Dynamic Data Mining Based on Cloud Model

Zhenguo Zhu[1,2], Ying Kong[2], Gen Pei[2]
1. School of Information Science, Southwest Jiaotong University, Chengdu, Sichuan, 610031, China
2. School of Information Science and Engineering, Chongqing Jiaotong University
e-mail:zhuzhg@qq.com, kongying168@qq.com, hurrican6@gmail.com

***Abstract*—With its continuing in-depth research, data mining has become gradually mature. In comparison with static data mining, dynamic data mining is being appreciated for its capacities of dealing with the data which are uncertain, irregular and obviously time-related. This paper, based on the cloud model theory, proposes a new method of dynamic data mining, which effectively helps handle the dynamic data and achieve some practical values.**

Keywords -data mining, dynamic data, cloud model, cloud transformation

I. INTRODUCTION

As is known, static data mining [1~5],which is still short of such attributes as time, is now a mature technology. However, there also are lots of dynamic data [6~8] such as phone call logs, credit card transaction, share dealing, web logs and so on that needs exploration. These data are usually ever-changing, and it's necessary to uncover the data dynamically from the database to achieve the latest information. In such case, we proposed a new method named dynamic data mining based on cloud model (DDM-CM) as well as existing achievements to deal with dynamic data, and achieve some practical values.

II. DYNAMIC DATA MINING (DDM)

Data in dynamic database contain not only the information to be covered in the static data but also such time-series attributes as documents, pictures or videos related to the time behavior. The key value of dynamic data mining is to pick up and analysis the data from the Dynamic Data Source (DDS) dynamically in order to obtain corresponding knowledge and information. The operational procedure [7] are supposed to include dynamic data acquisition, data handling, data mining, data evaluating as well as data application. Some relevant definitions are as follows:

The data in dynamic data source (DDS) are marked as d_i (where i means data labels. $i \in Z+$).

Set the current time as T, and define one constant δ ($\delta \in R+$). All d_i which are generated till $T-\delta$ from history data set, marked as D_{old}.

Set the current time as T, and define one constant δ ($\delta \in R+$). All d_i which are generated between $T-\delta$ and T from history data set, marked as $D_{current}$.

Set the current time as T, and define one constant δ ($\delta \in R+$). All d_i which are generated after T from history data set, marked as D_{new}.

In terms of the definitions above, we can discriminate the difference between static data set and dynamic data set obviously. Now, let see the traditional dynamic data mining architecture [7] (Figure 1).

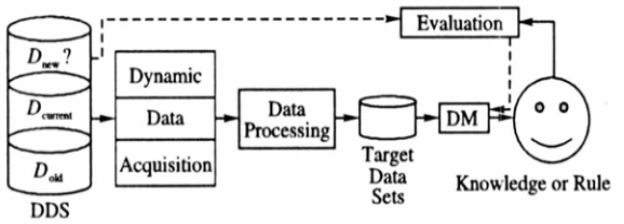

Figure1.Dynamic Data Mining Structure

The DDS (Dynamic Data Source) stores history data set, current data set and new data set.

And DDA means dynamic data acquisition. It obtains history data set, current data set from the DDS, this processing may be finished one time. But, the processing to obtain the new data set should be executed separately. DP means Data processing. It deals with the problems about pretreatment of the data set, which include treatment of missing data, data cleaning, data type conversion as well as dimensionality reduction.

III. CLOUD MODE

The Cloud Model provides a mechanism to convert qualitative concepts described with linguistic values to the quantitative expressions. It reflects the fuzziness and the randomness of the knowledge [9]. The model contains three numerical characters [9], Ex (Expected Value), En (Entropy), and He (Hyper Entropy). Where "Ex" reflects the barycenter of cloud droplets, which stands for the value of the qualitative concepts. " En " is used to measure the fuzziness and probability of the qualitative concepts and it reveals the relationship between the fuzziness and randomness. While "He" is a measure of uncertainty, which reflects the vergence of the cloud droplets. Figure 2 is the cloud model display of the linguistic values about "20 years old ".

