IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-9 E-ISSN: 2347-2693

Regression Based Data Mining Techniques for Frequent Data Stream
(One Dimensional and two Dimensional Stream Data)

Pinki Sagar

Dept. of Computer Science and Engineering,
Manav Rachna International University, INDIA,

www.ijcseonline.org

Received: Sep /07/2015 Revised: Sep/14/2015 Accepted: Sep/26/2015 Published: Jan/30/ 2015

***Abstract*—** Data mining in the stream data handles quality and data analysis using extremely large and infinite amount of data and disk or memory with limited volume[2]. 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 analyze a way to predict frequent items using linear regression model[5] to the continuously incoming one dimensional stream data like the time series data. By establishing the regression model from the stream data, it may be used as a prediction model to uncertain items. The proposing way will exhibit its effectiveness through experiment in stream data.

Keywords—*Data mining, Time Series Data, Regression Techniques, Stream Data*

I. Time Series Data Prediction

Time series refers to data observed with fixed time interval about one or several incidents [3]. Stream data is also the data collected with fixed time interval in time as time series data. The examples of the latter include the daily fluctuating composite stock exchange index, monthly sales of certain consumer goods, and annual production of the crops. Since these time series data are historical series displaying changes in time on certain economical or natural phenomena, time series observed at one point usually depends on its previous data. Therefore, forecasting through time series analysis analyzes previously observed data, finds a rule, model it and forecast a rule in the future.

II. Frequent Item Prediction Method(single dimesional stream data)

Chai, Eun Hee Kim and Long Jin[2] proposed a method FIPM (frequent item prediction method) which is used to predict frequent items. Using linear regression model. It can be used as a prediction for the stream data. Stream data has nature of continuous and infinite. They proposed a method(fipm) that can predicts frequent items using simple linear regression method for one Dimensional stream data. it is only possible to access stream data temporarily because Stream data is continuous and complex in time . 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 FIPM method first one dimensional stream data is preprocessed to establish simple 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 in which each one dimensional data is inputted .In the rearranging of data we collect the input time of same data and calculate the difference in time at which data is accrued. In the next step pairing is to be done with time which is calculated from the rearranging or reorganization of data .

III. Regression Model Estimation

Regression analysis is a method to predict an output based on an input value upon finding a functional relationship between values with given data. The analysis on the linear relationship between an input and output values is known as the linear regression analysis. When there are two or more input values, it is known as the multi regression analysis. In this paper, we estimate the regression model using the simple Linear regression analysis.

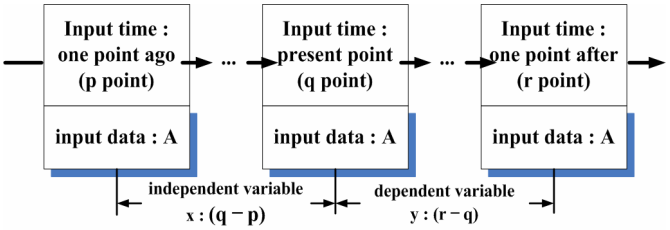

Figure 1: Ordered pair between an independent variable and a dependent variable in the example of the value, A

In this paper, two variables ar defined x and y , these are independent and dependent variables. Value of dependent and independent variable are calculated as follows:
P: input time at one point ago before current time

Q: input at one point after the current time.

Variable x(independent variable): difference of value p and q (q-p), (as shown in figure 1) means that difference of time between current time and one point ago before current time.

Variable y(dependent variable): difference of value q and r (r-q) (as shown in figure 1 means that difference of time between current time and one time after current time(future time).

Now We are able to estimate the regression model from the differences with a time in the before and a time in the after. Prediction of support value of an item in the future by estimating the time difference between a time (one point ago and one point after)in using the regression model for the relation of independent variable and dependent variable in the above

IV. Method of FIPM

Step1: Reorganize the stream data according to the occurrence of particular time.

