Ratio Rule Mining from Multiple Data Sources

Jun Yan yanjun@math.pku.edu.cn
LMAM, Department of Information Science, School of Mathematical Science, Peking University, Beijing, 100871, P. R. China

Ning Liu liun01@mails.tisnghua.edu.cn
Department of Mathematical Science, Tsinghua University, 100084, P. R. China

Qiang Yang qyang@cs.ust.hk
Department of Computer Science, Hong Kong University of Science and Technology, Hong-Kong

Benyu Zhang byzhang@microsoft.com
Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, P. R. China

Qiansheng Cheng qcheng@pku.edu.cn
LMAM, Department of Information Science, School of Mathematical Science, Peking University, Beijing, 100871, P. R. China

Zheng Chen zhengc@microsoft.com
Microsoft Research Asia, 49 Zhichun Road, Beijing, 100080, P. R. China

Abstract. Both multiple source data mining and streaming data mining problems have attracted much attention in the past decade. In contrast to traditional association-rule mining, to capture the quantitative association knowledge, a new paradigm called Ratio Rule (RR) was proposed recently. We extend this framework to mining ratio rules from multiple source data streams which is a novel and challenging problem. The traditional techniques used for ratio rule mining is an eigen-system analysis which can often fall victim to noises. The multiple data sources impose additional constraints for the mining procedure to be robust in the presence of noise, because it is difficult to clean all the data sources in real time in real-world tasks. In addition, the traditional batch methods for ratio rules cannot cope with data streams. In this paper, we propose an integrated method to mining ratio rules from data streams from multiple data sources, by first mining the ratio rules from each data source respectively through a novel robust and adaptive one-pass algorithm (which is called Robust and Adaptive Ratio Rule (RARR)), and then integrating the rules of each data source in a simple probabilistic model with a rule-clustering procedure. In this way, we can acquire the global rules from all the local information sources incrementally. We show that the RARR can converge to a fixed point and is robust as well. Moreover, the integration of rules is efficient and effective. Both theoretical analysis and experiments illustrate that the performance of RARR and the proposed information integration procedure is satisfactory for the purpose of discovering latent associations in multiple-source data streams.

Keywords: Multiple Source Data mining, Data Stream mining, Ratio Rule, Robust Statistics, Eigen System Analysis.

1 Introduction

Large organizations are faced with the problem of mining multiple data sources that are distributed at different locations (Zhang, Wu et al. 2003; Wu, Zhang et al. 2005). An example is the World Wide Web, which consists of many data sources such as Web pages and user logs. In addition, many multiple-data-source mining problems are inherently streaming data problems. In other words, we can not get the data of each data source altogether. As a result, the multiple-source data-stream mining problems have attracted much attention of researchers recently.

Association rule mining, which is one of the most important approaches in data mining (Agrawal, Imielinski et al. 1993; Aumann and Lindell 1999), is one of the major representations for knowledge discovered from large databases (Han and Fu 1995). Association-rule mining from multiple databases has attracted more and more attention of researchers nowadays. To mining the association rules, most prevalent approaches assume that the database transactions only carry Boolean information while ignoring the knowledge inherent in the quantities of the items. To capture the quantitative association knowledge, several effective and efficient algorithms for mining quantitative association rules have been proposed recently (Srikant and Agrawal 1996; Korn, Labrinidis et al. 1998). Among them, a new knowledge representation known as ratio rules is presented and proven to be effective (Korn, Labrinidis et al. 1998). A classical example contrasting ratio rules with traditional association rules is as follows:

Association rule: *{bread, milk} => butter (80%)*

Ratio rule: *bread: milk: butter = 3: 2: 1.*

The former states that customers who buy "bread" and "milk" also tend to buy butter with 80% confidence. The latter means that for each 3 amounts spent on bread, 2 amounts spent on milk, a customer normally spends 1 amount of butter. Ratio rules are different from association rules in many qualitative aspects and have many advantages (Korn, Labrinidis et al. 2000). For example: ratio-rule mining can help reconstruct lost data and repair damaged data and forecast future data. Though ratio rule has been proved to be effective, how to mine ratio rules from multiple-source data streams becomes a novel challenging problem to the modern data-mining research.

One way to deal with this problem is the partitioned approach which aims to mine the rules at different sources and transform the rules to a centralized system to integrate global rules (Zhang, Wu et al. 2003; Wu, Zhang et al. 2005). This can avoid the problem of transporting all the data from different data sources to a centralized location. In this paper, we address this challenging issue as two sub-problems: 1) mining local-ratio rules from each data stream data source adaptively; 2) integrating the local rules into global rules. In this paper, we address the first problem by proposing a novel one-pass incremental ratio-rule mining algorithm and address the second problem by integrating local rules of each data source using a simple probabilistic algorithm. In this way, we can obtain the global rules from all the local information sources adaptively.

