IJCSC International Journal of Computer Science & Communication Vol. 1, No. 2, July-December 2010, pp. 285-288

Sequence Forecast Algorithm Based on Nonlinear Regression Technique for Stream Data

Sonali Tiwari

Department of Computer Science and Engineering, Faculty of Engineering & Technology, Faridabad
Email: sonalitiwari78@gmail.com

ABSTRACT

Data mining is the process of extracting knowledge structures from continuous, rapid and extremely large stream data which handles quality and data analysis. In such traditional transaction environment it is impossible to perform frequent items mining because it requires analyzing which item is a frequent one to continuously incoming stream data and which is probable to become a frequent item. This paper proposes a way to predict frequent items using regression model to the continuously incoming real time stream data. By establishing the regression model from the stream data, it may be used as a prediction model to uncertain items. After gathering real-time stream data through sliding window, the proposed algorithm computes support for appointed sequence and describes non linear equation to forecast sequence trends in the future.

Keywords: Frequent Items; Stream Data; Non-linear Regression; Sliding Window

1. INTRODUCTION

Stream data is generated continuously in a dynamic environment, with huge volume, infinite flow, and fast changing behavior. There is a significant body of research on data stream query methodologies[11,12] and mining technologies, such as association rules, frequent patterns, clustering, classification, forecast, sequential analysis, both in the fields of data stream management system and knowledge discovery.

In recent years, many time-series mining problems have been explored for a data stream environment. Because sequence forecast is an import subject of those, some methods were designed to develop sequence trend analysis in a dynamic environment. In [13], Yixin Chen etc. investigated methods for multi-dimensional regression analysis of time-series stream data; they built up a partially materialized data cube model and took an exception-guided regression approach to analysis stream data. In [14], Wei-Guang Teng etc. devised a FTP-DS algorithm to mine frequent temporal patterns of data stream. The FTP-DS algorithm uses linear regression to perform trend detection, it's an effective method, but it always omits some exceptions, which are too pivotal to lead to failure.

In this paper, we present an algorithm that use nonlinear regression approach to ameliorate FTP-DS algorithm, so the sequence trends can cover those key exceptions. Our experiment on comparing between the FTP-DS algorithm and the proposed algorithm shows that coverage of proposed algorithm is much wider. The rest of the paper is organized as follows. In Section II, we introduce the basic concepts of sequence forecast and concerned problem, Section III describes the FTP-DS algorithm. In Section IV, we present the proposed algorithm based on non-linear regression technique. Section V shows our experiments and performance analysis and our study and future research is concluded in Section VI.

2. RELATED CONCEPTS

In this section, we introduce the basic concepts of sequence forecast in data stream and give a brief look at FTP-DS algorithm. The fundamental difference in the analysis in a dynamic environment from in a static one is that the former focused on trends analysis instead of static aggregation. Sequence in data stream is essentially a sequence of sets of events, which conform to the given timing constraints [15]. As an example, the sequential pattern (A) (C, B) (D), encodes an interesting fact that event D occurs after an event-set(C, B), which in turn occurs after event A. If the support of sequence is no less than the threshold, this sequence is called frequent sequence. Sequence forecast aims at displaying a trend curve to indicate the general direction of the appointed frequent sequence. Sequence forecast plays an important role in gene analysis, finance forecast, network security, web information extraction, communication statistic, and so on. Typical methods for sequence forecast includes regression method, weighted moving average method and the least squares method.

3. PREVIOUS WORK

A Brief Look at FTP-DS Algorithm

Algorithm FTP-DS was presented by Wei-Guang Teng on 29th VLDB conferences at 2003. There were two key definitions for FTP-DS algorithm. One is the support of frequent sequence; the other is the pattern representation that called a compact ATF. We will give a brief look at those in below.

The support of frequent sequence X at a specific time t is denoted by the ratio of the number of events having sequence X in the current sliding window to total number of events. Supposing the sliding window size is 3, there are 3 sliding windows in Figure1, i.e., [0,3], [1,4], [2,5], the support of sequence <bd> is computed as: in [0,3], the <bd> is in event {ID:2, ID:5}, so its support is 2/5=0.4; in [1,4], the <bd> is in event {ID:1, ID:4}, so its support is 2/5=0.4; in [2,5], the <bd> is in event {ID:1}, so its support is 1/5=0.2.

