Ratio Rules Mining in Concept Drifting Data Streams

Wei Fan *Toyohide Watanabe †Koichi Asakura ‡

Abstract—**Ratio rules mining in data streams is a challenging problem in terms of two issues: concept drifting and continuous large amount of data. In this paper, we propose to estimate distribution of each data stream as time progresses, and to detect partially coagulated intervals in the distribution of each data stream as emerging trends. Then we mine ratio rules from subsequences data at these emerging trends. Traditional techniques cannot be applied to process continuous large amount of data in data streams because of time and space constraints. In this paper, we propose an incremental Principle Component Analysis method, as well as a multiple regression measurement to mine ratio rules incrementally and adaptively. Methods are proposed to detect the change of data trends and to mine ratio rules in a single on-line scan of the data streams. Our extensive experiments on synthetic and real datasets verify efficiency and effectiveness of our proposed methods.**

Keywords: Ratio rules mining, data streams, emerging trends of data, incremental Principal Component Analysis, multiple regression measurement

1 Introduction

Different from association rule mining [1], ratio rule mining is proposed to capture quantitative association knowledge [2, 3, 4, 5]. A classical example is $\{bread : milk : butter\} = 1 : 2 : 1$. This example attempts to characterize purchasing activity:"if a customer spends 1 amount on bread, then s/he is likely to spend 2 amounts on milk and 1 amount on butter". Although techniques for mining ratio rules in static database have been proved to be effective, how to mine ratio rule in concept drifting data streams is still a challenging problem.

Data streams environment imposes additional constraints for the mining procedure: presence of concept drifting and continuous large amount of data. Take the example of market basket data streams. If purchasing activities show an increasing trend in the number of customers with a certain combination of demographic characteristics in a short period, then the manager may be interested in mining ratio rules at this emerging trend in order to describe the purchasing behavior of this group. On the other hand, traditional techniques cannot be applied to process continuous large amount of data in data streams because of time and space constraints.

In order to address the issues, we propose an approach to mine ratio rules at emerging trends of data streams incrementally and adaptively. In order to detect the emerging trends of data streams, we estimate the distribution of the continuous data in each data stream, and detect partially coagulated intervals in the distribution of data as the emerging trends. The emerging trends of all data streams can be detected synchronously in a single on-line scan of data. Then we mine ratio rules from subsequences data at these emerging trends. Here, in addition to adopt an automated incremental Principal Component Analysis method for generating ratio rules, we also propose a generalized multiple regression measurement which attempts to assess how good the generated ratio rules are at each new arrival data point.

This paper is organized as follows. In Section 2, we discuss related work for mining ratio rules. Section 3 gives a formal definition of ratio rules. In Section 4 and Section 5, we elaborate our proposed approaches for tackling two issues of our ratio rules mining problem. In Section 6, we present empirical results of synthetic and real data. Finally, Section 7 concludes the paper.

2 Related Work

In [2], ratio rules are defined as eigenvectors of a database (matrix) whose eigenvalues are the largest. This model has the advantage to estimate missing values in the database. While, it also limits the application of ratio rules mining in data streams: first, the technique of eigen-analysis cannot be applied to process large volume of continuous data because of time and space constraints. Second, because the mining results in [2] are very easy to fall victim to noise, it is difficult to mine "local" ratio rules which are derived from subsets of data. Third, it is hard to quantify a ratio rule, such as how well does a ratio rule fit to the data streams.

[3] and [4] select the ratio rules based on user-specified support and confidence: these methods are able to mine "local" ratio rules. Compared to [3] which is valid only for 2-dimensional data, [4] provide an efficient algorithm to

*Nagoya University, Graduate School of Information Science, Nagoya, 464-8603, Japan. Email: fan@watanabe.ss.is.nagoya-u.ac.jp
†Nagoya University, Graduate School of Information Science, Nagoya, 464-8603, Japan. Email: watanabe@is.nagoya-u.ac.jp
‡Daido University, School of Informatics, Nagoya, 457-8530, Japan. Email: asakura@daido-it.ac.jp

mine ratio patterns from the multidimensional database. However, for data streams, the traditional algorithms for counting support in static database are not efficient.

