TIME SERIES PREDICTION USING INCREMENTAL REGRESSION

Sung Hyun Kim, Yongmi Lee, Long Jin, Duck Jin Chai, Keun Ho Ryu

Database/Bioinformatics Laboratory, Chungbuk National University, Korea
{hyun, ymlee, kimlyong, djchai, khryu}@dblab.chungbuk.ac.kr

ABSTRACT ... Regression of conventional prediction techniques in data mining uses the model which is generated from the training step. This model is applied to new input data without any change. If this model is applied directly to time series, the rate of prediction accuracy will be decreased. This paper proposes an incremental regression for time series prediction like typhoon track prediction. This technique considers the characteristic of time series which may be changed over time. It is composed of two steps. The first step executes a fractional process for applying input data to the regression model. The second step updates the model by using its information as new data. Additionally, the model is maintained by only recent data in a queue. This approach has the following two advantages. It maintains the minimum information of the model by using a matrix, so space complexity is reduced. Moreover, it prevents the increment of error rate by updating the model over time. Accuracy rate of the proposed method is measured by RME(Relative Mean Error) and RMSE(Root Mean Square Error). The results of typhoon track prediction experiment are performed by the proposed technique IMLR(Incremental Multiple Linear Regression) is more efficient than those of MLR(Multiple Linear Regression) and SVR(Support Vector Regression).

KEY WORDS: Time Series Prediction, Incremental, Regression, Data Mining

1. INTRODUCTION

Data mining tasks are generally divided into two major categories: Predictive tasks and Descriptive tasks. There are two types of predictive modeling tasks: classification, which is used for discrete target variables, and regression, which is used for continuous target variables[1]. Regression is a predictive modeling technique where the target variable to be estimated is continuous[1]. Example of applications of regression include predicting a stock market index using other economic indicators, forecasting the amount of precipitation in a region based on characteristics of the jet stream, projecting the total sales of a company based on the amount spent for advertising. Regression of conventional prediction techniques in data mining uses the model which is generated from the training step. This model is applied to new input data without any change.

Like typhoon track time series method uses the past data for prediction. If the conventional regression is applied directly to time series, the rate of prediction accuracy will be decreased because characteristic of time series may be changed over time.

This paper proposes an incremental regression for time series prediction. This technique considers the characteristic of time series which may be changed over time. It is composed of two steps. The first step executes a fractional process for applying input data to the regression model. The second step updates the model by using its information as new data. Additionally, the model is maintained by only recent data in a queue. It maintains the minimum information of the model by using a matrix, so space complexity is reduced. The typhoon track prediction experiment is performed by the proposed technique IMLR and MLR, SVR. Accuracy rate of these methods is measured by RME and RMSE.

2. RELATED WORKS

As instance based learning[2-6], incremental modeling will update the model when new data are input.

[2] introduced the concept of instance based learning for classification through the use of stored examples and nearest neighbor techniques. The idea of instance based prediction of real-valued attributes was introduced by [3]. It describes an approach which uses a form of local linear regression. This idea of local linear regression was explored more detailed in [4], but no effort is made to limit the growth of the stored database. [5] discussed two major classes of LWL(Locally Weighted Learning), memory-based LWL and purely incremental LWL that do not need to remember any data explicitly. [6] introduced the algorithm that approximated safely to the value function for continuous state control tasks, and learnt quickly from a small amount of data. HEDGER is the instance based learning algorithm, based on LWR. The variation of standard linear regression techniques is training points which close to the query point have more influential over the fitted regression surface than those further away.

Incremental modeling about time series was studied by [7]. This technique performed prediction by using sequencial bayesian evolutionary. Its prediction is based on model of previous step and the current model is evaluated when new data are input. The current model is replaced if new model is better.

History data must be maintained to update the model in conventional study. The cost of updating model is much

and it is difficult to find the best updated point. This paper proposes the efficient technique for updating model that needs only the maintenance of recent data.