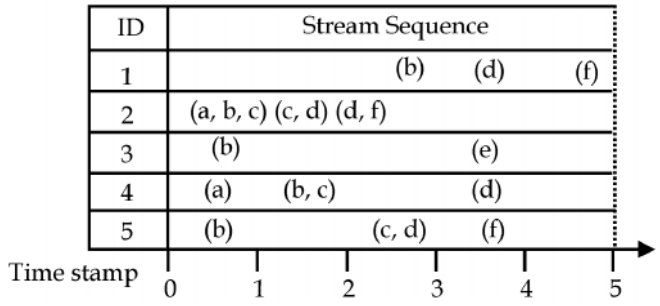

Fig. 1: An Example of Support Computation

The ATP form of time series was defined as

$$(t_s, \sum tf, \sum f, \sum f^2), \quad (1)$$

$$\hat{f} = \hat{\alpha} + \hat{\beta}_t, \quad (2)$$

and FPT-DS algorithm extracted a regression fit line: with the following equations:

$$S_{tt} = \sum t^2 - \frac{(\sum t)^2}{n}, \quad (3)$$

$$S_{ff} = \sum f^2 - \frac{(\sum f)^2}{n}, \quad (4)$$

$$S_{tf} = \sum tf - \frac{(\sum t)(\sum f)}{n}, \quad (5)$$

$$\hat{\beta} = \frac{S_{tf}}{S_{tt}} \quad (6)$$

$$\hat{\alpha} = \overline{f} - \hat{\beta}\overline{t} = \frac{\sum f}{n} - \hat{\beta} \times \frac{\sum t}{n} \quad (7)$$

4. PROPOSED WORK

Frequent Itemset(Sequence) Mining Scheme for Real Time Data Stream

As we know stream data is continuous and complex in time so it is only possible to access such data temporarily. Stream data has sequential characteristics that can be considered as time series data. Prediction of time series data gathers useful data estimating future through the analysis of data from the past. In the proposed method first the real time stream data is preprocessed to establish non- linear regression model. When the regression model is generated, prediction process on the possibility of frequent items is performed based on the regression model stream data is reorganized with the time and the support for each data is calculated X in the current sliding window to total number of events.

Algorithm

Step1: Input frequent sequence of real time stream data, window size N, and the support threshold.

Step2: Calculate Support which is no less than the threshold.

Step3: Here we take time as independent variable and support as dependent variable.

Step4: Calculate the co efficient of regression model: b_0, b_1 and b_2, Inputs are (xi, yi), $i = (1,2,3,4\ldots\ldots\ldots n)$ calculate the following variable $Y_1, Y_2, Y_3, X_1, X_2, X_3$ and X_4

$$Y_1 = (\Sigma y_i)/n, \quad Y_2 = (\Sigma x_i y_i)/n, \quad Y_3 = (\Sigma x_i^2 y_i)/n$$

$$X_1 = (\Sigma x_i)/n, \quad X_2 = (\Sigma x_i^2)/n, \quad X_3 = (\Sigma x_i^3)/n, X_4 = (\Sigma x_i^4)/n$$

$$b_2 = \frac{(Y_2 - X_1 Y_1)(X_3 - X_1 X_2) - (Y_3 - Y_2 Y_1)(X_2 - X_1^2)}{(X_3 - X_1 X_2)^2 - (X_4 - X_2^2)(X_2 - X_1^2)}$$

$$b_1 = \frac{(Y_2 - X_1 Y_1) - b_2 (X_3 - X_1 X_2)}{(X_2 - X_1^2)}$$

$$b_0 = y_1 - b_1 x_i - b_2 x_i^2$$

Step5: Fit the non-linear regression model in the equation:

$$\hat{Y}_i = b_0 + b_1 x_i + b_2 x_i^2 \ldots\ldots n + \varepsilon$$

b_0, b_1, b_2 are the regression coefficients of non linear regression model.

Step6: Calculate the error

$$\varepsilon_i = Y_i - \hat{Y}_i$$

5. EXPERIMENT AND EVALUATION

In C++ programming environment with Microsoft windows-XP in support and also the MS Excel 2003 for plotting the graphs, we performed the experiments and evaluated the results in terms of errors.

We took time as independent variable (or predictor variable) and support as dependent variable (or response variable). We perform the preprocessing of data stream using sliding window and calculated the support (or occurrence frequency). Using these two variable values we calculated the regression coefficients b0 and b1, to create a model of linear regression for forecasting of frequent items.

Here, in this mining scheme we found that the errors or the residuals that we encounter are little less as compare to the residuals of the algorithm FTP-DS. The comparisons of errors are shown below in the form of line graphs. The residual (or error) comparison has been done in FTP-DS algorithm applying linear regression and nonlinear (or parabolic) regression technique.