To solve the first problem, the traditional technique designed to mining ratio rules from a single data source is an eigen-system analysis that is similar to Principal Component Analysis (PCA) (Korn, Labrinidis et al. 2000; Jolliffe 2002). It aims at finding the geometrical structure of data set by finding out the directions along which the data have maximal variances. Through Singular Value Decomposition (SVD) (Golub and Van 1996), the eigen-system analysis can find an optimal set of directions in the sense of least square reconstruction error. Its computational complexity is $O(m^3)$, where m is the minor value between the sample number and the data dimension. Although the eigen-system analysis is an effective approach, it has several shortcomings for real world problems. First, we observe that the traditional method is very sensitive to outliers. This is a major shortcoming for ratio-rule mining from multiple-source data since we cannot clean all the data sources before the mining computation in real applications, due to the requirement of data streams which arrives continuously. Second, the traditional PCA-like eigen-system analyses are batch algorithms in nature, which requires that the data must be available in advance and all at once. However, this type of batch algorithms can no longer satisfy the applications in which the data are incrementally received. Thus the ratio rule mining algorithm to our multiple source data stream problem should be robust and have the ability of adaptive learning. In other words, a robust and adaptive ratio rule mining algorithm is highly desired.

A naive method to making the ratio-rule mining robust is by introducing sophisticated data cleaning operations. Many works propose algorithms for cleaning a database (Wang, Storey et al. 1995; Chaudhuri, Ganjam et al. 2003; Yan, Zhang et al. 2004). However, the cleaning procedures can slow down the mining procedure greatly and are not suitable to streaming data problems. Some other works have addressed the problems of robust eigen-system analysis (Liano 1996; Huber 2003) or adaptive eigen-system analysis (Kushner and Clark 1978; Weng, Zhang et al. 2003). However, few of them have considered the robust and adaptive eigen-system at the same time (Li 2004). Moreover, to the best of our knowledge, none of the existing algorithms that address both the robust and adaptive issues provide any in-depth theoretical analysis to verify the robustness and convergence formally in the area of data mining.

In this paper, we propose a Robust and Adaptive Ratio Rule (RARR) mining algorithm for multiple source data stream. Our method can find the ratio rules robustly and adaptively from each single data source. Moreover, it is highly scalable in that it generates the ratio rules incrementally. We transform the traditional single source ratio-rule mining problem into an optimization problem of a non-negative energy function under the square criterion (Xu 1993; Jolliffe 2002). In other words, the ratio rules of a database are the minimal points of a non-negative energy function from our point of view. To make the minimization procedure robust, we use a class of criteria called Steady Criterion Function (SCF) (Huber 2003) to replace the square criterion. We can prove that the new produced energy function under SCF can find similar solutions as the square criterion but it has the additional advantage that it is not sensitive to outliers. We propose to use stochastic approximation approach (Ljung 1977; Kushner and Clark 1978) to minimize the new energy function under the SCF *adaptively*. We also give the convergence and robustness proofs of this incremental computation procedure. To solve the second problem of rule integration, we propose to cluster all the local ratio rules from each data source using clustering and a probabilistic model.

The rest of this paper is organized as follows. In Section 2, we give an overview of our proposed system for ratio rule mining on multiple source data streams. In section 3, we introduce some background knowledge of single source ratio rule mining and formulate our problem of robust and adaptive ratio rule mining. In Section 4, we show the detailed proposed technique used to mine ratio rules robustly and adaptively from single source data streams. In Section 5, we give the integration approach used by us to integrate all the rules mined from different data sources. In section 6, we give some experimental results on some real word datasets. Conclusions and future works are given in Section 7. The detailed proof of some theorems which ignored in the regular paper could be found in the appendix.

2 Overview of the Proposed Ratio Rule Mining System

As introduced above, to mine ratio rules from multiple source data streams, the whole system is decomposed into two parts. The first part mines ratio rules from each single data source independently. In this paper, we propose a novel approach called Robust and Adaptive Ratio Rule (RARR) mining algorithm to solve this problem. The details of this algorithm are given in the section 4. Once we get the local ratio rules from each data source, the next problem is to integrate them into global rules. In this paper, we propose to first cluster similar rules from all sources by identifying the rules focusing on the same group of objects, and then integrating all the information implied by the rules using a linear combination. The weight of this linear combination is computed in a probabilistic model. The detailed approach is introduced in Section 5. Figure 1 gives an intuitive explanation of our proposed system, where we mine a group of local rules from each data source and analyze all the local rules by clustering, and then integrate these local rules by a simple linear combination procedure.

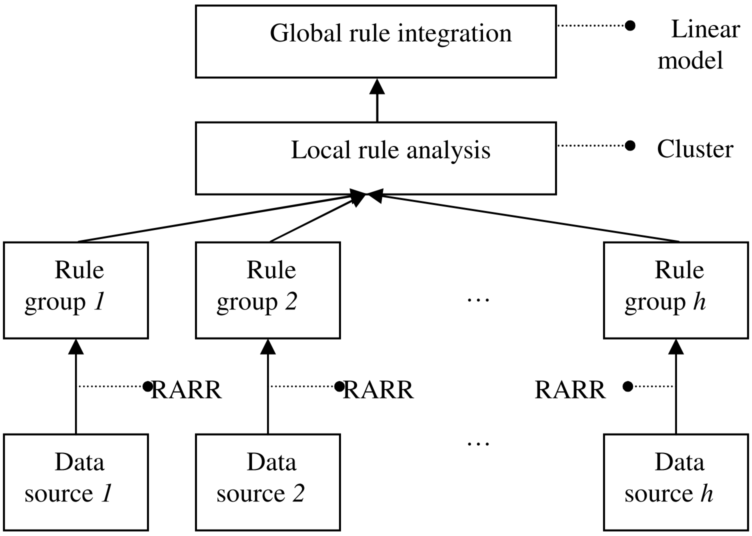


Figure 1. An overview of the proposed system

3 Background Knowledge for Single Source Ratio Rule Mining

In this section, we introduce some background knowledge such as what is ratio-rule mining and what is eigen-system analysis. Moreover, we give some intuitive motivation examples to illustrate our problem. For better comprehension, some mathematical notations and definitions used throughout this paper are shown firstly in Table 1.