3. MULTIPLE REGRESSION

Linear regression analyzes the linear relationship between two variables, input variable and target variable. It is multiple regression if input variable is over 2. Formula (1) is multiple linear regression(MLR) which input variables is k.

$$y = b_0 + b_1 x_1 + \cdots + b_k x_k + \varepsilon \quad (1)$$

Variable y is predicted by variables x in the function as formula (1). Coefficient b is computed by method of least squares. It shows in formula (2). Variables x and variable y is transformed into the matrix by formula (2-a). B is computed by using the *transposed matrix* of X and *inverse matrix* by formula (2-b). B is the set of b.

(a) Matrix of variables x, variable y

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (2)$$

(b) Compute B

$$B = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = (X^T \times X)^{-1} \times X^T \times y$$

4. TIME SERIES PREDICTOIN

4.1 Fractal

General regression is the technique for predicting target variable by input variables. Hence, input variables and target variable need to make the model and prediction. Time series in this paper assume single variable. Time series needs the preprocessing to be applied to regression. Single variable is divided into many variables by fractal.

Figure 1 shows how to divide input variable into three variables by time t. Tuple is composed of current step(n), one previous(n-1) step and two previous(n-2) steps. For example, current tuple(n) indicates the value of current step to be divided. The future such as x_{n+1} is predicted by past data such as x_n, x_{n-1}, x_{n-2}. Because, the future value is predicted by the past value in time series. So we must to find the relational function between the future data and the past data. Formula (3) represents MLR model in formula (1) which is transformed into the time series form.

Input data → x : <n, n-1, ..., 2, 1, 0>

t	x_n	x_{n-1}	x_{n-2}
n	n	n-1	n-2
n-1	n-1	n-2	n-3
n-2	n-2	n-3	n-4
n-3	n-3	n-4	n-5
n-4	n-4	n-5	n-6
⋮	⋮	⋮	⋮

Figure 1. Example of Fractal

$$\begin{aligned} & y = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4 + b_5 x_5 + \varepsilon \\ & \downarrow \\ & y \rightarrow x_{n+1},\ x_1 \rightarrow x_n,\ x_2 \rightarrow x_{n-1} \\ & \downarrow \\ & x_{n+1} = b_0 + b_1 x_n + b_2 x_{n-1} + b_3 x_{n-2} + b_4 x_{n-3} + b_5 x_{n-4} + \varepsilon \end{aligned} \quad (3)$$

The number of variable(L) to be divided must be in which selected. A selection method is described following. N and threshold are user definition parameters.

1. L is repeated from 1 to N(maximum)
2. Compute training error rate(E)
3. L is chosen if E is lower than threshold

4.2 Incremental Multiple Linear Regression

Regression of conventional prediction techniques uses the model which is generated from the training step. This model is not changed. If conventional regression is applied directly to time series then the rate of prediction accuracy will be decreased because characteristic of time series may be changed over time.

This paper proposes an incremental multiple linear regression(IMLR) for time series prediction. This technique considers the characteristic of time series which may be changed over time. Additionally, the model is maintained by only recent data in a queue. For example, if queue size is 10, the tenth input data will be inserted in the model and almost all of old first data are eliminated in the model. Hence, the model always is made by ten recent data. High predicting accuracy is expected because the model is made by recent data. IMLR is proposed for updating formula (2-b) to decrease cost. This technique considers an insertion data and deletion data for a model updating. IMLR reduces a space complexity because matrixes $X^T X$, $X^T y$ have fix size. Formula (4-a) shows how to transform initial data to the matrix X, *transposed matrix* X^T of matrix X, matrix y. Formula (4-b) shows the multiplication result of matrix X^T and matrix X, matrix X^T and y. Formula (5) shows a

model updating formula. q is the size of queue and $(n\text{-}q)$ indicates the first data of queue.

(a) Matrix X^T, X, y

$$X^T = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ (x_1)_1 & (x_1)_2 & \cdots & (x_1)_n \\ (x_2)_1 & (x_2)_2 & \cdots & (x_2)_n \\ (x_3)_1 & (x_3)_2 & \cdots & (x_3)_n \\ (x_4)_1 & (x_4)_2 & \cdots & (x_4)_n \\ (x_5)_1 & (x_5)_2 & \cdots & (x_5)_n \end{bmatrix}$$

$$X = \begin{bmatrix} 1 & (x_1)_1 & (x_2)_1 & (x_3)_1 & (x_4)_1 & (x_5)_1 \\ 1 & (x_1)_2 & (x_2)_2 & (x_3)_2 & (x_4)_2 & (x_5)_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & (x_1)_n & (x_2)_n & (x_3)_n & (x_4)_n & (x_5)_n \end{bmatrix} \quad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