The authors of [5] proposed an integrated method to mine ratio rules from distributed and changing data source. Similar to our adopted method for mining ratio rules incrementally, a novel robust and adaptive one-pass algorithm (RARR) is proposed. However, in our proposed approach, we declare the ratio rules at emerging trends of data, and give a measurement to evaluate the goodness-of-fit of the ratio rules. In our experimental results, we compare the quality of our algorithm with that of RARR.

3 Problem Definition

Let $A = \{a_1, a_2, \cdots, a_m\}$ be a set of attributes which are also the names of the multiple data streams. An m-dimensional data point T_i arrives at i-th time point, and the value of attribute a_j is represented by $v(a_j)$.

We define a ratio rule P is a set of emerging trend in terms of a reference attribute and a ratio among attributes which is generated from data points at the emerging trend. For example,

$$P = \{(a_1 : a_2 : a_3) = (1 : 2 : 3) \text{ at } a_1 \in [1, 3]\}$$

represents that at the emerging trend of attribute a_1 as $[1, 3]$, ratio relationships among attributes a_1, a_2 and a_3 are $v(a_2)/v(a_1) = 2 : 1$, and $v(a_3)/v(a_1) = 3 : 1$. Here, the attribute a_1 is called reference attribute whose values appear an emerging trend. We omit the attributes whose values are very close to 0, and attributes are ordered in alphabetic order.

In the following sections, we will elaborate the methods to detect emerging trends and to mine ratio rules at the emerging trends incrementally and adaptively.

4 Detection of Emerging Trends

Based on the technique proposed in [6] for diagnosing evolution of data streams, we detect partially coagulated intervals in the distribution of data as emerging trend intervals.

4.1 Change Diagnosis of Data Streams

When a data stream shows high level of evolution, it is expected that the relative data concentrations at various spatial locations may change over time. The work of [6] is able to capture such changes using the concept of velocity density which measures the rate of change of data concentration at a given spatial location over a user-defined time horizon h_t.

Let T be the current instant and S be the set of data points which have arrived in the time window $(T - h_t, T)$.

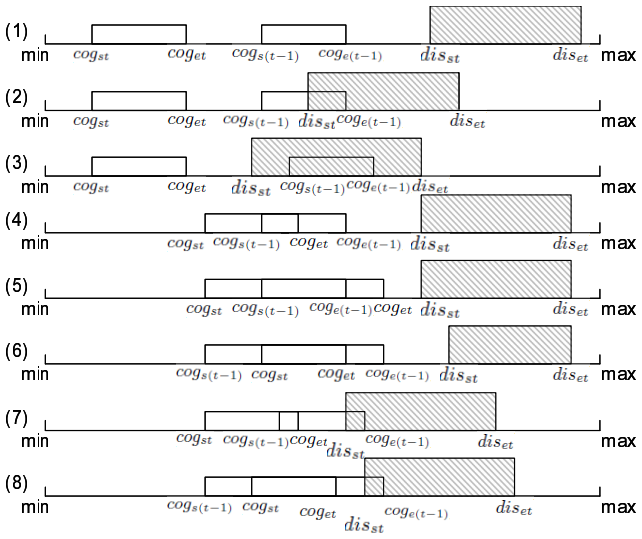


Figure 1: The possible relative positions of coagulation intervals at t and $(t-1)$, and dissolution interval at t.

We intend to estimate the rate of increase in density at spatial location X and time T by using two estimations: *forward time slice density estimation* $F_{(h_s,h_t)}(X, T)$ and *reverse time slice density estimation* $R_{(h_s,h_t)}(X, T)$. Intuitively, $F_{(h_s,h_t)}(X, T)$ measures the density function for all spatial locations at a given time t based on the set of data points which have arrived in the past time window $(T-h_t, T)$. Similarly, $R_{(h_s,h_t)}(X, T)$ measures the density function at a given time t based on the set of data points which will arrive in the future time window $(T, T + h_t)$. Note that both functions can be calculated using the same data points from the interval $(T-h_t, T)$, except that one is calculated assuming time runs forward, whereas the other is calculated assuming time runs in reverse.

Therefore, *velocity density estimation* $V_{(h_s,h_t)}(X, T)$ at a given location X and time T is defined as:

$$V_{(h_s,h_t)}(X, T) = \frac{F_{(h_s,h_t)}(X, T) - R_{(h_s,h_t)}(X, T - h_t)}{h_t}$$

Note that the velocity density is positive, if in the interval $(T - h_t, T)$ a greater number of data points which are closer to X have arrived at the end of the interval. On the other hand, when a greater number of data points which are closer to X are at the beginning of the interval $(T - h_t, T)$, then the velocity density is negative. If the trends have largely remained unchanged, then the velocity density at the location X will be almost zero.