Table 1. Notations and definitions

d	Number of attributes
n	Number of items
k	Number of ratio rules retained
$X \in R^{d \times n}$	$d \times n$ data matrix
(x_{ij})	Value at row *i*, column *j* of data matrix *X*
x_i	The i^{th} item (a column vector)
$\overline{x}_i$	The mean of *i* leading items $\overline{x}_i = \frac{1}{i} \sum_{j=1}^{i} x_j$
x_i^T	The transpose of x_i
$< x_i, x_j >$	Inner product of two vectors
$E\{\cdot\}$	Expectation (Kallenberg 2002)
I_k	Identity matrix of order *k*
$\|\cdot\|$	Euclidean norm (Golub and Van 1996)
$span\{\phi_1, \cdots, \phi_k\}$	Linear space spanned by $\phi_1, \cdots, \phi_k$ (Hamilton 1990)
$diag\{\lambda_1, \cdots, \lambda_d\}$	Diagonal matrix whose diagonal elements are $\lambda_1, \cdots, \lambda_d$
h	Number of data sources

3.1 Mining Ratio Rules from Static Single Data Source

In general, the problem of ratio-rule mining is as follows. Given a data matrix $X \in R^{d \times n}$, (x_{ij}) gives the amount spent by customer *j* on product *i*, the goal is to find all ratio rules of the form,

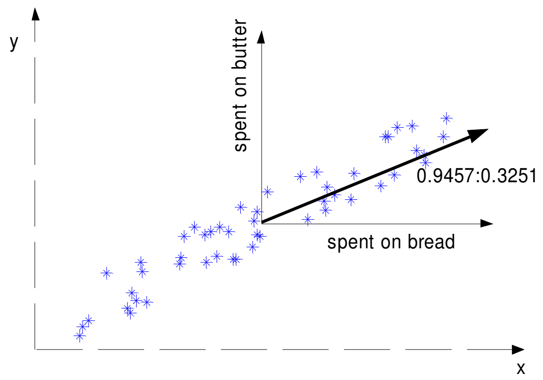
product 1: product 2: …: product d= $v_1 : v_2 : \cdots v_d$

where the rule means that customers who buy the products will spend $v_1 : v_2 : \cdots : v_d$ respectively on each product.

Figure 2 gives an intuitive example. Figure 2(a) lists a set of n $(n=50)$ customers and d $(d=2)$ products organized in an $d \times n$ matrix X . Each row vector of the matrix X^T can be thought of as a d-dimensional point. In Figure 2(b), each point is a customer, the x-axis is the dollars spent on bread and the y-axis is the dollars spent on butter by customers. The straight line is the ratio rule which means the ratio of dollars spent on these two products by most customers. It can be seen that the ratio rule describes the geometrical data distribution, i.e. the trends in which most customers buy the products. The main technique used to mine the ratio rules are eigen-system analysis, for which we give the detail in the next section.

	bread ($)	*butter ($)*
C1	4.75	1.63
C2	3.95	2.07
…	…	…
C50	8.97	3.11

(a). A data matrix in table form



(b). A data matrix and ratio rule in graphical form

Figure 2. An example of ratio-rule mining

3.2 Eigen-system Analysis

Eigen-system analysis algorithms such as Principal Component Analysis (PCA) (Korn, Labrinidis et al. 2000; Jolliffe 2002) and Linear Discriminant Analysis (LDA) (Duda, Hart et al. 2000) have been widely used in statistical data analysis, pattern recognition, digital signal processing and machine learning. To the ratio-rule mining problem, if $l_1, l_2, \cdots, l_k$ are the k leading eigenvectors of covariance matrix c, where $c = {}^{1}\!/_{n} \sum_{i=1}^{n} (x_i - \overline{x}_n)^T (x_i - \overline{x}_n)$, suppose $l_1 = (a_1, a_2, \cdots, a_d)$, then the first ratio rule of the form *product 1: product 2: …: product d*= $v_1 : v_2 : \cdots v_d$ should be l_1, i.e. $v_1 = a_1, v_2 = a_2, \cdots, v_d = a_d$. Due to the same reason, all the ratio rules should be $l_1, l_2, \cdots, l_k$. The ratio rules can be calculated by solving the eigenvectors of covariance matrix *c*. This can be done through Singular Value Decomposition (SVD) (Golub and Van 1996) which has a complexity of $O(m^3)$, where *m* is the minor value between the sample number and the data dimension.

3.3 Motivation and Problem Definition

Though the eigen-system analysis algorithms have been proved to be effective for ratio rule mining problems, there are two big problems we must address for the multiple source data streams. Firstly, we observed that the traditional eigen-system analysis is very sensitive to outliers. In other words, the covariance matrix of a data matrix *X* is very sensitive to outliers (Huber 2003). However, the real data sources usually have noises in most of the real applications. Since it is very hard to clean all the data sources or even impossible to clean the noises on data streams, we need the mining algorithm to be robust. As an example, using the same data matrix with Figure 2, we add three outliers (represented by "+") to this data set and show the ratio rule with outliers in Figure 3. The bold straight line is the ratio rule with these three outliers and the dashed line is the ratio rule without outliers. It can be seen that if a few more customers happen to buy a small quantity of bread and lots of butter, the ratio rules that are computed by traditional approached will make no sense. Then once the mining on some data sources fail, the global rules mined will be wrong. In addition, the streaming data problem of each data source needs the mining approach to be adaptive.