(b) $X^T \times X$, $X^T \times y$ (4)

$$X^T X = \begin{bmatrix} z_{11} & z_{12} & z_{13} & z_{14} & z_{15} & z_{16} \\ z_{21} & z_{22} & z_{23} & z_{24} & z_{25} & z_{26} \\ z_{31} & z_{32} & z_{33} & z_{34} & z_{35} & z_{36} \\ z_{41} & z_{42} & z_{43} & z_{44} & z_{45} & z_{46} \\ z_{51} & z_{52} & z_{53} & z_{54} & z_{55} & z_{56} \\ z_{61} & z_{62} & z_{63} & z_{64} & z_{65} & z_{66} \end{bmatrix}$$

$$X^T y = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \\ c_5 \\ c_6 \end{bmatrix}$$

• Model updating formula

$$X^T X = \sum_{i=1}^{k}\sum_{j=1}^{k}(X^T X[i,j] + (x_{n+1}[i] \times x_{n+1}[j]) - (x_{n-q}[i] \times x_{n-q}[j]))$$
$$X^T y = \sum_{i=1}^{k}(X^T y[i] + (x_{n+1}[i] \times y_{n+1}) - (x_{n-q}[i] \times y_{n-q})) \quad (5)$$

Inverse matrix computation needs to get matrix B. Cholesky LU decomposition[8] is used for *inverse matrix* computation. Cholesky LU decomposition is convenient when matrix is symmetric, positive definite. In this case, matrix $X^T X$ is symmetric, positive definite because it is multiplication result of matrix X^T and matrix X. Formula (6) shows *inverse matrix* computation procedure.

$$\begin{aligned} &X^T \times X = M : \text{symmetric, positive definite} \\ &M = L \times L^T : \text{L is lower triangular matrix} \\ &(L \times L^T) \times x = b, \quad L^T \times x = y \\ &\text{Compute } y \text{ by } L \times y = b \\ &\text{Compute } \textit{inverse matrix } x \text{ by } L^T \times x = y \end{aligned} \quad (6)$$

Table 1 shows IMLR algorithm.

Table 1. Incremental Multiple Linear Regression Algorithm

Algorithm IMLR
input: time series(xadd[], yadd, xout[], yout)
output: coefficient of function(b)
Begin
Step 1: apply input data xadd[] and first data xout[] to matrix $X^T X$
For i = 1 to r Do // r = row size of $X^T X$
For j = 1 to c DO // c = column size of $X^T X$
matrix[i][j] += xadd[i] * xadd[j];
matrix[i][j] -= xout[i] * xout[j];
End of inner For
End of outer For
Step 2: apply input data xadd[], yadd[] and first data xout[], yout to matrix $X^T y$
For i = 1 to r Do // r = row size of $X^T y$
ymatrix[i] += xadd[i] * yadd;
ymatrix[i] -= xout[i] * yout;
End of For
Step 3: compute inverse matrix of $X^T X$ (Cholesky LU decomposition)
inverse[][] = LU(matrix[][]);
Step 4: compute b (coefficient of function)
For i = 1 to r Do // r = row size of inverse
For j = 1 to c DO // c = column size of inverse
beta[i][j] += inverse[i][j] * ymatrix[j];
End of inner For
End of outer For
End

5. EXPERIMENTAL RESULT AND EVALUATION

The past typhoon data[9] was used in the comparison experiment of IMLR. Which is 2005-4 NESAT. It has 49 samples. Data were observed to 6 hour intervals. The typhoon track is predicted by the predicting of latitude, longitude and pressure.

Latitude, longitude and pressure of typhoon data were applied to MLR, SVR and IMLR. 10 data were used for training the model and 39 data were used for testing the model. Number of variable is three. mySVM[10] was used for SVR experiment. RME, RMSE was used for an error rate measurement. Formula (7) shows RME and RMSE.

$$RME = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - y^*_i}{y_i}\right| \qquad RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y^*_i)^2} \quad (7)$$

The result of typhoon track prediction experiment which is performed by the proposed technique IMLR is more efficient than those of MLR and SVR. Table 2 shows results of experiment.