4.2 Emerging Trends Intervals

After discussing change diagnosis of continuous data, we define specific trends in given spatial locations.

Definition 1 *A data coagulation for time point t and*

user-defined threshold min_{coag} is defined to be a connected region R in the data space, so that for each point $X \in R$, $V_{(h_s,h_t)}(X,t) > min_{coag} > 0$.

Definition 2 *A data dissolution for time point t and user-defined threshold min_{dissol} is defined to be a connected region R in the data space, so that for each point $X \in R$, $V_{(h_s,h_t)}(X,t) < -min_{dissol} < 0$.*

From Definition 1 and 2, we determine emerging trend intervals to represent emerging trends of continuous data. For current time point t, we denote the coagulation interval at t and the previous time point $(t-1)$, as $[cog_{st}, cog_{et}]$ and $[cog_{s(t-1)}, cog_{e(t-1)}]$ respectively. Additionally, the dissolution interval at t is denoted as $[dis_{st}, dis_{et}]$. In order to determine the emerging trend interval at t, it is necessary to examine the relative positions of the three above intervals. Figure 1 concludes all of the eight possible relative positions of the three intervals within the range of the reference attribute $[min, max]$:

(1) Because there is no overlap between $[dis_{st}, dis_{et}]$ and $[cog_{s(t-1)}, cog_{e(t-1)}]$, there is no dissolution occurred in the last coagulation interval, then the emerging trend interval at time t equals $[cog_{st}, cog_{et}] \cup [cog_{s(t-1)}, cog_{e(t-1)}]$.

(2) There is an overlap between $[dis_{st}, dis_{et}]$ and $[cog_{s(t-1)}, cog_{e(t-1)}]$: it means that a dissolution occurred in $[dis_{st}, cog_{e(t-1)}]$. Therefore the emerging trend interval at time t equals $[cog_{st}, cog_{et}] \cup [cog_{s(t-1)}, dis_{st}]$.

(3) In this case, the coagulation interval at $(t-1)$ is included in the dissolution interval at t. It seems there is a shift of data concentration from $[cog_{s(t-1)}, cog_{e(t-1)}]$ to $[cog_{st}, cog_{et}]$ at time t. Therefore, the emerging trend interval at time t equals $[cog_{st}, cog_{et}]$.

(4) Although there is an overlap between $[cog_{s(t-1)}, cog_{e(t-1)}]$ and $[cog_{st}, cog_{et}]$, there is no overlap between $[cog_{s(t-1)}, cog_{e(t-1)}]$ and $[dis_{st}, dis_{et}]$. Therefore, there is no dissolution in last coagulation interval, and the emerging trend interval at time t equals $[cog_{st}, cog_{e(t-1)}]$.

(5) Similar to (4), except that $[cog_{s(t-1)}, cog_{e(t-1)}]$ is included in $[cog_{st}, cog_{et}]$. Because there is no dissolution in last coagulation interval, the emerging trend interval at time t equals $[cog_{st}, cog_{et}]$.

(6) Contrasting to (5), $[cog_{st}, cog_{et}]$ is included in $[cog_{s(t-1)}, cog_{e(t-1)}]$. Although it seems that from t, data records are beginning to concentrate on detail interval, because at t there is no dissolution in last coagulation interval, the emerging trend interval at time t equals the last coagulation interval.

(7) There is an overlap between $[cog_{s(t-1)}, cog_{e(t-1)}]$ and $[cog_{st}, cog_{et}]$. Additionally, there is a dissolution in last coagulation interval. Therefore, the emerging trend interval at t equals $[cog_{st}, dis_{st}]$.

(8) Compared with (6), in case a dissolution exists in the last coagulation, the emerging trend interval at t equals $[cog_{s(t-1)}, cog_{et}]$.

Note that the method proposed to diagnose changes of trends and to determine the emerging trend intervals in terms of one reference attribute in a single-scan of data stream: therefore, it is possible to determine all of the emerging trend intervals of all attributes synchronously.

