データストリームから局所的相関を発見する多変量回帰の測定方法の提案

范　　薇† 小柳　佑介† 朝倉　宏一†† 渡邉　豊英†

† 名古屋大学大学院情報科学研究科社会システム情報学専攻　〒 464–8603 名古屋市千種区不老町
†† 大同工業大学情報学部情報システム学科　〒 457–8530 名古屋市南区滝春町
E-mail: †{fan,koyanagi,watanabe}@watanabe.ss.is.nagoya-u.ac.jp, ††asakura@daido-it.ac.jp

あらまし　近年、大規模データであるデータストリームが注目を集めている。時間的に変化するデータレコードの集積であるため、ストリーム間の相関関係も変動である。我々は局所的相関パターンを検出する方法を提案した。この論文では、多変量間の線形回帰関係を表す相関パターンの変化を測定する測定量を提案した。この測定量はインクリメンタル計算することより、効率的にストリーム処理ができる。画像の系列データに対し、分類と識別の実験結果により、提案方法の妥当性を示す。

キーワード　変動的データストリーム，局所的相関関係，多変量線形回帰

Generalized Regression Measure for Local Correlation Tracking in Evolving Data Streams

Wei FAN†, Yusuke KOYANAGI†, Koichi ASAKURA††, and Toyohide WATANABE†

† Department of Systems and Social Informatics, Graduate School of Information Science, Nagoya University
Furo-cho, Chikusa-ku, Nagoya, 464–8603, Japan
†† School of Informatics, Daido Institute of Technology　Takiharu-cho, Minami-ku, Nagoya, 457–8530, Japan
E-mail: †{fan,koyanagi,watanabe}@watanabe.ss.is.nagoya-u.ac.jp, ††asakura@daido-it.ac.jp

Abstract　Many applications deal with multiple data streams, such as supermarket streams of transactional data, stock rating information, observation data from sensor network and so on. In the data stream environment, correlation patterns among data streams generated at different time points may be different due to data evolution. In this paper, we propose a generalized measurement to detect the change of multiple continuous data streams. The generalized measurement produces goodness-of-fit scores online to evaluate whether the new arrival data samples satisfy the existing correlations or not, and tracks local correlations among data streams which enable the user to glean valuable insights into emerging trends in the underlying activity of data. We demonstrate usefulness, robustness and efficiency of our proposed measurement on a series of image data. As illustrated in empirical results, our proposed measurement can detect the change of correlations efficiently to classify the images and recognize the images accurately.

Key words　local correlation tracking, evolving data streams, generalized regression measurement.

1. Introduction

Recently, a large number of applications deal with multiple data streams, such as supermarket transactions data, stock rating information, observation data from sensor network and so on. Often, the volume of such data streams may easily range in the millions on a daily basis, and the data may show important changes in the trends over time because of fundamental changes in the underlying phenomena. This is referred to as *data evolution*. By understanding the nature of such changes, a user may be able to glean valuable insights into emerging trends in the underlying activity.

There is a considerable amount of work which focus on incremental maintenance of data mining models in the context of evolving data streams [1–3]. The focus in our paper is correlation among data streams. The notion of correlation among data streams is important since it allows us to discover groups of data streams with similar behavior and, consequently, to discover potential anomalies which may be revealed by a change in correlation. This paper stud-

ies the problem of data evolution in terms of capturing and tracking local correlations among data streams.

In the data stream environment, correlation patterns among data streams generated at different time points may be different due to data evolution. In this paper, we propose a generalized regression measurement for evaluate the goodness-of-fit of existing correlations among multiple evolving data streams incrementally. Integrated with our previous work of incremental Principal Component Analysis [6], the generalized regression measurement produces a goodness-of-fit score to evaluate whether the new arrival data samples satisfy the existing correlations or not. Therefore, our method can reflect the changes of correlations robustly and accurately. In this paper, we mine the correlations among data streams in an online fashion; therefore, the technique to be discussed in this paper achieves to process an incoming data point efficiently in terms of space usage and execution time, detect changes of correlations dynamically and report the changes automatically.