Table 2. Result

Methods \ Data		NESAT		
		Latitude	Longitude	Pressure
MLR	RME	0.0826	0.0106	0.0041
	RMSE	22.9628	15.545	5.2285
SVR	RME	0.0127	0.0048	0.0039
	RMSE	4.2361	8.1920	4.8677
IMLR	RME	0.0102	0.0026	0.0034
	RMSE	3.4808	6.3365	4.1698

The latitude predicting result, of MLR has 8.26% relative error rate, of SVR has 1.27% relative error rate. Error rate of MLR is higher than different techniques because the model of MLR is not changed. Moreover, IMLR which has 1.02% relative error rate against SVR is better in performance. IMLR prevents the increment of error rate by updating the model over time. The result of longitude, pressure is similar to those of latitude as show in table 2. Figure 2 shows actual track of typhoon NESAT and predicting track by IMLR. The track for predicting is almost correct.

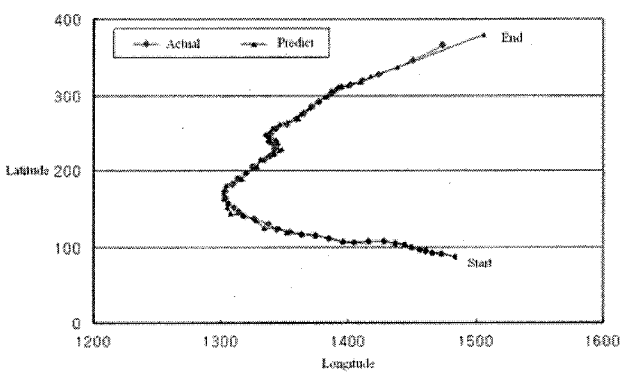

Figure 2. actual track VS predicting track

6. CONCLUSION

Applying time series to conventional regression is undesirable because characteristic of time series may be changed over time. This paper proposes the incremental multiple linear regression(IMLR) for time series prediction. IMLR prevents increment of error rate by updating the model over time. Moreover, IMLR maintains the minimum information of the model by using a fixed matrix, so space complexity is reduced. The result of typhoon track prediction experiment is performed by the proposed technique IMLR is more efficient than different techniques. IMLR can be applied to the domain which the characteristic of data may be changed over time.

REFERENCES

[1] P. N. Tan, M. Steinbach, and V. Kumar, (2005). *INTRODUCTION TO DATA MINING*. Addison Wesley.

[2] D. W. Aha, D. Kibler, and M. K. Albert, (1991). Instance-Based Learning. Machine Learning, pp.37-66.

[3] D. Kibler, D. W. Aha, and M. Albert, (1989). Instance-based prediction of real-valued attributes. Computational Intelligence, pp.51-57.

[4] C. G. Atkeson, A. W. Moore, and S. Schaal, (1997). Locally weighted learning. Artificial Intelligence Review, pp.11-73.

[5] S. Schaal, C. G. Atkeson, and S. Vijayakumar, (2000). Real-Time Robot Learning With Locally Weighted Statistical Learning. Proc. of the IEEE International Conference on Robotics and Automation, pp.288–293.

[6] W. D. Smart, L. P. Kaelbling, (2000). Practical reinforcement learning in continuous spaces. Proc. of the 17th International Conference on Machine Learning, pp.903-910.

[7] D. Y. Cho, B. T. Zhang, (2000). Time Series Prediction using Sequencial Bayesian Evolutionary. KISS autumn conference, Vol.27, No.2, pp.311-313.

[8] S. C. Chapra, and R. P. Canale, (1999). *Numerical Methodd for Engineers, Third Edition*. McGraw-Hill Korea.

[9] http://www.typhoon.or.kr

[10] S. Ruping, (2000). mySVM. Computer Science Dep. AI Unit Univ. of Dortmund.

ACKNOWLEDGEMENT

This research was supported by ETRI(Telematics & USN Research Division) and the Regional Research Centers Program of Ministry of Education & Human Resources Development in Korea.