5 Mining Ratio Rules

In this section we aim to mine ratio rules from data points within emerging trend intervals. We adopt our previously proposed incremental Principal Component Analysis method (IPCA) [7] to mine ratio rules. In addition, because a emerging trend interval is determined with respect to one reference attribute, it is possible that there are more than one kind of ratio rules within an emerging trend interval. Therefore, we utilize our previously proposed generalized multiple regression measurement (GR) [8] to evaluate the goodness-of-fit of existing ratio rules.

5.1 Incremental Ratio Rules Mining

We mine ratio rules incrementally using IPCA [7] as shown in Figure 2. The main idea is to read in a new data point $\boldsymbol{x_{t+1}}$ and perform three steps:

- Compute the projection $\boldsymbol{y_i}$, based on the current projection vector $\boldsymbol{e_i}$, by projecting $\boldsymbol{x_{t+1}}$ onto $\boldsymbol{e_i}$. Note that $\boldsymbol{e_i}$ is the i-th principal component coefficient vector, and represents the i-th ratio rule of the past data sets;
- Estimate the reconstruction error $\boldsymbol{u}$ and the energy $\boldsymbol{d}$ based on $\boldsymbol{y}$; and
- Update the estimations of $\boldsymbol{e_i}$.

Intuitively, the goal is to update $\boldsymbol{e_i}$ adaptively and quickly based on the new data point. The larger the reconstruction error $\boldsymbol{u}$, the more $\boldsymbol{e_i}$ is updated. However, the magnitude of this update should also take into account the past data currently "captured" by $\boldsymbol{e_i}$. For this reason, the update is inversely proportional to the current energy $\boldsymbol{d}$.

5.2 Evaluation of Ratio Rules

Here, we utilize GR [8] to evaluate the goodness-of-fit of existing ratio rules at the new data point $\boldsymbol{x_{t+1}}$:

$$GR = \frac{\mathrm{e}^T \mathrm{Se}}{\sum_{i=1}^{t+1} \parallel \boldsymbol{x_i} - mean \parallel^2}$$

Incremental Principal Component Analysis:
0. Initialize $\mathbf{e}_i$ to a unit vector, d_i to a small positive value $(i = 1, \ldots, m)$
1. for a new data point $\mathbf{x}_{t+1}$ arrives
2. $\hat{\mathbf{x}}_1 := \mathbf{x}_{t+1}$
3. for $1 < i < K$
4. $y'_{t+1,i} := \mathbf{e}_i^T \hat{\mathbf{x}}_i$ // compute i - th PC of the new arrival data
5. $d_i := \delta d_i + y'^2_{t+1,i}$ // energy of i -th PC
6. $\mathbf{u}_i := \hat{\mathbf{x}}_i - y'_{t+1,i}\mathbf{e}_i$ // reconstruction error based on i-th PC
7. $\mathbf{e}_i := \mathbf{e}_i + \frac{1}{\mathbf{d}_i} y'_{t+1,i}\mathbf{u}_i$ // update coefficient of i-th PC at time point $t+1$
8. $y_{t+1,i} := \mathbf{e}_i^T \hat{\mathbf{x}}_i$ // output the actual i - th PC
9. $\hat{\mathbf{x}}_{i+1} := \hat{\mathbf{x}}_i - y_{t+1,i}\mathbf{w}_i$ // repeat with remainder PCs of $\mathbf{x}_{t+1}$
10. Endfor
11. Endfor

Figure 2: Incremental Principal Component Analysis.

where $\boldsymbol{e}$=$[e_1, e_2, \ldots, e_K]^T$ is one of the existing ratio rules which calculated by IPCA algorithm in Section 5.1;

$$\mathrm{S} = \sum_{i=1}^{t+1} (\boldsymbol{x_i} - mean)(\boldsymbol{x_i} - mean)^T$$

is denoted as scatter matrix and *mean* is the average vector of the data. As discussed in [8], the result of GR varies within $[0, 1]$. Larger the result is, better the ratio rule fits to the new data. The skeleton of mining ratio rules within an emerging trend interval is illustrated in Figure 3.

