The Forecasting of Evolutionary Multi-Source Data limits

Guoqi Chen[1, a], Fang Yu*[2,b]

[1]School of Automation, Wuhan University of Technology, Wuhan 430070, P.R.China.

[2] Department of Science, Wuhan University of Technology, Wuhan 430070, P.R. China.

[a]3341883515@qq.com, [b]yufang9977@163.com

Keywords: Data Evolution, Dynamic Evolution Factor, Evolutionary Matrix, Limit Distribution

Abstract. This paper gives a method of forecasting data evolution in a long term, its practical significance is predicting data evolution direction and targeting the data that need to be updated.With the assumption of no absolutly real data in the world, we could view the data updating process as a data evolution in order to infinitely close to an objective entity. The EVO elements could be considered as the most close to real data in a specified period. In the long term, data evolution constitutes a Markov chain. According to Chapman-K Olmogorov Equation, we derive the Markov chains of data evolution in the long term. Furthermore ,we greatly simplifies the evolutionary computation by a properties of limiting distribution and the EVO elements. Experiments show that after a great times of updating , it still has a high prediction precision for the initial price.

Introduction

At present, many researches, on discovery of data true value and data inconsistent, assume that there is a only correct data. However, in reality, the real data is constantly changing, for example, with the improvement of measurement technology, we could measure the radius of the earth more accurately.This paper regard the data updating process as a data evolution which infinitely close to an objective entity.

The EVO elements can be considered as the most close to real data within a specified period. In the long term, data evolution constitutes a markov chain. According to Chapman-Kolmogorov Equation, we derive the markov chains of data evolution in the long term, what is more ,we greatly simplifies the evolutionary computation by the properties of limiting distribution and EVO elements.

Equivalent Group Set

The section headings are in boldface capital and lowercase letters. Second level headings are typed as part of the succeeding paragraph (like the subsection heading of this paragraph). Assuming that under the relational schema R, an entity E is described by n data source, they generate a list of n records $R=\{e_1,e_2...e_n\}$, where the R is called record list and the elements $e_1,e_2...e_n$ are records from n data source. Obviously, the elements in R are random sequence, however, we more concernes that how many of these elements are the same, and put these elements as a group. $G=\{g_1,g_2...g_t\}$, where the G is an equivalent group set if the elements $g_i,g_j,(i,j=1,2,...,t)$ meet the following properties:1. $g_i,g_j\subset R$, 2. $g_i\bigcap g_j=\varnothing$, 3. $\forall e_{i_1},e_{i_2}\in g_i$ satisfy the conditions $e_{i_1}=e_{i_2}$ 4. $t\le n$

Which means the quivalent group set is an set of that each elements is a sublist from R, and each sublist contains records are equal. In this way, we map the list R into the set G

Counting Vector

For an equivalent group set, we more interested in the number of equal records in g_i

$$C(G)=[c(g_1),c(g_2),...,c(g_n)]. \tag{1}$$

Where the function $C(*)$ is a counting function which maps an equivalent group set G into a counting vector and each element in vector refer to the number of equal records in g_i .Counting vector only reflect the number of the same record, the dimensions of vector represent the number of unequal record groups of G .Therefore, Counting vector could map data with a variety of structure. It only concern about that if these record values are equal, and don't care what kind of structure these data have been.

In fact, Counting vector reflects how many data sources adopted the data with the same value, the size of each element in counting vector refer to the number of data sources that adopt data with specific value. For example, the frequency of a certain type of CPU, many E-commerce sites may give different frequency for this type of CPU .We could use a Counting vector to record the number of occurrences of different frequency.

Usually, when we examine the authenticity of a data, we not only examine how many data source showing these data, but also examine the weight of the data source itself. At the initial time, the weight of data source is given by the credit of data organization, as time goes on,the weight will change over time.