This paper is organized as follows. In Section 2, we provide a basic background of the traditional regression models. In Section 3, we propose the generalized regression measurement for multiple regression model and discuss how it can be used in order to detect changes of correlations among evolving data streams. In Section 4, empirical results illustrate the efficient and robust performance of our approach. Conclusion and future work are presented in Section 5.

2. Traditional Regression Models

In order to evaluate the linear relationship between sequences Y and X, the traditional linear regression model is established as follows:

$$Y = \beta_0 + \beta_1 X + u. \tag{1}$$

The variable u is called the *error term*. Given a set of data points, $X = [x_1, x_2, \ldots, x_N]$ and $Y = [y_1, y_2, \ldots, y_N]$, β_0 and β_1 can be estimated in the sense of minimization of

$$\sum_{i=1}^{N} {u_i}^2 = \sum_{i=1}^{N} (y_i - \beta_0 - \beta_1 x_i)^2. \tag{2}$$

Using the first order conditions, we can solve β_0 and β_1 as follows:

$$\begin{aligned} \beta_0 &= \bar{y} - \beta_1 \bar{x}, \\ \beta_1 &= \frac{\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \end{aligned}$$

where $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$ and $\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$, are the average of sequences Y and X, respectively. After obtaining β_0 and β_1, *R-squared* measurement is defined as follows for evaluating goodness-of-fit of the linear regression line

$$R^2 = 1 - \frac{\sum_{i=1}^{N} u_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{3}$$

From expression (3) we can further derive:

$$R^2 = \frac{[\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})]^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2 \sum_{i=1}^{N}(y_i - \bar{y})^2}. \tag{4}$$

The value of R^2 is always between 0 and 1. The closer the value is to 1, the better the regression line fits to the data points. We observe that R^2 gives absolute score for evaluating how well the sequences Y and X linearly correlated with each other, unlike distance-based similarity measurement. In addition, according to expression (4), R^2 is invariant to the regression order of two sequences.

The above regression model is called *Simple Regression Model*, since it involves only one independent variable X and one dependent variable Y. We can add more independent variables to construct *Multiple Regression Model* as follows:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_K X_K + u. \tag{5}$$

$\beta_0, \beta_1, \ldots, \beta_K$ can be estimated similarly using the first order conditions.

In order to address our problem of extracting correlations among multiple sequences, the Simple Regress Model has to be run for evaluating linear regression between each pair of sequences. The performance cannot be efficient in the context of multiple sequences. We also cannot use R^2 in the Multiple Regression Model to test whether multiple sequences are linearly correlated with each other or not, because R^2 in the Multiple Regression Model is sensitive to the order of sequences. When we randomly choose X_i to substitute Y as dependent variable and let Y be independent variable, then the regression becomes

$$X_i = \beta_0 + \beta_1 X_1 + \ldots + \beta_i Y + \ldots + \beta_K X_k + u. \tag{6}$$

The R^2 here will be different from that in expression (5). Therefore, in this paper, we propose a generalized regression measure for detecting correlations among multiple sequences.

3. Approach

In this section, we discuss our approach for tracking local correlation among data streams based on our proposed generalized multiple regression measurement. In section 3.1, we illustrate the generalized measurement and derive some important properties of the measurement for our purpose of online mining correlations among multiple dimensional streams. In section 3.2, we give the algorithm for tracking local correlations due to data evolution integrated with our previous work.