6 Experimental Results

Synthetic data streams. In the 2-dimensional synthetic data (x_t, y_t), firstly, we generate continuous x_t from α different Gaussian distribution: $x_t \sim N(\mu_i, \sigma_i^2), (i = 0, 1, \cdots, \alpha)$. We assume that the number of data points with the same distribution is V, and μ_i continuously changes as $\mu_{i+1} = \mu_i + (-1)^s \cdot V$, where $s \in \{1, 2\}$ specifies the direction of the movement and has a probability of 10% to be 1, which makes the streams flow reversely. Secondly, we generate y_t by β ratio relationships, as $y_t = b_i \cdot x_t + \varepsilon, (i = 0, 1, \cdots, \beta)$, where $\varepsilon \sim N(0, 0.1^2)$ denotes random noise.

Additionally, we specify the number of data points satisfying each ratio is B. Therefore, we have the following concept drifting scenarios: (1) $V < B$, then we get the same ratio rules in different emerging trend intervals; (2) $V = B$, then we get different ratio rules at different emerging trend intervals; (3) $V > B$, then we can find different ratio relations in one emerging trend interval.

For each new arrival data point:
1. Calculate $GR_i (i = 1, \ldots, v)$ with respect to the existing v ratio rules.
2. If the maximum $GR_j \geq threshold (j \leq v)$
   the j-th ratio rule fits to the new data point;
3. Else
   the new data point represents a new ratio rule, and initializes coefficients of (v+1)-th ratio rule
4. Update the coefficients of the objective ratio rule according to IPCA algorithm in Figure 2.

Figure 3: Generalized multiple regression measure.

Real data sequences. We downloaded all the NASDAQ stock prices from yahoo website [1] starting from $05-08-2001$ and ending at $05-08-2003$. It has over 4000 stocks. We used the daily closing prices of each stock as sequences. These sequences are not of the same length for various reasons, such cash dividend, stock split, etc. We made each sequence length 365 by truncating long sequences and removing short sequences. Finally, we have 3140 sequences all with the same lenght 365. The 365 daily closing proces start from $05-08-2001$, ending at some date (not nessaritly at $05-08-2003$, since there are no proces in weekends).

Bench methods. We compare performance our approach with that of eigen-analysis based approach [2] and RARR method [5]. We choose $E_{h_s,h_t}(t)/h_t$ to be the value of min_{coag} and min_{dissol}. Here, $E_{h_s,h_t}(t) = h_t \int_{All\ X} |V_{(h_s,h_t)}(X,t)| \delta X$ is the total rate of change over the entire spatial locations at t. Additionally, the threshold of GR measurement is set to be 0.98.

Performance measurements.: we measure the quality of ratio rules by "Guessing Error" which is defined in [2]. Given a set of ratio rules R on a $n \times m$ data matrix D, "single-hole guessing error" is defined as the reconstructing error as in equation (1). Correspondingly, "h-hole guessing error" is defined as in equation (2).

$$GE = \sqrt{\frac{1}{nm} \sum_{i}^{n} \sum_{j}^{m} (\hat{d}_{ij} - d_{ij})^2} \qquad (1)$$

$$GE_h = \sqrt{\frac{1}{nh|H_h|} \sum_{i}^{n} \sum_{H \in H_h} \sum_{j \in H} (\hat{d}_{ij} - d_{ij})^2} \qquad (2)$$

where H_h contains some subset of the ${}_hC_m$ combination of sets H with h "holes".