Step2: Calculate the time difference between the previous time and next time of variable when the data arrived.

Step3: create the pairs from the array of difference of time for example:

(x1,x2,x3,x4): {x1andx2},{x2and x3),{x3andx4}

(a1,a2,a3,a4) : (a1,a2),(a2,a3),(a3,a4)

Variable of x independent variable and y dependent variable

X: a1,a2,a3

Y: a2,a3,a4

Step4: calculation of co efficient of regression model of FIPM b0 and b1

$$b_0 = \bar{y} - b_1 \bar{x} \qquad (1)$$

$$\hat{b}_i = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sum x_i^2 - n\bar{x}^2} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \qquad (2)$$

Step5: fit the regression model of FIPM:

$$\acute{y} = b_0 + b_1 x + E \qquad (3)$$

Step6: calculate the error

$$E_i = y_i - \acute{y}_i \qquad (4)$$

V. Preprocessing of Stream Data in FIPM

In FIPM[2] we have two type of variable Dependent and independent variables. For example A is accruing at time interval of 2(T15-T13),3(t13-t10) and s on in figure (3) .Independent variable is calculated by(Input time at Present – input time at one time ago)and dependent variable is calculated as (Input time after one point- input time at one time ago). In the regression model in the preprocessing we put the value of dependent and independent variables in the linear regression model, Using linear regression model and preprocessing method. Since stream data has a characteristic of being continuously transmitted in time, we define the changes in time with two variables. These two variables indicate an independent variable and a dependent variable respectively. The independent variable means the time interval between a time in the past and the present time whereas the dependent variable indicates the time interval between a time in the future to be entered and the present time. When a value is entered, its entry point in the future may be predicted by estimating a regression model from the relationship between these two variables. Therefore, the support of stream data item may be calculated by using this regression model. Frequent item mining methods of stream data must be researched using statistical methods egression model from one dimensional stream data and used it; however, the regression model must be corrected considering the distribution of data changing in time for the accuracy of the model. Therefore, a method on changing a regression model gradually in regards to the distribution of stream data must be studied as well.

Example:

VALUES	TIME	VALUES	TIME
B	T16	E	T7
A	T15	D	T6
B	T14	B	T5
A	T13	A	T4
C	T12	C	T3
E	T11	D	T2
A	T10	E	T1
C	T9	A	T0
A	T8		

Figure 2: example of one dimension stream data

in this example A is appearing continuous at t-15,t-13 ,t-10,t-8,t-4, and at t-1 in figure(2)

Reorganization of the sample stream data

I/P VALUES	INCOMING TIMES
A	T15,T13,T10,T 8,T4,T0
B	T16,T14,T5
C	T12,T9,T3
D	T6,T2
E	T11,T7,T1

Figure 3: Reorganization of stream data

time interval for the input of A from the time of Previous to the present time: t: {2, 3, 2, 4, 4}

Pairs are: {2and 3},{3and 2},{2 and 4},{4 and 4}and so on. Arrange the variables x(independent) and y(dependent) in to form of table figure(3)

I/P TIME SEQUENCE	1	2	3	4	5	6	7	8	9
INDEPENDENT VARIABLE(X)	2	3	2	4	4	3	5	7	6
DEPENDENT VARIABLES(Y)	3	2	4	4	3	5	7	6	5

Figure 4: Arrange the pairs in the form(x and y)in the table.