3.1 Generalized Regression Measurement

Given $K (K \geqq 2)$ sequences $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_K$ as

$$\begin{pmatrix} \mathrm{X}_1 \\ \mathrm{X}_2 \\ \vdots \\ \mathrm{X}_K \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1N} \\ x_{21} & x_{22} & \ldots & x_{2N} \\ \vdots & \vdots & \vdots & \vdots \\ x_{K1} & x_{K2} & \ldots & x_{KN} \end{pmatrix}. \tag{7}$$

We first organize them into N data points in the K-dimensional

space:

$$p_1 = \begin{pmatrix} x_{11} \\ x_{21} \\ \vdots \\ x_{K1} \end{pmatrix}, p_2 = \begin{pmatrix} x_{12} \\ x_{22} \\ \vdots \\ x_{K2} \end{pmatrix}, \ldots, p_N = \begin{pmatrix} x_{1N} \\ x_{2N} \\ \vdots \\ x_{KN} \end{pmatrix}.$$

Here, in order to construct a multiple regressioin model for the K sequences, that is, we seek to find a regression line in the K-dimensional space that fits to the N data points. Here, we define the error term u_i as the *vertical* distance from data points $(x_{1i}, x_{2i}, \ldots, x_{Ki})$ to the regression line.

Appendix gives in detail how to determine the regression line in K-dimensional space as follows:

$$\frac{p(1) - \bar{\mathrm{X}}_1}{e_1} = \frac{p(2) - \bar{\mathrm{X}}_2}{e_2} = \ldots = \frac{p(K) - \bar{\mathrm{X}}_K}{e_K}. \tag{8}$$

where $p(i)$ is the value of i-th dimensional element of data point p, $[e_1, e_2, \ldots, e_K]^t$ is the eigenvector corresponding to the maximum eigenvalue of the scatter matrix. $\bar{\mathrm{X}}_j (j = 1, 2, \ldots, K)$ is the average of sequence X_j. If any data point p in K-dimensional space satisfies expression (8), it must lie on the line.

Similar to the traditional regression model, after determining the regression line, we define a generalized measurement for evaluate the goodness-of-fit of the regression line as follows:

$$GR^2 = 1 - \frac{\sum_{i=1}^{N} u_i^2}{\sum_{j=1}^{K} \sum_{i=1}^{N} (x_{ji} - \bar{\mathrm{X}}_j)^2}. \tag{9}$$

We can derive following important properties of GR^2:

（1） $GR^2 = \frac{\lambda}{\sum_{i=1}^{N} \|p_i - m\|^2}$ and $0 \leqq GR^2 \leqq 1$.

（2） $GR^2 = 1$ means the K sequences have exact linear correlation with each other.

（3） GR^2 is invariant to the order of $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_K$, i.e., we can arbitrarily change the order of the K sequences, while the value of GR^2 does not change.

Proof.

（1） According to Appendix, we have:

$$\sum_{i=1}^{N} u_i^2 = -\mathrm{e}^t \mathrm{S} \mathrm{e} + \sum_{i=1}^{N} \| p_i - m \|^2 . \tag{10}$$

Thus,

$$\begin{aligned} GR^2 &= 1 - \frac{\sum_{i=1}^{N} u_i^2}{\sum_{j=1}^{K} \sum_{i=1}^{N} (x_{ji} - \bar{\mathrm{X}}_j)^2} \\ &= 1 - \frac{\sum_{i=1}^{N} u_i^2}{\sum_{i=1}^{N} \| p_i - m \|^2} \\ &= \frac{\mathrm{e}^t \mathrm{S} \mathrm{e}}{\sum_{i=1}^{N} \| p_i - m \|^2} \\ &= \frac{\mathrm{e}^t \lambda \mathrm{e}}{\sum_{i=1}^{N} \| p_i - m \|^2} \\ &= \frac{\lambda}{\sum_{i=1}^{N} \| p_i - m \|^2}. \end{aligned} \tag{11}$$

図 1 tracking process of local correlation based on generalized regression measure

Tracking Local Correlations

For each new arrival data point:

1. Calculate $GR_i^2 (i = 1, \ldots, u)$ by substituting eigenvectors corresponding to the largest eigenvalues of existing u kinds of correlations.
2. If the maximum $GR_j^2 \geq threshold\ (j \leq u)$
   the new data point satisfies the j correlation, update the coefficients of j correlation according to algorithm [6];
3. Else
   the new data point represent a new kind of correlation (u+1) among data streams, initialize coefficients of the u+1 kind of correlation.