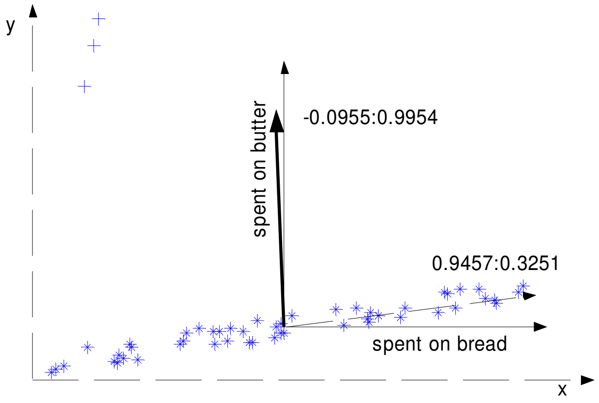


Figure 3. Ratio-rule mining with outliers

Motivated by these, we propose to design a novel robust and adaptive ratio-rule mining method to meet the requirements of this problem. Moreover, the automatic synchronization of multiple source data streams maybe requires the designed algorithm to be a one pass approach. Suppose that x_i, *i=1,2,…* are data items obtained in a data stream. The covariance matrix *c* of the first *n* data points are estimated by $c(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$, where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. To learn ratio rules from this data stream, our problem is to calculate the *k* leading eigenvectors of $c(n)$ adaptively. Moreover, suppose that $\hat{c}(n)$ are covariance matrices that contain the outliers. We need to ensure that the variance between leading eigenvectors of $c(n)$ and the leading eigenvectors of $\hat{c}(n)$ be smaller than a small constant which serves as a threshold value. Moreover, we add a constraint to our algorithm that the algorithm could scan the data stream only once.

4 Robust and Adaptive Ratio Rule Mining

We give the detail of our proposed robust and adaptive ratio-rule mining algorithm in this section for mining from a single data source. The main technique of our algorithm is transforming the ratio rule problem the minimization of an energy function. We use Steady Criterion Function (SCF) (Huber 2003) to replace the square criterion. We can prove that the new produced energy function under SCF can achieve similar solution with square criterion and is robust. Then we optimize the new energy function by a novel application of stochastic approximation adaptively. Through this way, each data source could be mined robustly. Otherwise, the global rules will be wrong no matter what kinds of integration techniques used since the local rules are wrong. Note that we also can look the multiple source data as a single data stream which could be mined just by the algorithm proposed here.

4.1 Robust Ratio Rule Mining

As discussed above, given a data matrix $X \in R^{d\times n}$ with *n* items, the ratio rules are the *k* leading eigenvectors of covariance matrix $c(n) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)(x_i - \bar{x}_n)^T$, where $\bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i$. For better comprehension, we consider only the first ratio rule, i.e. the primary eigenvector of $c(n)$ firstly. Define an energy function $J(m) = E\left\|X - mm^T X\right\|^2 = E\{X^T X\} - (2 - \|m\|^2)m^T c(n)m$, where $m \in R^d$ is a *d* dimensional vector. Theorem 1 below tells us that the first ratio rule is in fact the solution of a minimization problem $\arg\min J(m)$. Before introduce Theorem 1, we give two hypotheses H1 and H2 as below.

H1: $E\{X\} = 0,\ E\{\|X\|^2\} < \infty$;

H2: $c(n) = \Phi diag(\lambda_1,\cdots,\lambda_d)\Phi^T$, where $\Phi = (\varphi_1,\cdots,\varphi_d)$ is an orthogonal matrix, which the columns are orthogonal to each other. $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$.

Note that $E\{X\} = 0$ can be satisfied by centralizing data (Korn, Labrinidis et al. 1998) and $E\{\|X\|^2\} < \infty$ is always true. The only condition that can not be always satisfied is the requirement that the eigenvalues of the covariance matrix are such that $\lambda_1 > \lambda_2 > \cdots > \lambda_d > 0$. The work (Weng, Zhang et al. 2003) gives an example that this constraint can be extended to $\lambda_1 \ge \lambda_2 \ge \cdots > \lambda_d > 0$ which is easier to be satisfied. Note this extended constraint is always true in our experiments.

Theorem 1: Suppose that H1 and H2 can be satisfied. If m^* is the solution of the minimization problem $m^* = \arg\min J(m)$, i.e.

$$m^* = \arg\min E\{\left\|X - mm^T X\right\|^2\} ,$$

then m^* is the primary eigenvector of covariance matrix $c(n)$.

The proof of Theorem 1 is a special case of Theorems 2 and 3 in reference (Xu 1993) by setting p=1. This theorem shows that mining the first ratio rule is in fact equivalent to the optimization of the energy function $J(m)$.