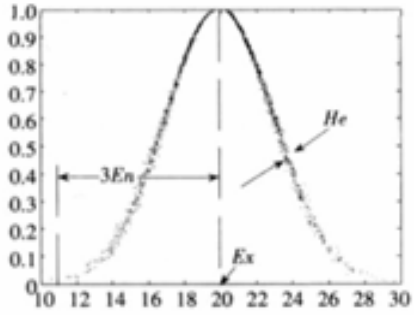

Figure 2. cloud model about 20 years old

The Cloud Generator [10] is just a kind of arithmetic, and used to transform the qualitative concepts into the quantitative expressions or achieve the opposite function. It includes normal cloud generator and backward cloud generator. Normal cloud generator (Figure 3.) converts the qualitative concepts into the quantitative expressions. It produces the cloud droplets and the cloud droplets will be cloud cluster when their number is sufficient enough. The input data set of CG is Ex, En, He and the number of the cloud droplets, N. The output is the quantitative value of the cloud droplets as well as the certainty that the cloud droplets stands for. And the one-dimensional Cloud Generator can be so easily extrapolate to two-dimensional , that we drop it here[11].

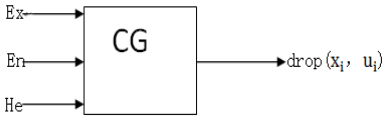

Figure 3. Cloud Generator

IV. THE COMBINATION OF DDM STRUCTURE AND CLOUD MODEL

We'd like to improve the traditional dynamic data mining structure in order to couple it with the cloud model. The new model is described as in Figure 4.

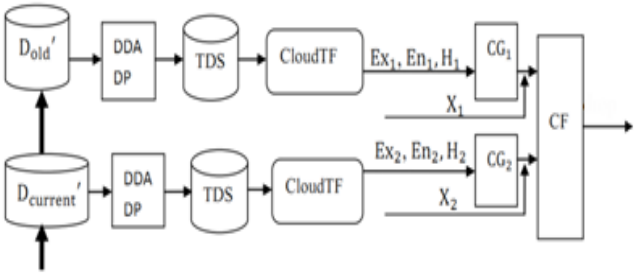

Figure 4. Dynamic Cloud Model

Below are new definitions based on the new model:

The data in the dynamic data source (DDS) is marked as d_i, where i means data labels. $i \in Z+$).

Set the current time as T, and define one constant δ ($\delta \in R+$). All d_i which are generated till $T-\delta$ from history data set, marked as D_{old}.

Set the current time as T, and define one constant δ ($\delta \in R+$). All d_i which are generated between $T-\delta$ and $T+\delta$ from history data set, marked as $D_{current}$.

Here, we do not define the new data set (D_{new}), because the current data set can most reflect the features of the current event effectively, and the user may only interested in these data set. And with the data coming continually, the new data set will be the old one finally.

To be simply, we combine the DDA module and DP module together. The CloudTF means Cloud Transformation [12]. It samples the serial data to discrete data, and generates the three parameters of the cloud model. The CF means Curve Fitting [13]. It deals with the problems about curve fitting. "X_1"and "X_2"means the threshold. They are used to control the proportion of the curve set when fitting the history cloud droplets. And $0 \le X_1 \le X_2 \le 1$, $X_1 + X_2 = 1$.

The model above we can call it "Dynamic Cloud Model (DCM) ".

One critical problem of the dynamic data mining is to monitor the data stream and produce a structure that can present the changing data set in memory. So we can use the TCP slide windows [14] as temporary storage in the DDA and DP process. See Figure 5.

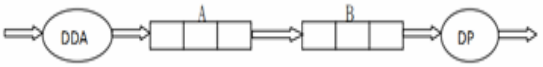

Figure 5. slide window

There is a slide window in each port of DDP and DP (A and B). And each window has two flags, where "F" means the volume of the window is full, and "E" means it is empty. The realization of this mechanism can be referenced in [14] detailed. We can set different weights to the data set before processes start. The older the data the lighter the weight. The common way to set the weight is to define an attenuation function [7]. It expresses as:

$$f(t) = 2^{-\lambda t} \quad (1)$$

Where "λ" is a kind of regulation parameter, the larger the value the lower the importance of the data set, "t" is the time when the data set arrival, and the larger the value the lower the importance of the data set as the same. The idea that set different weight to the data sets reflects the timeliness and the unlimitedness of dynamic data. Then when the data set from TDS through Cloud Transformation are made discrete, and several different granular cloud droplets through Cloud Transformation arithmetic [12] are generated. It expresses as:

$$f(X) = \sum_{i=1}^{n} ((\alpha_i * (Ex_i, En_i, Ee_i)) \quad (2)$$

Where "α_i" is the coefficient and "n" is the number of the discrete concept. In this way, it can change any irregular distribution into several superimposed cloud. The more the

number of this cloud the more precise the result. Input "Ex_i, En_i, He_i" generated by the history data set and current data set to the CG to generate the cloud droplets. Lots of droplets form the cloud curve. Then fit the data on the curve or fit the curve with the "Curve Fitting Module " to generate more precise result. Figure 6 shows the sales volume curve of cars resulted from the cloud curve set.

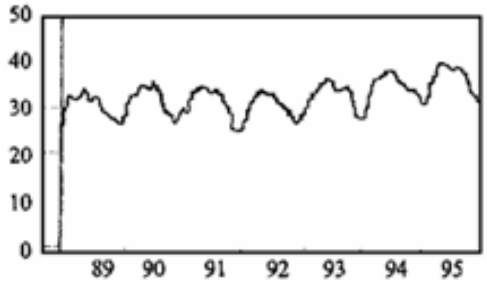

Figure 6. sales volume curve

It can be used to forecast the Temporal sequence [15, 19~20], to mine the Association Rules [16] with the cloud curve and some relevant data. It can also assess the existing model [17] and detect the support [18] by feeding the cloud curve to dynamic cloud model.

V. APPLICATION OF THE DYNAMIC CLOUD MODEL

Now we can mine the potential and valuable information from the ever-changing real-time data set with this new model. The case below is about the stock data of a company, of which we will make a brief analysis. The data shown below is obtained from January 1, 2000 to April 1, 2011.

TABLE 1. STOCK DATA

Date	Open	Low	High	Close	Amount	Volume
2001-1-1	0.4	0.4	0.41	0.4	8160	330000
			…			
2011-4-1	1.48	1.46	1.51	1.47	11600	1722600

Here, we just analysis the tendency curve of the stock data during a period of time, so we select the"Open","Close" (mark the average of the two as Mid_{OC}) , "High","Low" (mark the average of the two as Mid_{HL}) and the"Date", and ignore the "Volume" and "Amount". Let us suppose that current time T is November 1, 2010, and $\delta = 5$ (month as the unit). We define the data between January1, 2000 and Mary 31, 2010 as history data set, the data between June 1, 2010 and April 1,2011 as current data set. The data set which processed by DDA module and DP module should generate the target data set, then input the target data set into the CloudTF to form the discrete data as well as the Ex_i, En_i, He_i. We just show the discrete data.distributing chart of the Mid_{HL},as in Figure 7.

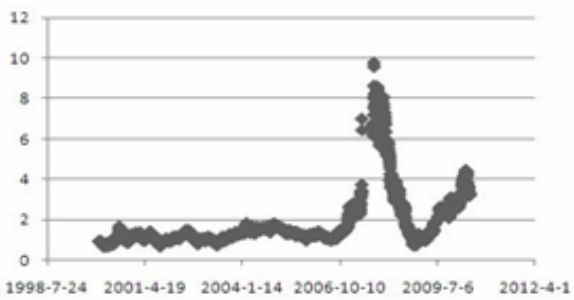

Figure 7. distribute of the Mid_{HL}

Then input the "Ex_i, En_i, He_i" into the cloud generator to generate the cloud droplets. Build the current data set droplets and history data set droplets into a set of smooth curve with the CF module. Now we can forecast the trend of the stock according to this curve [19~20]. And the Figure 8 is the result of the predictive curve and the distribution of the error between the real data and predictive value.