VI. Processing of Sequence Forecast Algorithm on Plane Regression algorithm

FPTDS[3] is used for time series stream data in the training data is presented according to their time and ids for example <ab> stream data sequence is appearing at id 2,4 and 9, time is t0 to t2 in figure(5), and we can analyzed Other stream data according to their presence. In this type of stream data we can analyze the frequency of appearing of particular stream data sequence.

id										
1	a		b		c	b	d	e	f	b
2		ab	d	c		d	e		f	b
3							b	c	d	
4	a	b	c	d	e	f				
5	c	d		a	d	f	f	g	b	d
6		b	c	d	e	f	g	h	b	d
7	c	d			b	d	f	g	c	d
8		a	c	d		b	d	e	f	g
9	a	b	c	d	e	f	a	b	d	
10		a	b	d	c	d	e	f	g	
0	1	2	3	4	5	6	7	8	9	10

Time---------->

Figure5: Training data for SFA-**PR**

VII. Preprocessing of training data for Sequence Forecast Algorithm on Plane Regression algorithm

Calculate the support sequence for particular data according to their ids, time and sliding windows for any specified stream data sequence. Sliding windows are 0-3, 1-4, 2-5, 3-6 and so on. Support or actual F(frequency for appearing the stream data) is calculated using the following method.

F= Num of id's at which stream data is presented
Total number of id's of stream data

VIII. Method of Sequence Forecast Algorithm on Plane Regression algorithm

With the help of preprocessing of training data, we go it the frequency as independent variable and time as an independent variables, and then For SFA-PR We calculate the coefficient using preprocessing and fit the regression model. For calculating the coefficients and regression model we used the various symbols:

$\sum f$: Sum of all frequencies (or support) Dependent variables (1,2----n)

$\sum t$: Sum of all times independent Variables (1, 2 -------n)

$\sum tf$: Sum of multiplication of time and Frequencies (1, 2, -------n)

For calculating the regression model we use the following equations so that we can predict the frequency at which data is appearing.

SFA-PR the regression model is:

$$Y= b_0+\mathbf{b_1}t+b_2f+\mathcal{E} \quad (5)$$

There exists constant n and matrix Y, X, and β, let

$$Y= \begin{bmatrix} YI \\ Y2 \\ . \\ . \\ YN \end{bmatrix} \quad B= \begin{bmatrix} B0 \\ B1 \\ .B2 \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & t1 & f1 \\ 1 & t2 & f2 \\ 1 & tn & fn \end{bmatrix}$$

We can have following forecasting equations:

$$Y= Xb + \mathcal{E} \quad (6)$$

then, the regression **plane** can be

$$B = (x`x)^{-1} x`y \quad (7)$$

$$X`x`= \begin{bmatrix} n & \sum t & \sum f \\ \sum t & \sum t^2 & \sum tf \\ \sum f & \sum tf & \sum f^2 \end{bmatrix}$$

And

$$X`y= \begin{bmatrix} \sum y \\ \sum y \\ \sum fy \end{bmatrix}$$

IX. Analysis of FIPM and SFA-PR

FIPM and SFA-PR both algorithms are used in prediction for the stream data (continuous data)

Analysis of prediction curve: In the figure (6) it is to be analyzed that prediction curve for the SFA-PR is near to linear regression curve line. In the case of FIPM prediction curve is not frequent. It is clear that prediction for stream data in SFA-PR is better than prediction using FIPM.

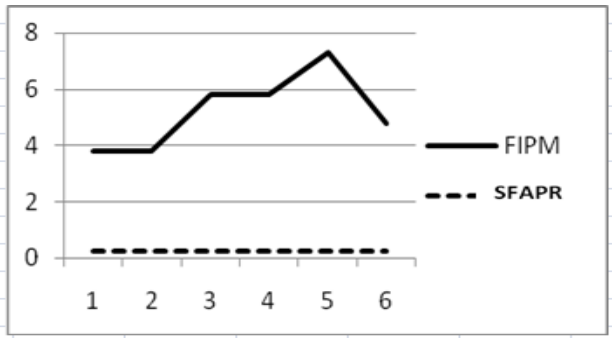

Figure6: Prediction curve for FIPM and SFA-PR

X. Analysis of error curve

Errors are generated during the prediction of stream data in SFA-PR is more less than the errors ,which are generated during the prediction of stream data in the FIPM.