Now define a function $d(X,m) = \left\|X - mm^T X\right\|$, then $J(m) = E\{\rho_s(d(X,m))\}$, where $\rho_s(t) = t^2$ is called the square criterion. It has been proved that (Huber 2003) functions under the square criterion are not robust. As a result, the ratio-rule mining procedure is sensitive to outliers as shown in our example. To solve this problem, we consider using other criteria to replace the square criterion. There is a class of criteria known as Steady Criterion Function (SCF), which is not sensitive to outliers (Huber 2003). These are the functions that we use.

Definition 1: A function $\rho(t)$ is called Steady Criterion Function (SCF) if and only if it satisfies:

(1), $\rho(t) \in C^1[0,\infty)$;

(2), $\rho(t) \ge 0$; $\rho(t) > 0$ when $t > 0$; $\rho(t)$ is critical monotone increasing;

(3), $w(t) = \rho'(t)/2t$ monotone decreasing; $w(t) \to 0$ when $t \to \infty$ and $\lim_{t \to 0^+} w(t)$ exists.

where $w(t)$ is called the Influence Function of $\rho(t)$.

Two of the widely used Steady Criterion Functions are:

(1) Cauchy SCF, $\rho_c(t) = \ln(c + t^2)$ where c is a positive constant;

(2) Exponential SCF, $\rho_e(t) = 1 - e^{-ct^2}$ where c is a positive constant.

We use the SCF $\rho(t)$ to replace the square criterion in the energy function $J(m)$. As an example, we choose the Exponential SCF $\rho_e(t)$ in this paper. The energy function can be rewritten as $J_e(m) = E\{\rho_e(d(X,m))\}$. In other words, we use $\hat{m}^* = \arg\min J_e(m)$ as the estimation of $m^* = \arg\min J(m)$. We can prove that $\hat{m}^*$ should be very close to m^* when the data is clean and $\hat{m}^*$ is not sensitive to outliers while m^* is. Figure 4 shows the solution of $\hat{m}^* = \arg\min J_e(m)$ in the same example of Section 2. The dashed line is the ratio rule without outliers by square criterion, the dotted line is the ratio rule with outliers by square criterion and the solid line is the ratio rule with outliers by exponential criterion. It can be seen that $\hat{m}^*$ on noised data is not sensitive to outliers and is close to m^* computed without outliers.

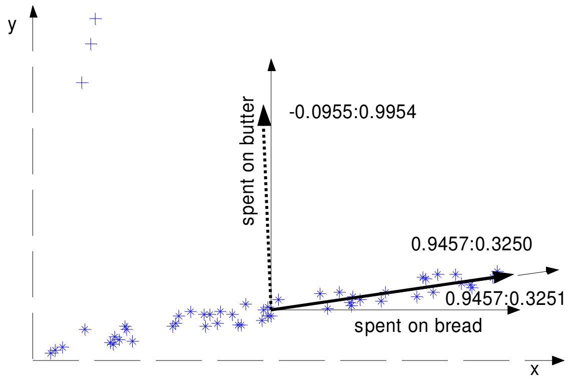


Figure 4. Robust and adaptive ratio-rule mining with outliers

For all k ratio rules including the primary one, suppose that $M=(m_1,m_2,\cdots,m_k)\in R^{d\times k}$, which is similar to the first ratio rule, the conclusion of Theorem 1 still holds; i.e. $M^*=\arg\min E\{\left\|X-MM^TX\right\|^2\}$ should return all the ratio rules. Using the exponential SCF to replace square criterion, we have the robust energy function,

$$J_e(M)=E\{\rho_e(d(X,M))\}\ .$$

We assume that the data obeys the normal distribution, which is often a good approximation to the real world due to the Central Limit Theorem (Kallenberg 2002). Theorem 2 below shows that the transformation of the criterion will not affect the minimal value of the energy function if the data is "clean."

Theorem 2: Suppose that the items of X obey the normal distribution, X satisfies H1 and H2 and $\rho_e(\bullet)$ is an exponential Steady Criterion Function. Then $J_e(M)$ achieves its minima if and only if $M=(\varphi_1,\cdots,\varphi_k)O_p$, where O_p is an orthogonal matrix.

The proof of Theorem 2 can be found in the appendix. Note that the $(\varphi_1,\cdots,\varphi_k)$ are leading eigenvectors of $c(n)={}^1\!/_n\sum_{i=1}^n(x_i-\bar{x}_n)(x_i-\bar{x}_n)^T$. We can draw the conclusion that the minimization of $J_e(M)=E\{\rho_e(d(X,M))\}$ can achieve the similar solution with $M^*=\arg\min E\{\left\|X-MM^TX\right\|^2\}$ without outliers. Theorem 3 below shows that the solution of $\arg\min J_e(M)$ will not be sensitive to outliers.

Before introducing Theorem 3, we must introduce some additional symbols. Suppose $F(R^d)$ is a set composed of all the probability distributions on R^d, if the distribution function of X is $f(\cdot)$. We denote $J_{\rho,f}(m)=E\{\rho(d(X,m))\}=\int_{R^n}\rho(d(x,m))df(x)$. For any $f\in F(R^d)$ and a nonempty set $C\subseteq F(R^d)$, for any $\varepsilon\in[0,1]$, we denote $F_C^{\varepsilon}(f)=\{(1-\varepsilon)f+\varepsilon h\mid h\in C\}$ Let $C_1=\{m\in R^d\mid\|m\|=1\}$ and for any $dis\in[0,\sqrt{2}]$ we define $S_{\rho,f}(dis)=\inf_{\substack{\|m\|=1\\ d(x,\varphi_1)\geq dis}} J_{\rho,f}(m)$.

