining benchmarks: A survey and empirical demonstration. In: Proc. of the 8th ACM SIGKDD Int'1 Conf. on Knowledge Discovery and Data Mining. New York: ACM Press, pp.102~111.2002

[2] Li AG, Qin Z and He SP, "Extracting similar patterns in time series data". Journal of Xi'an Jiaotong University, Vol.36, pp.1275~1278, 2002

[3] Bin-xiang Zheng and Xiu-hua Du, "A New algorithm of similarity mining in time Series data", Information and control, Vol.31,pp.1002~0411, 2002

[4] C Faloutsos, M. Ranganathan, Y M anolopoulos Fast subsequence matching in time series databases in Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'94), pp.419~429, ACM Press, 1994

[5] Xiao-shuai Xing,and Li-cheng Jiao, "Clustering Method in the Field of Data Mining", JOURNAL OF CIRCUITS AND SYSTEMS, Vol.8, pp.1007~0249, 2003

[6] Mihael Ankerst, Markus M. Breuning, "OPTICS: Ordering Points to Identify the Clustering", Proc, ACM SiMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999