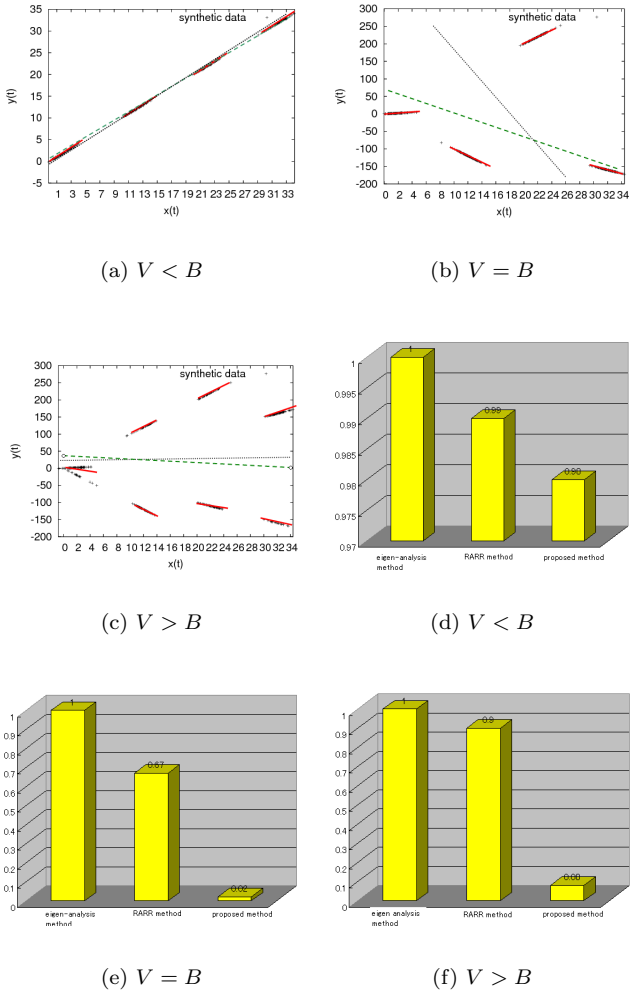
6.1 Sensitivity Analysis

We generate $x(t)$ from $\alpha = 4$ trend intervals. In the case of $V < B$ as shown in Figure 4(a), our proposed method (in red solid segments) achieves the ratio in every trend intervals. Because that all of the data satisfy the same ratio relationship, the result of eigen-analysis method (in black dashed line) and RARR (in green dashed line) are also able to get the exact ratio rules. In Figure 4(b) where $V = B$, there are $\beta = 4$ kinds of different ratio relationship in different intervals of X. Our proposed method is able to capture the different ratio rules in different intervals. While, the results of eigen-analysis method and

[1] http://table.finance.yahoo.com/

RARR method are sensitive to noise, and did not capture the exact ratio relationship in subsets of data. In the last case of $V > B$, it is expected to mine different ratios within a trend interval. As shown in Figure 4(c), we can see that in each trend interval, our proposed method mined the ratio rules according to GR measurement successfully, while, the eigen-analysis method and RARR method are failed.

In Figure. 4(d), 4(e) and 4(f), using the single-hole guessing error measurement, we compare the effectiveness of the three approaches for each concept drifting scenarios. Here, the ratio of results of RARR algorithm and our proposed approach to that of eigen-analysis based method is illustrated. We can find that the performance of our approach is best in all of the cases. Therefore, we can see that the "local" ratio rules generated by our approach in each emerging trend intervals of data describing the ratio relationship better.



(a) $V < B$ (b) $V = B$

(c) $V > B$ (d) $V < B$

(e) $V = B$ (f) $V > B$

Figure 4: Sensitivity analysis.

In Figure. 5, we depict the ratio of processing time of

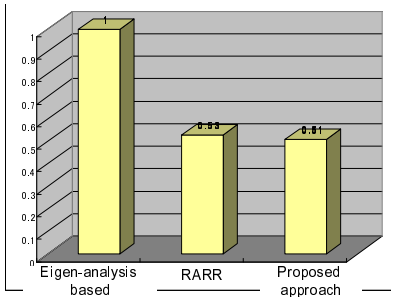


Figure 5: Comparison of processing time.

the three approaches for the case of $(V > B)$. We can see that the RARR method and our proposed approach achieve almost the same processing time. Both of these two approaches realize incremental generation of ratio rules. Although our approach includes the process for detecting emerging trends of data, this process is also incremental and has no effect to process time. The eigen-analysis based method is the most expensive.

6.2 Scalability Analysis

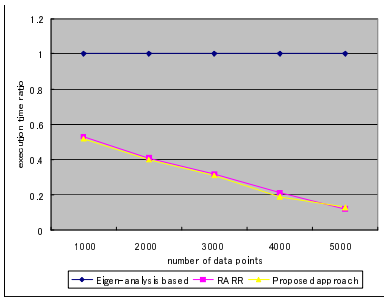
As shown in Figure 6(a), as the number of data points increases from 1000 to 5000, execution time of eigen-analysis based method increases fast. Additionally, in terms of 'h-hole guessing error", as shown in Figure 6(b), eigen-analysis based approach performances the most worst. For the reason that our approach achieves ratio rules mining at emerging trends of data, the guessing error is the least.