Theorem 3: Suppose that $C=F(R^d)$, $D_f=\{m\mid J_{\rho,f}(m)=\arg\min J_{\rho,f}(w)\}$, D is the distance between two sets C_1 and D_f. $\rho(\cdot)$ is a SCF and the distribution function f is nonsingular. If $D_f\neq C_1$, then $\forall dis\in(0,D]$, $\exists\sigma>0$ such that for any $\varepsilon<\sigma$, $\tilde{f}\in F_c^{\varepsilon}(f)$, $\forall\tilde{m}\in D_{\tilde{f}}$, the distance between $\tilde{m}$ and D_f should be smaller than *dis*. Moreover, suppose that f is the distributed function of the d dimensional normal distribution *N(c(n),0)*, $\forall dis\in(0,\sqrt{2}]$, if $\tilde{f}\in F_c^{\varepsilon}(f)$ and $\varepsilon<\min\{1/3,(S_{\rho,f}(dis)-S_{\rho,f}(0))/3\}$, for any $m\in D_{\tilde{f}}$, we have the distance between m and φ_1 which is the primary eigenvector of *c(n)* must smaller than *dis*.

The proof is given in the appendix. It is easy to be extended to all k ratio rules. The intuition of this theorem is that if a dataset contains noise, the distance between $\hat{m}^*$ and the target ratio rule on clean data should be smaller than a constant *dis*.

4.2 Robust and Adaptive Ratio Rule Mining

Given the energy function $J_e(M)=E\{\rho_e(d(X,M))\}$, the problem that remains is to optimize this function adaptively. We propose to use a stochastic approximation approach (Ljung 1977; Kushner and Clark 1978; Oja and Karhunen 1985) to obtain this solution. Stochastic approximation was proposed by Robbins and Monro (H. and S. 1951). It is often used to optimize an unknown formed objective function. For better comprehension, we modify the algorithm to solve the primary ratio rule first; i.e. we show how to minimize $J_e(m)=E\{\rho_e(d(X,m))\}$.

Suppose that the X is an array of discrete random variables with identical distribution on $x_i, i=1,2,\cdots$. We have,

$$J_e(m)=E\{\rho_e(d(X,m))\}=\frac{1}{n}\sum_{i=1}^{n}\rho(d(x_i,m)).$$

It can be seen that,

$$-\frac{\partial J_e(m)}{\partial m}=2[(2-\|m\|^2)\Sigma(m)m-m^T\Sigma(m)mm],$$

where $\Sigma(m)=\frac{1}{n}\sum_{i=1}^{n}w_e(d(x_i,m))x_ix_i^T$, $w_e(t)$ is the influence function of $\rho_e(t)$ (see Definition 1). The Gradient Descent algorithm of the energy function $J_e(m)$ then should be,

$$m_{i+1}=m_i+\alpha_i\left[-\frac{\partial J_e(m)}{\partial m}\right]_{m=m_i}$$

i.e.

$$m_{i+1}$$

$$= m_i + \alpha_i \frac{1}{n} \sum_{j=1}^{n} 2 w_e(d(x_j, m_i))\{(2 - \|m_i\|^2) x_j x_j{}^T m_i - m_i^T x_j x_j{}^T m_i m_i\} \quad \text{where} \quad y_j = m_i^T x_j \quad ,$$

$$= m_i + \beta_i \frac{1}{n} \sum_{j=1}^{n} w_e(d(x_j, m_i))\{ y_j (x_j - u_j) + (y_j - y_j')x_j \}$$

$u_j = y_j m_i$, $y_j' = m_i^T u_j$, $\beta_i = 2 / n\alpha_i$, α_i is a positive constant. This is a batch algorithm to solve the first ratio rule. Formally, it can be rewritten as an adaptive algorithm:

$$m_{i+1} = m_i + \alpha_i w_e(d(x_i, m_i))\{ y_i (x_i - u_i) + (y_i - y_i')x_i \} , \qquad (*1)$$

where $\{x_i\}_{i=1}^{\infty}$ is the item stream in a database, $y_i = m_i^T x_i$, $u_i = y_i m_i$, $y_i' = m_i^T u_i$, α_i is a positive constant called the learning rate. Moreover, we require that $\{\alpha_i\}_{i=0}^{\infty}$ satisfy $\alpha_i > 0$, $\sum_{i=0}^{\infty} \alpha_i = \infty$, $\alpha_i \to 0$ when $i \to \infty$. Note we can choose the first item as the initial value of this iteration procedure.

To solve all the *k* ratio rules, observe that

$$-\frac{\partial J_e(M)}{\partial M} \quad ,$$
$$= E\{2 w_e(d(X, M))[2XX^T M - XX^T MM^T M - MM^T XX^T M]\}$$

We can formally get the adaptive iteration algorithm:

$$M_{i+1} = M_i + \alpha_i w_e(d(x_i, m_i))\{ (x_i - u_i) y_i^T + x_i (y_i - y_i')^T \} , \qquad (*2)$$

where $y = M^T x$, $u = My$, $y' = M^T u$. Our algorithm can be conducted by computing equation (*2) iteratively from some given initial value. The algorithm summary is listed in Table 2.