We know expression (10) $\geqq 0$, therefore:

$$\sum_{i=1}^{N} \| p_i - m \|^2 \geqq \mathrm{e}^t \mathrm{S} \mathrm{e} = \lambda \geqq 0, \tag{12}$$

so we conclude $0 \leqq GR^2 \leqq 1$.

（2） If $GR^2 = 1$, then $\sum_{i=1}^{N} u_i^2 = 0$, which means the regression line fits to the N data points perfectly. Therefore, the K sequences have exact linear correlation with each other.

（3） According to expression (9), this is obvious.

∎

In the next subsection, we utilize the generalized regression measurement GR^2 for detecting changes of correlations among multiple evolving data streams.

3.2 Tracking Local Correlations

Korn *et al.* introduced the problem of discovering *ratio rules* for representing linear correlations among large collections of numeric variables in a database, and proposed an efficient method based on eigensystem analysis (like Principal Component Analysis [4]) for extract k ratio rules (combinations of variables) with k greatest variances [5]. However, this method cannot be applied to detect correlations among multiple evolving data streams dynamically due to its batch process. In addition, the method of [5] is sensitive to noise. In data streams environment, due to data evolution, correlation patterns among data streams generated at different time points may be different due to data evolution. Therefore, the tracking of local correlations is more important for the understanding of evolving data streams. In order to track evolution of correlations, we aim to detect changes of correlation dynamically and report the local correlations automatically.

In this paper, we propose to utilize the generalized measurement GR^2 to produces a goodness-of-fit score to evaluate whether the new arrival data samples satisfy the existing correlations or not. According to the algorithm which we have proposed in [6] for calculating eigenvectors incrementally, we can calculate the score of GR^2 incrementally from expression (11). Fig. 1 illustrates the tracking process of local correlations. The following empirical results illustrate the efficiency and effectiveness of our approach.

表 1　Classification ratio of our proposed approach

Ratio of noise / Types of noise	20%	30%	40%	50%
1	0.88	0.78	0.90	0.86
2	0.86	0.89	0.90	0.83
3	0.87	0.83	0.87	0.87
4	0.86	0.84	0.82	0.87

4. Empirical Results

Our experiments illustrate that the proposed generalized measurement GR^2 can reflect the changes of correlation among multiple evolving data streams efficiently and accurately.

4.1 Data Set and Experiment Environment

The experiments are run on image sequences data. Here we choose images from "Columbia Object Image Library" [7]. There are 20 objects rotated about their vertical axis, resulting 72 images per object. In our experiments, our approach track the correlations of the image sequences, then the accuracy of our approach can be evaluated in terms of the classification ratio of objects and recognition rate of objects.

Key steps of experiments. Our experiments of visual learning and object recognition are conducted the following steps:

（1） Random choose 50 images of one kind of object as the clean dataset.

（2） We interchange images of the clean dataset with images of other objects. Here the images of other objects are "noise". There are two parameters for generating "noise". One is the number of images of "noise", and the other is the number of kind of "noise". As show in Table 2, we randomly interchange clean image datasets with the ratio of 20%, 30%, 40%, 50%, respectively. On the other hand, the number of types of "noise" is set to be 1, 2, 3, 4, respectively.

（3） From one point of view, our approach is used for detecting the change of correlations, and the efficiency of our approach can be evaluated by the correct classification ratio.

（4） From the other point of view, we test the local correlations extracted by our approach with other 10 images for illustrating the accuracy of our approach. Comparing to the batch process of [5] and our previous work of incremental analysis [6], the recognition rate of the generalized measurement is higher.

（5） The results are the average evaluation of 10 runs for each combination of parameters.

4.2 Object Classification

As illustrated in Table 1, we can see that our approach is able to detect the changes of correlations accurately and efficiently.