Evolutionary Operation and Dynamic Evolution Factor

Definition1:

$$R(T) = \max[C(G) \circ \bar{W}(G)] \Big/ \left\| \max[C(G) \circ \bar{W}(G)] \right\|_2 . \tag{2}$$

The Eq. 2 is an Evolutionary Operation, where G is equivalent group set. $\bar{W}(G)$ refers to the vector in which each element of it is a representative of the average weight . $C(G) \circ \bar{W}(G)$ refers to the Hadamard product of $C(G)$ and $\bar{W}(G)$. $R(T)$ is called evolutionary degree. $\max[C(G) \circ \bar{W}(G)]$ means choose the maximum elements in vector $C(G) \circ \bar{W}(G)$. The maximum elements are called EVO elements. When there are more then one equivalent maximum elements in the vector, they should merge into one. $\| * \|_2$ means 2-norm operation, we take 2-norm to normalize counting vector and average weight vector .

The EVO elements and it's evolutionary degree can be get through Evolutionary Operation, however there are no absolutly real data in the world, we can view the data updating process as a data evolution approaching an infinitely close approximation to an objective entity. The EVO elements are maximum number of the integration of quantity and weight. it can be considered as the most close to real data within a specified period.

The data in the reality is dynamic, we assume that the evolution of the data is caused by the following three factors

1. The proportion of EVO elements in the vector $[C(G) \circ \bar{W}(G)]$.

Definition

$$R(G) = \max[C(G) \circ \bar{W}(G)] \Big/ \left(C(G)^{\mathrm{T}} \bullet \bar{W}(G) \right). \tag{3}$$

Where the $R(G)$ refer to the proportion of EVO elements in the vector $[C(G) \circ \bar{W}(G)]$.

2. The proportion of threatening elements in the vector $[C(G) \circ \bar{W}(G)]$.

Definition

$$RS(G) = \max\{C(G) \circ \bar{W}(G) - \widetilde{\max}[C(G) \circ \bar{W}(G)]\} \Big/ \left(C(G)^{\mathrm{T}} \bullet \bar{W}(G) \right). \tag{4}$$

Where the $RS(G)$ refer to the proportion of second largest elements in the vector $[C(G) \circ \bar{W}(G)]$, function $\widetilde{\max}[*]$ means in addition to a maximum element in a vector, take other elements return to zero.

3. The span of equivalent data quantity change.

The transition probability of evolution

Definition

$$\begin{cases} P_{ij} = g[R(G_i), RS(G_i), |j-i|^{\alpha}] = f(i, j-i) \\ \sum_{j=1}^{n} P_{ij} = 1 \end{cases} . \tag{5}$$

Where the P_{ij} is said to be a transition probability of evolution from i to j. $g[R(G_i), RS(G_i), |j-i|^{\alpha}]$ is a function that relate to the three factors above. i, j refer to the number of the same record from a certain data source in different times. $\alpha \geq 0$ is a real number that to control the way of influence probability by span from i to j.Function $g[R(G_i), RS(G_i), |j-i|^{\alpha}]$ can be a variety of specific analytic expression, let $f(*)$ is a certain expression.

The whole evolution probabilities can be written as transition matrix form

$$P = \begin{pmatrix} f(1,0) & f(1,1) & f(1,2) & \cdots & f(1,n-1) \\ f(2,-1) & f(2,0) & f(2,1) & \cdots & f(2,n-2) \\ f(3,-2) & f(3,-1) & f(3,0) & \cdots & f(3,n-3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(n,1-n) & f(n,2-n) & f(n,3-n) & \cdots & f(n,0) \end{pmatrix}. \tag{6}$$

Suppose the probability of the data sources number that evolves from i to j in one interval is only related to the state i at the beginning of this interval, those evolutionary process satisfy markov property.

Setting a rules that it is a one step transition matrix, that is to say all the transition occur in a time interval. Let $\phi^{(0)}$ is an initial distribution vector in n-dimension space, than we can get the transfering probability distribution in any after any time intervals $\phi^{(x)}$. according to Chapman-Kolmogorov equation

$$\phi^{(x)} = \phi^{(0)} P^{x-1} . \tag{7}$$

In practice, Evolution Factor will change over time, that cause the variation of analytic expression in function $f(*)$.We suppose that in m intervals, the change are small, than , the $\phi^{(x)}$ can be written as Eq. 8

$$\phi^{(x)} = \phi^{(0)} P_1^m P_2^m \cdots P_{[x/m]-1}^m P_{[x/m]}^{x\%m} = \phi^{(0)} \left(\prod_{l=1}^{[x/m]-1} P_l^m \right) P_{[x/m]}^{x\%m} . \tag{8}$$

The $\phi^{(x)}$ is the probability distribution after in *x* times evolution, where the times of evolution *x* is greater than Stabilization intervals *m*. The P_l is the constant stochastic evolutionary matrix in the *l*th timescale. The $[x/m]$ is the round off function of x/m, $x\%m$ means taking the remainders of x/m.