Table 2. Algorithm summary

```
Input: data item stream $x_i$ , i=1,2,… and initial value $M_1$
Initialization $y_1 = M_1{}^T x_1$ , $u_1 = M_1 y_1$ , $y_1' = M_1{}^T u_1$ , $\overline{x}_1 = x_1$ ;
for i= 2, 3, …
    update data mean $\overline{x}_i = {}^1\!/_n \sum_{j=1}^{i} x_j = \frac{(i-1)}{i} \overline{x}_{i-1} + \frac{1}{i} x_i$
    centralize the data $x_i = x_i - \overline{x}_i$ [1].
    $M_{i+1} = M_i + \alpha_i w_e(d(x_i, M_i))\{ (x_i - u_i) y_i^T + x_i (y_i - y_i')^T \}$
    $y_{i+1} = M_{i+1}{}^T x_{i+1}$ , $u_{i+1} = M_{i+1} y_{i+1}$ , $y_{i+1}' = M_{i+1}{}^T u_{i+1}$
Output Robust and Adaptive Ratio Rules $m_j$ , j=1,2,…k at step i, i.e. $M_i$
```

[1] This is used to satisfy the constraint that $E\{X\} = 0$ in H1.

The initial values M_1 used in our experiments are the first k data items. As introduced above, the learning rate $\{\alpha_i\}_{i=0}^{\infty}$ must satisfy the constraints $\alpha_i > 0$, $\sum_{i=0}^{\infty}\alpha_i = \infty$, $\alpha_i \to 0$ when $i \to \infty$. To select the parameters properly, we let $\alpha_i = \varepsilon/(i+1)$. If ε is too large, the algorithm is difficult to converge. If the ε is too small, the convergence speed will be very slow. To give a proper set of parameters, the theorem below proposes an approach to estimate.

Theorem 4: Suppose that $w(t) \le T$, where $T > 0$ is a constant, $\exists C$ such that $\|x_i\| < C$ with probability 1. For any *i*, if $\alpha_i < 1/uTC^2$, then we have $\|m_i\|^2 \le u+1$ with probability 1, u is a constant satisfies $\|m_0\|^2 \le u+1$, $u^3 - u^2 \ge 2$, $u^3 - 4u \ge 4$.
Through Theorem 4, we can estimate the parameters by estimate *u, T* and *C* from the data previously. The proof of it is ignored in this paper.

4.3 Robust and Adaptive Ratio Rule Mining

Lennart and Ljung have studied the convergence of stochastic approximation algorithms in 1977 (Ljung 1977) . Our proposed algorithm can satisfy all the constraints of their work via some simple transformation. A special case of their work is that if an iterative algorithm $m_i = m_{i-1} + \alpha_i Q(i, m_{i-1}, \theta(i))$, *i*=1,2,… can satisfy the constraints A1~A9, it can converge to v (v is the leading eigenvectors of $c = \lim_{i\to\infty} c(i)$ to our problem), where $Q: R \times R^d \times R^d \to R^d$, $\theta(i) = A(m_{i-1})\theta(i-1) + B(m_{i-1})\xi_i$, and $A(\cdot): R^d \to R^{d\times d}$, $B(\cdot): R^d \to R^{d\times d}$ are projections.

A1. $\{\xi_k\}$ is an independent random-variable sequence;

A2. there exist a constant C such that for any k , $\|\xi_k\| < C$ with probability 1;

A3. $Q(i,m,\theta)$ is continuous and differentiable in the domain of m , the partial derivative $\partial Q/\partial m$ and $\partial Q/\partial\theta$ are bounded to some given m and θ ;

A4. $A(\bullet)$ and $B(\bullet)$ are Lipschitz continuous in their domain;

A5. $\forall$ $\bar{m}$ in its domain, $\lim_{i\to\infty} EQ(i,\bar{m},\bar{\theta}(i,\bar{m}))$ exists and is noted as $f(\bar{m})$;

A6. $\alpha_i > 0, \sum_{i=1}^{\infty}\alpha_i = \infty$;

A7. $\exists$ $p > 0$ such that $\sum_{i=1}^{\infty}\alpha_i^{\,p} < \infty$;

A8. α_i monotone decreases with *i*;

A9. $\limsup_{i\to\infty}[\,{}^{1}\!/_{\alpha_i} - {}^{1}\!/_{\alpha_{i-1}}] < \infty$.

If we let $Q(i,m,x) = w_e(d(x,m))\{y(x-u) + (y-y')x\}$, Equation (*1) can be rewritten as,

$$m_i = m_{i-1} + \alpha_i Q(i, m_{i-1}, \theta(i)) ,$$

where $\theta(i) = A(m_{i-1})\theta(i-1) + B(m_{i-1})\xi_i$, $A(\cdot) = 0$, $B(\cdot) = I_d$, $\xi_i = x_{i-1}$. It can be easily proved that (*1) satisfies A1~A9. From Theorem 1 of work (Xu 1995) we can draw the conclusion that (*1) can converge to the primary eigenvector of $c = \lim_{i \to \infty} c(i)$. Due to the same reason, it is easy to be extended to the case of (*2), i.e. (*2) can converge to the leading eigenvectors of $c = \lim_{i \to \infty} c(i)$.