6.3 Experiments on Real Data

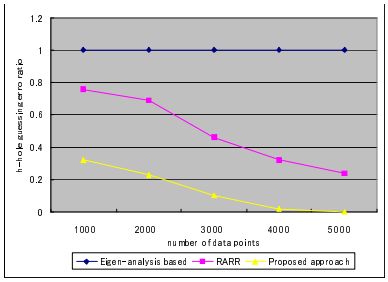
Figure 7 shows real stock data from four companies. From mined ratio rules among these 4 sequences, we find that sequence A and sequence B are independent with other sequences, while sequence C is related with sequence D. We plot the data points of sequence A and sequence B in Figure 8, and we can verify that there is no linear relationship between sequence A and sequence B. On the other hand, the data points of sequence C and sequence D are plotted in Figure 9, and we can mine ratio rules (in red solid segments) between the two sequence in terms of intervals of sequence C, as well as that in the last interval of sequence C, no linear relationship exists, and the real data of these two last subsequences are not similar with each other in Figure 7.

7 Conclusion

In this paper we discussed an approach for mining ratio rules in concept drifting data streams. The velocity density estimation technique is used to detect emerging trend intervals. Then ratio rules are mined within each emerging trend interval in order to provide insight into



(a) number of data points



(b) number of data points

Figure 6: Scalability analysis.

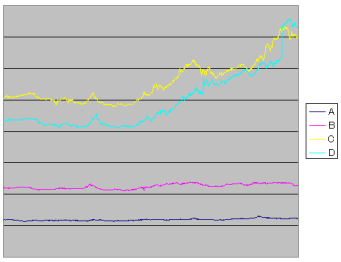


Figure 7: 4 real stock data sequences.

the nature of the pattern in the underlying data characteristics. Breaking the limitation of traditional batch process of eigen-analysis, we propose an incremental process for generating ratio rules. Innovation of the GR measurement addresses the problem for deriving ratio rules from subsets of data.

References

[1] Agrawal, R., Imielinski, T., Swami, A. N.: "Mining Association Rules between Sets of Items in Large Databases". *ACM SIGMOD International Conference on Management of Data*, Washington, D.C.pp. 207–216 93

[2] Korn, F., Labrinidis, A., Kotidis, Y., Faloutsos, C.: "Quantifiable Data Mining Using Ratio Rules". *The International Journal on Very Large Data Bases*. V8, N3-4, pp. 254–266. 00

[3] Hamamoto, M. and Kitagawa. H.: "Ratio Rules Mining with Supprot and Confidence Factors", *IEEE conference proceeding*, pp. 500-505, 06

[4] Zhang, M., and Hsu, W., and Lee. M.: "Mining Prevalence-based Ratio Patterns", *IEEE conference proceeding*, pp. 140-147, 07

[5] Yan, J., Liu, N., Yang, B. Y., Cheng, Q. S., Chen, Z.: "Mining Adaptive Ratio Rules from Distributed Data Sources". *Data Mining and Knowledge Discovery*. V12, N2-3, pp. 249–273. 06

[6] Aggarwal, C.C.: "A Framework for Diagnosing Changes in Evolving Data Streams", *ACM SIGMOD conference proceeding*, pp. 575-786, 03

[7] 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, Aichi, O-511, 08

[8] Fan, W., Koyanagi, Y., Asakura, K., Watanabe. T.: "Generalized Regression Measure for Local Correlation Tracking in Evolving Data Streams". *1st Forum on Data Engineering and Information Management*. Shizuoka, E6-3, 09

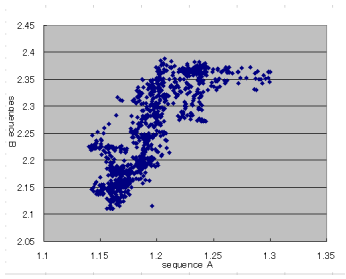


Figure 8: Data sequence A and data sequence B.

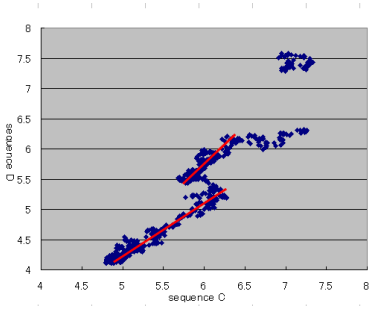


Figure 9: Mined ratio rules between sequcence C and D in terms of trend intervals of C.