Proof:

$$\begin{aligned} \phi^{(x)} &= \phi^{([x/m]-1)} P_{[x/m]}^{x\%m} = \phi^{([x/m]-2)} P_{[x/m]-1}^m P_{[x/m]}^{x\%m} = \phi^{(1)} P_2^m \cdots P_{[x/m]-1}^m P_{[x/m]}^{x\%m} \\ &= \overline{\phi}^{(0)} P_1^m P_2^m \cdots P_{[x/m]-1}^m P_{[x/m]}^{x\%m} \end{aligned}$$

we can see that if the analytic expression of transfer probability is defined, then the probability of any length of time and any transition span can be calculated, however, this kind of calculation is very difficult. We tried to simplify this operation, there are some commonality within each time interval.

If the limit distribution of each interval can be found, it can use to instead the high power matrix within each interval. In fact, in a discrete Markov chain, if the probability of data sources numbers is ergodic, the stationary distribution of transition probability distribution is equal to its limit distribution, it means that we can get $\lim_{m\to\infty}\phi^{(0)}P_i^m=\pi_i$ by $\pi_i P_i^m=\pi_i$. It can be found that the limit distribution do not dependent on the initial distribution. In this way can be simplify as following:

$$\lim_{m\to\infty}\phi^{(x)}=\lim_{m\to\infty}\underbrace{\underbrace{\underbrace{\phi^{(0)}P_1^m}_{\pi_1}P_2^m}_{\pi_2}\cdots P_{[x/m]-1}^m}_{\pi_{[x/m]-1}}P_{[x/m]}^{x\%m}=\pi_{[x/m]-1}P_{[x/m]}^{x\%m} \quad . \tag{9}$$

Sometimes, data evolution does not have ergodicity, its transition matrix is reducible, but according to the Eq. 9， it is only need to make sure that the previous period has a stationary distribution, the current limit distribution is independent of earlier distribution.

$$\lim_{m\to\infty}\phi^{(x)}=\underbrace{\phi^{(0)}P_1^mP_2^m}_{\phi^{[x/m]-2}}\cdots\lim_{m\to\infty}P_{[x/m]-1}^m{}^{x\%m}_{[x/m]}=\pi_{[x/m]-1}P_{[x/m]}^{x\%m} \quad . \tag{10}$$

In fact, When a EVO element was determined, the corresponding number i was decided. So, a nonzero row in transition probability of evolution Matrix is decided. At this point, the other rows in the matrix are 0, Therefore there is one, and only one nonzero row in the Evolutionary Matrix.

$$P\xrightarrow[i=i^*]{Evolutionary\ Operation}P_*$$

$$P_*=\begin{pmatrix}0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ g[R(i^*),RS(i^*),i^*-1] & g[R(i^*),RS(i^*),i^*-2] & \cdots & g[R(i^*),RS(i^*),i^*-n]\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{pmatrix}. \tag{11}$$

The whole collaborative forecasting process is as Fig.1.

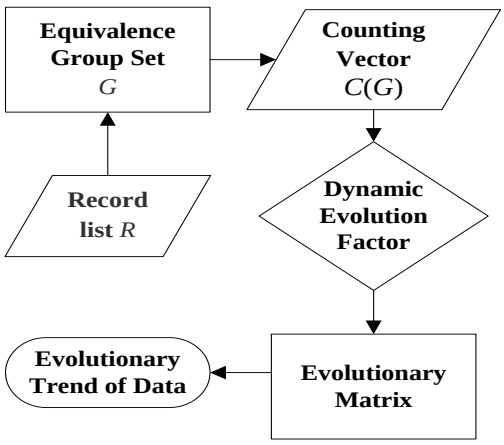

Fig. 1. Forecast flow chart

Table 1 SOURCES OF DATA SETS

Type	JD	YIXUN	YHD	AMAZON
Mobile	13140	12757	10611	12791
Digital	44324	69786	49108	59240
Computer	14102	19683	9799	13725
Appliance	96800	105379	66881	60497
Houseware	22284	35436	40482	25152
Jewelry	82452	68328	78481	62113
Cosmetics	25215	21326	15205	21742
Sports	5483	5567	2492	4532
Food	4164	6060	6285	3894
Dailyuse	17236	26409	11567	20686

Experiment

The experimental idea is collecting the commodity price data of e-commerce sites. Forecast the probability distribution of price movements by Evolutionary Matrix and limit distribution.