4.3 Object Recognition

As illustrated in Table 2, we can see that our approach outperforms the batch process and our previous incremental process.

表 2　Recognition rate comparison

	20%			30%			40%			50%		
	GR^2	Batch	Incremental PCA	GR^2	Batch	Incremental PCA	GR^2	Batch	Incremental PCA	GR^2	Batch	Incremental PCA
1	0.87	0.90	0.85	0.93	0.94	0.92	0.88	0.94	0.89	0.96	0.97	0.96
2	0.84	0.86	0.83	0.86	0.88	0.84	0.87	0.90	0.85	0.98	0.99	0.97
3	0.82	0.84	0.80	0.84	0.86	0.83	0.86	084	0.96	0.86	0.88	0.84
4	0.81	0.87	0.80	0.83	0.86	0.83	0.84	0.86	0.83	0.85	0.86	0.83

5. Conclusion

In this paper, we have proposed a generalized regression measure for tracking local correlations among multiple evolving data streams. As illustrated in empirical results, our approach can mine the correlations in an online fashion, therefore, it can detect changes of correlations dynamically and report the changes automatically. For our future work, we aim to expreiment our approach on more data sets, such as numerical streams of transactional data in order to detect spatial information of local correaltions and this kind of infomation is useful for prediction, data compression and so on.

文　献

[1] Tarek, F.G., Mohamed, T. and Hamed, N.: An efficient technique for incremental updating of association rules. ACM International Journal of Hybrid Intelligent Systems, Vol. 5, No. 1, pp. 45–53. ACM Press, New York (2008)

[2] Wang, H.X., Fan, W., Philip, S.Y. and Han, J.W.: Mining concept-drifting data streams using ensemble classifiers. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 226–235. ACM Press, Washington, D.C. (2003)

[3] Yeh, M.Y., Dai, B.R. and Chen, M.S.: Clustering over Multiple Evolving Streams by Events and Correlations. IEEE Trans. Knowledge and Data Eng., Vol. 19, No. 10, pp. 1349–1362. IEEE Press, New York (2006)

[4] Jolliffe, I.: Principal component analysis. Springer Verlag (1986)

[5] Korn, F., Labrinidis, A., Kotidis, Y. and Faloutsos, C.: Quantifiable Data Mining Using Ratio Rules. VLDB Journal, Vol. 8, Morgan Kaufmann, pp. 254–266 (2000)

[6] Fan, W., Koyanagi, Y. S., Asakura, K.Y., Watanabe T. H.: An Incremental PCA for Stream Analysis Based on NLMS Adaptive Filter. Tokai-Section Joint Conference on Electrical and Related Engineering, O-511, (2008)

[7] It is available at http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php

Appendix

Determine Regression Line in K-dimensional space

Suppose $p_1, p_2, \ldots, p_N$ are the N data points in K-dimensional space, we assume the regression line l expressed as:

$$l = m + \alpha \mathrm{e}, \tag{13}$$

where m is a point in the K-dimensional space, α is an arbitrary scalar, and e is a unit vector in the direction of l.

When we project points $p_1, p_2, \ldots, p_N$ to line l, we have point $m + \alpha_i \mathrm{e}$ corresponding to $p_i (i = 1, \ldots, N)$. The squared error for point p_i is:

$$u_i^2 = \| (m + \alpha_i \mathrm{e}) - p_i \|^2 . \quad (14)$$

Thus, the sum of all the squared-error is to be:

$$\begin{aligned}\sum_{i=1}^{N} u_i^2 &= \sum_{i=1}^{N} \| (m + \alpha_i \mathrm{e}) - p_i \|^2 \\ &= \sum_{i=1}^{N} \| \alpha_i \mathrm{e} - (p_i - m) \|^2 \\ &= \sum_{i=1}^{N} \alpha_i^2 \| \mathrm{e} \|^2 - 2\sum_{i=1}^{N} \alpha_i \mathrm{e}^t (p_i - m) \\ &\quad + \sum_{i=1}^{N} \| p_i - m \|^2 \\ &= \sum_{i=1}^{N} \alpha_i^2 - 2\sum_{i=1}^{N} \alpha_i \mathrm{e}^t (p_i - m) \\ &\quad + \sum_{i=1}^{N} \| p_i - m \|^2 .\end{aligned}$$