The standard error estimate of Y on X

$$S_{Y,X} = \frac{\sqrt{(Y_i - \hat{Y}_i)^2}}{\sqrt{N}}$$

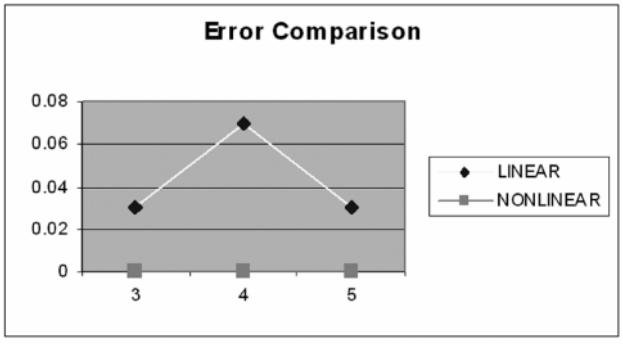

Fig. 2: Graph to Show the Residual Comparison

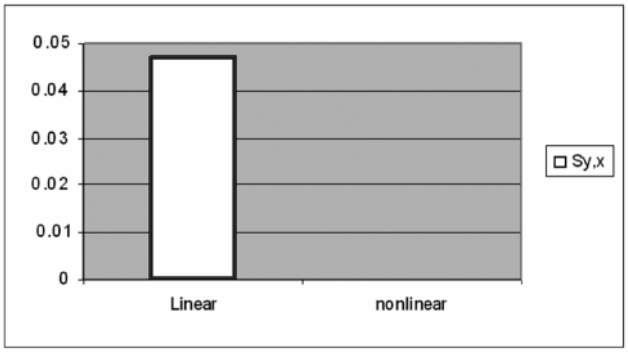

Fig. 3: Graph to Show the Standard Error Estimate Comparison

6. CONCLUSION AND FUTURE SCOPE

In this method, the least square method is used to estimate a regression line based on the nonlinear regression analysis. Consisted of continuously generating transactions. The regression in this method is a nonlinear regression model estimating a regression model with changes in two different variables. Since stream data has a characteristic of being continuously transmitted in time, they define the changes in time with two variables. These two variables indicate an independent variable and a dependent variable respectively.

Here, in this method I have presented an overview of some vulnerabilities of algorithm FTP-DS and designed nonlinear regression based algorithm to predict frequent sequence for real time stream data, in which I reviewed the concept of preprocessing from the algorithm FTP-DS. Our experimental result shows the merit of our algorithm in terms of errors.

Let us analyze a few open problems of the proposed algorithm from data stream background. Stream data is so huge, infinite, and fast changing that it always contains lots of noisy, overflowing, incomplete data items. As my work is concerned, it is not dealing with this problem that too at the cost of execution time, my future work is to resolve this problem.

REFERENCES

[1] B. Babcock, S. Babu, M. Datar, R. Motwani, and J. Widom, "Models and Issues in Data Stream Systems," *In Proc. of PODS*, March 2002.

[2] N. Davey, S. P. Hunt, and R. J. Frank, "Time Series Prediction and Neural Networks," *In Journal of Intelligent and Robotic Systems*, 2001.

[3] C. Giannella, J. Han, J. Pei, X. Yan, and P. S. Yu, "Mining Frequent Patterns in Data Streams at Multiple Time Granularities," In H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yeshar (eds.), *Next Generation Data Mining*, AAAI/MIT, 2003.

[4] L. Golab, M. Tamer Ozsu, "Issues in Data Stream Management," *In SIGMOD Record*, **32**, Number 2, 2003.

[5] X. Hao, D. Xu, "Time Series Prediction based on Non-Parametric," *In SIGMOD Record*, **32**, Number 2, 2003.

[6] S. Sarkka, A. Vehtari, and J. Lampinen, "Time Series Prediction by Kalman Smoother with Cross-Validated Noise Density," *In Proc. of IJCNN*, pp.1653-1658, 2004.

[7] D. F. Specht, "A General Regression Neural Network," *IEEE Trans. on Neural Networks*, **2**, No.6, pp.568-576, Nov. 1991.

[8] D. Wacherly, W. Mendenhall, R. L. Scheaffer, "Mathematical Statistics with Applications", 5th Edition, 2005.

[9] O. B. Yaik, C. H. Yong, and F. Haron, "Time Series Prediction using Adaptive Association Rules," *In Proc. Of DFMA05*, pp.310-314, 2005.

[10] Duck Jin Chai, Eun Hee Kin, Long Jin, Faridabad "Frequent Items Prediction Method to one Dimensional Stream Data ", *IEEE -2007*, Fifth International Conference on Computational Science and Applications.

[11] S.Guha and N.Koudas, "Approximating a Data Stream for Querying and Estimation: Algorithms and 16th ICDE Conference, Florida, pp.3 14, March 2002.

[12] M.Datar, A.Gionis, P.Indyk et, al, "Maintaining Stream Statistics Over Sliding Window", *In ACM-SIAM Symposium on Discrete Algorithms (SODA)*, Chicago, pp.406-417, June 2002.

[13] Y.Chen, G.Dong, J.Han et,al, "Multi-Dimensional Regression Analysis of Time-Series Data Stream", *Proceedings of the 28th International Conference on Very Large Data Base*, Berlin, pp. 323-334, August 2002.

[14] Wei-Guang Teng, Ming-Syan Chen, Philip S.Yu, "A Regression-Based Temporal Pattern Mining Scheme for Data Stream", *Proceedings of the 29th International Conference on Very Large Data Base*, Berlin, pp.607-617, August 2003.

[15] Joshi M, Karypis G, "A Universal Formulation of Sequential Patterns", Research Report No.99-021, *Department of Computer Science*, University of Minnesota, Minnesota, 1999.