The experimental tool is R language in version 3.2.4. It running on Windows 7 64-bit operating system. The hardware configuration for AMD FX-8350 4.0GHz CPU， DDR3 1600 8GB RAM. The data set is commodity price from January 2014 to August 2015 period, those from four B2C platform, the number of every commodity type are shown in Table 1.

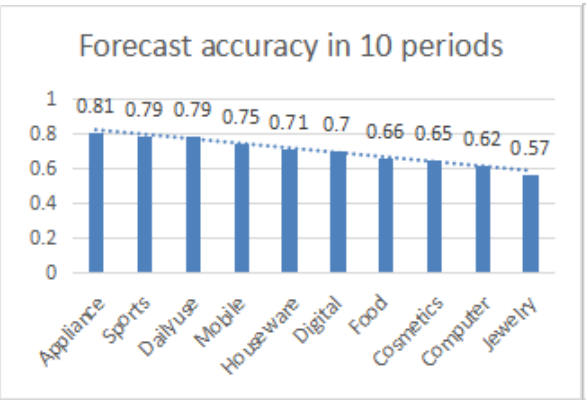

Fig. 2. Forecast accuracy in 10 periods

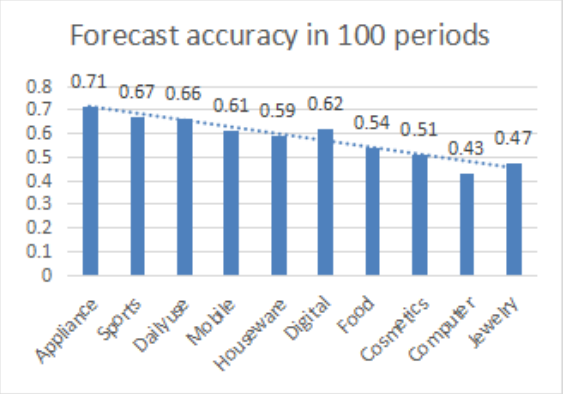

Fig. 3. Forecast accuracy in 100 periods

Conclusions

1.This paper gives a method of forecasting data evolution in a long term.

2.We analysis the limit properties of data evolution for the first time, and use the of limit properties of evolution to predict the behavior of data in long-term evolution.

3.We view the data updating process as a data evolution approaching an infinitely close approximation to an objective entity,that is a completely different perspectives for the past research.

4.Experiments show that after a great times of updates , it still has a high prediction precision for the initial price.

Acknowledgements

This work was financially supported by the National Natural Science Foundation of China （61502350）

References

[1] Wang Y, Liu S X, Feng J, et al. *Mining Naturally Smooth Evolution of Clusters from Dynamic Data*. Siam International Conference on Data Mining. 2007:125--134.

[2] Yu Fang, Chen Shengshuang Li Shijun Yu wei “Research on Evolution and Updating Among Multi-Source Data Based on Big Data”(in chinese) *Computer Science*, 2016, 12.

[3] John Odenckantz. "Nonparametric Statistics for Stochastic Processes: Estimation and Prediction." Technometrics 42.4(2012):429-430.

[4] Hedeler C, Belhajjame K, Fernandes A A A, et al. Dimensions of Dataspaces.*Dataspace the Final Frontier*, 2009, 5588:55-66..

[5] Panse, F., et al. "Duplicate detection in probabilistic data." IEEE, International Conference on Data Engineering Workshops IEEE, 2010:179-182.

[6] Panse, Fabian, M. V. Keulen, and N. Ritter. "Indeterministic Handling of Uncertain Decisions in Deduplication." Journal of Data & Information Quality 4.2(2012):1-25.