The sum of squared error must be minimized. Note that $\sum_{i=1}^{N} u_i^2$ is a function of m, α_i and e. Partially differentiating it with respect to α_i and setting the derivative to be zero, we can obtain:

$$\alpha_i = \mathrm{e}^t (p_i - m) \quad (15)$$

Now, we should determine vector e to minimize $\sum_{i=1}^{N} u_i^2$. Substituting (15) to it, we have:

$$\begin{aligned}\sum_{i=1}^{N} u_i^2 &= \sum_{i=1}^{N} \alpha_i^2 - 2\sum_{i=1}^{N} \alpha_i \alpha_i + \sum_{i=1}^{N} \| p_i - m \|^2 \\ &= -\sum_{i=1}^{N} \alpha_i^2 + \sum_{i=1}^{N} \| p_i - m \|^2 \\ &= -\sum_{i=1}^{N} [\mathrm{e}^t (p_i - m)]^2 + \sum_{i=1}^{N} \| p_i - m \|^2 \\ &= -\sum_{i=1}^{N} [\mathrm{e}^t (p_i - m)(p_i - m)^t \mathrm{e}] + \sum_{i=1}^{N} \| p_i - m \|^2 \\ &= -\mathrm{e}^t \mathrm{S} \mathrm{e} + \sum_{i=1}^{N} \| p_i - m \|^2,\end{aligned}$$

where $\mathrm{S} = \sum_{i=1}^{N} (p_i - m)(p_i - m)^t$, called *scatter matrix*.

Obviously, the vector e that minimizes above equation also maximizes $\mathrm{e}^t \mathrm{S} \mathrm{e}$. We can see Lagrange multipliers to maximize $\mathrm{e}^t \mathrm{S} \mathrm{e}$ subject to the constraint $\| \mathrm{e} \|^2 = 1$. Let:

$$\mu = \mathrm{e}^t \mathrm{S} \mathrm{e} - \lambda (\mathrm{e}^t \mathrm{e} - 1). \quad (16)$$

Differentiating μ with respect to e, we have:

$$\frac{\partial \mu}{\partial \mathrm{e}} = 2\mathrm{S}\mathrm{e} - 2\lambda \mathrm{e}. \quad (17)$$

Therefore, in order to maximize $\mathrm{e}^t \mathrm{S} \mathrm{e}$, e must be the eigenvector of the scatter matrix S:

$$\mathrm{S}\mathrm{e} = \lambda \mathrm{e}. \quad (18)$$

Since S generally has more than one eigenvector, we should select the eigenvector e which corresponds to the largest eigenvalue λ.

Finally, we need m to complete the solution. $\sum_{i=1}^{N} \| p_i - m \|^2$ should be minimized since it is always non-negative. To minimize it, m must be the average of $p_1, p_2, \ldots, p_N$.

With m as the average of the N points and e from (18), the regression line l is determined. The line in form of (13) is not easy to understand. Suppose $\mathrm{e} = [e_1, e_2, \ldots, e_K]^t$ is the eigenvector corresponding to the largest eigenvalue and $m = [\bar{\mathrm{X}}_1, \bar{\mathrm{X}}_2, \ldots, \bar{\mathrm{X}}_K]^t$, l can be expressed as:

$$\frac{p(1) - \bar{\mathrm{X}}_1}{e_1} = \frac{p(2) - \bar{\mathrm{X}}_2}{e_2} = \ldots = \frac{p(K) - \bar{\mathrm{X}}_K}{e_K}, \quad (19)$$

where $p(j)$ is the element of j-th dimension of data point p.