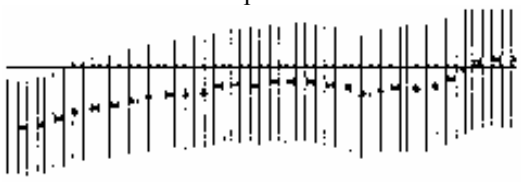

Figure 8. predictive curve and the distribution of the error

The curve describes the trend of the stock between the March 1, 2011 and April 1, 2011.

The horizontal line means the "X axis" and it stands for nothing.

The curve below the X axis means predictive value of the stock.

The vertical bar means the error between the real data and predictive value.

The Max error between the real data and predictive value is only just ± 0.25, which shows the preferable robustness and feasibility of this model strongly.

VI. SUMMARIES

In recent years, the achievements based on cloud model in data mining field have been approved by the experts at home and abroad. This article comes up with a new method named dynamic cloud model to couple the dynamic data mining structure with the cloud model. We can deal with the dynamic data which based on timing sequence together with the dynamic cloud model, but not when it comes to the multi-data stream. And this is what we're to cope with in future. There is no denying that the cloud as well as its mutation will have a glorious prospect in the data mining field for the time to come.

REFERENCES

[1] Guojun Mao,Lijuan Duan.Data Mining Theory and Algorithm. Tsinghua University Press, 2007.

[2] Jingmin Chen. Data Warehouse and Data Mining. Electronic Industry Press, 2002.

[3] Heng zhao. Cluster Analysis in Data Mining. XiAn: Ph.D. Thesis In XidianUniversity, 2005.
[4] Agrawal R,Imielinski T,Swami A.Mining association rules between sets of item in large database Proceding of the ACM SIGMID Conference on Management of Data, Journal of Computational and Applied Mathematics,1993:207~216.
[5] Chen M, Lapaugh A S, Singh J P. Predicting category accesses for a user in structured information space Proceding of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,Computers and Mathematics with Applications,2002:65~72.
[6] Henzinger M R, Raghavan P, Rajagopalan S. Computing on data streams, Dimacs Series in Discrete Mathematic and Theoretical Computer Science. 2005:2107~108.
[7] Zhiwei Ni. Dynamic Data Mining, Science Press, 2010.
[8] Jiawei Han. Data Mining Concepts and Techniques. Machine Press, 2005.
[9] Deyi Li, X M Shi, Gupta M M. Soft Interference Mechanism Based on Cloud Models, In: Proceeding of the 1st International Workshop on Logic Programming and Soft Computing: Theory and Application, Bonn, Germany, 1996.
[10] ZhaoHui Yang, DeYi Li. Planar Model and It's Application In Prediction. Chinese J. Computers, 1998.
[11] DeRen Li, ShuLiang Wang. Spatial Data Mining Theory and Application. Science Press, 2007.
[12] Yi Du. Aritificial Intelligence with Uncertainty. Beijing: National Defense Industry Press, 2005.
[13] Shiliang Xu. Computational Methods. Beijing: Posts and Telecom press, 2009.
[14] W. Richard Stevens.TCP/IP Illustrated. Beijing: China Machine Press, 2000.
[15] The STREAM Group. STREAM: the standford stream data manager. http://www.db.stanford.edu/stream.
[16] Han J W, Jian P. Mining frequent patterns without candidate generation: a frequent-pattern tree approach. Data Mining and Knowledge Discovery, 2004, 8(1): 5357.
[17] Guha S, Mishra N. Clustring data streams. In: Proceeding of the 41st IEEE Symposium on Foundations of Computer Science, 2000, (12):359~366.
[18] Muthukrishnan S. Data streams: algorithms and applications Proceedings of the 14th Annual ACM-SIAM Sympon Discrete Algorithms, 2003.
[19] Fan Yang, Ye Wang, Deyi Li. Cloud prediction based on time granularity. In: Proceedings of 5thPacific-Asia Conference on Knowledge Discovery and Data Mining, Hong Kong, 2001.
[20] Rong Jiang, Hui Chen. Time-series prediction with cloud models in DMKD. In: Proceedings of 3rd Pacific-Asia Conference on Knowledge Discovery and Data Mining, Beijing, 1999.

Published by Atlantis Press, Paris, France.