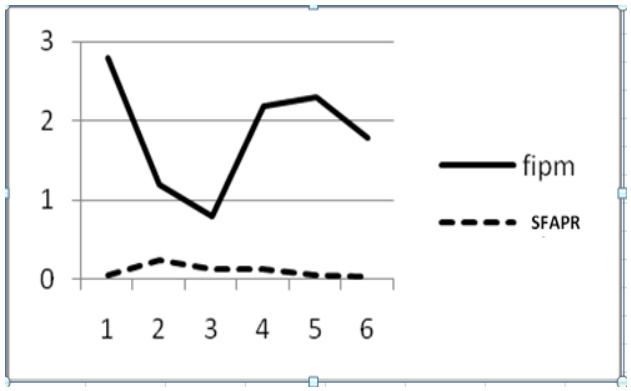

Figure7: Error curve for FIPM and SFA-PR

XI. Conclusion

FIPM[2] and SFA-PR[3]both algorithms are used for mining from stream data .both are based on regression technique. Each technique has its own importance. But according to analysis of FIPM and SFA-PR it is to be found that Prediction is very high for stream data in SFA-PR means that we can find that what are the frequencies for the appearing of stream data sequence. Execution time for the SFA-PR is very low in comparison of FIPM. Error detection is very high in SFA-PR in comparison of FIPM. In the end it is to be analyzed that SFA-PR is much efficient than FIPM.

References

[1] D.F. Andrews, :A robust method for multiple linear regression,**T***echnometrics* , vol 16, **1974**, pp 125 - 127.

[2]Chai, Eun Hee Kim and Long Jin**:pr**ediction of Frequent Items to OneDimensional Stream Data; Fifth International Conference on Computational Science and Applications ; page 353-360, **2001**

[3]Y. Chen, G.Dong, J.Han, B.W.Wah, and J.Wang **:** .Multi-Dimensional Regression Analysis of Time- Series Data Streams; Proc. Int. Conf. Very Large Data Bases;Hong Kong, China, Aug. **2002**.

[4]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**.

[5]R. Hayward; A Basic Approach to Linear Regression; RWJ linical Scholars Program; pp1-3,University of Michigan , **2005**.

[6]O.B.Yaik, C.H.Yong, and FHaron, Time Series Prediction using Adaptive Association rules,InProc.of DFMA05, pp.310-314, **2005**.

[7]Omid Rouhani-Kalleh; Algorithms for Fast Large Scale data Mining Using Logistic Regression; Proceedings of the 2007 IEEE Symposium on Computational Intelligence and Data Mining; pp 155-162, **2007**.

[8]Feng Zhao, Qing-Hua A Li **:**A Plane Regression Based Sequence Forecast Algorithms for Stream Data ; Proc. of the Fourth International Conference on Machine Learning and Cybernetics; pp-1559-1562 Guangzhou,18-21 August, **2005**.

[9]Y. Peng, G. Kou, Y. Shi, Z. Chen; A Descriptive Framework for the Field of Data Mining and Knowledge Discovery. International Journal of Information Technology and Decision Making, Volume 7, Issue 4: 639 – 682; **2000**

[10] Perlich, C,Provost, F., Simonoff, J. S. Tree Induction verses. Logistic Regression:A Learning-Curve Analysis. Journal of Machine Learning Research Vol. 4 pp-211- 255. **2003**.

[11]Amir Bar-Or, Daniel Keren, Assaf Schuster, and Ran Wolff: Hierarchical Decision Tree Induction in istributed Genomic Databases; IEEERANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 17;pp; 1138- 1150,**2007**.

[12]Qi Luo; Advancing Knowledge Discovery and Data Mining; Workshop on Knowledge Discovery and Data Mining pp;3-5, **2008**.

[13]Fayyad, Usama; Gregory Piatetsky-Shapiro, and adhraic Smyth; From Data Mining to Knowledge Discovery in Databases. -pp:12-17, June **2008**.