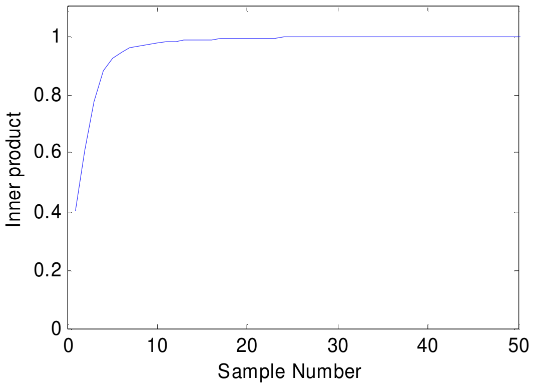


Figure 5. Convergence curve of the previous example

Since $\|m - m'\| = 2(1 - m \cdot m')$, and $m = m'$ iff $< m, m' >= 1$, the correlation between two unit eigenvectors is represented by their inner product, and the larger the inner product is, the more similar the two eigenvectors are. We analyze the example in Section 3 to show the convergence ability of the proposed iterative algorithm by Figure 5. The y-axis is the inner product between the eigenvector of some iteration step and the target eigenvector by batch algorithm without noise. The x-axis is the number of samples received. It can be seen that the convergence is very fast. Then both the theoretical analysis and the experiments tell us the algorithm proposed can converge. To save space, we ignore some convergence results of our experiments in the next section.

5 Multiple Source Ratio Rule Integration

Through the RARR algorithm proposed in Section 4, we can mine local ratio rules from different data sources robustly and adaptively. Without the robust miming, some of the local rules mined can be wrong and thus the global rules will make no sense no matter what kind of information-integration technique used. Due to this reason, the RARR is particularly important to the whole system on multiple source data. However, mining the local rules separately form each data source can still produce redundant rules. The information integration process is therefore very important for filtering out this redundant knowledge. Our next problem is to integrate these groups of local ratio rules mined by RARR to a group of global ratio rules. Moreover, once the local rules updated, the global rules should to be updated as well.

5.1 Local Rule Clustering

Suppose the rules mined from the multiple data sources $S_1, S_2, \cdots S_h$ are $\{\{v_1^1, v_2^1, \cdots v_{k_1}^1\}, \{v_1^2, v_2^2, \cdots v_{k_2}^2\}, \cdots \{v_1^h, v_2^h, \cdots v_{k_h}^h\}\}$ respectively, where v_i^j means the i^{th} rule of data source j. We do not know which combination could achieve the most reasonable global ratio rules from the local ones. One approach to solving this problem is to try all possible combinations with $k_1 \times k_2 \times k_h$ possible cases and try to explain all the possible cases. However, the complexity of this approach maybe too high to some real applications if both the number of data sources and the number of rules mined from different data sources are large. In this paper, we propose to cluster the rules found previously and integrate the local rules within each cluster using a greedy algorithm. As an example, the primary ratio rule of data source 1 (v_1^1) maybe focusing on the relationship between object *a* and object *b*, while the primary eigenvectors of other data sources maybe focusing other objects. Suppose that v_2^3 is also focusing on object *a* and *b*, then we can integrate v_1^1 and v_2^3 as a candidate to obtain the global rule while ignoring the combination of v_1^1 and $v_1^i, i = 1, 2, \cdots, h, i \neq 1$.

Table 3. Summary of local rule clustering

```
Input: local rules $\{\{v_1^1, v_2^1, \cdots v_{k_1}^1\}, \{v_1^2, v_2^2, \cdots v_{k_2}^2\}, \cdots \{v_1^h, v_2^h, \cdots v_{k_h}^h\}\}$
    Initialization: select seeds $\{v_1^l, v_2^l, \cdots v_{k_l}^l\}$ with the largest training sample number.
    for $i = 1, 2, 3, \ldots, h \ (i \neq l)$
      for $j = 1, 2, 3, \ldots, k_i$
        for $k = 1, 2, 3, \ldots, k_l$
            Compute the distance between $v_j^i$ and $v_k^l$ : $dis(v_j^i, v_k^l)$
            Cluster the $v_j^i$ to cluster $k^* = \arg\min dis(v_j^i, v_k^l)$
Output: $k_l$ clusters of local rules.
```

All the mined rules are *d* dimensional vectors. To cluster all local rules, we select the data source with the largest number of training samples as a seed data source first as the seed rules. This is because that the data source with the largest number of data samples can best fit the global data distribution among all other choices. . We then repeat the process by choosing the next rule sets to build clusters. Suppose that the seed rules are $\{v_1^1, v_2^1, \cdots v_{k_1}^1\}$, we then aim at clustering the local rules into k_1 clusters. The clustering procedure is summarized in Table 3.

In this algorithm, the distance between v_j^i and v_k^l , $dis(v_j^i, v_k^l)$ is calculated by the following process: 1) transform the v_j^i and v_k^l into binary vectors, i.e. given a threshold and if a entry of these vectors is larger than this threshold, we set it to one, otherwise, set it to zero; 2) compute the Euclidean distance of these two binary vectors. The intuition of this clustering procedure is that if two local rules from different data sources are focusing on the same group of objects, they should be clustered into one cluster. As a result, they should give a global rule focusing on this group of objects.

5.2 Multiple Source Rule Integration

Suppose all the local rules have been clustered following the approach introduced in the previous subsection. We now present the details of the rule integration algorithm in this section. Each cluster of local rules is focusing on the ratio relationship among a particular group of objects. Assume that $\{v_{i_1}^1, v_{i_2}^2, \cdots v_{i_h}^h\}$ is the i^{th} cluster of local rules, we can use a linear model to integrate the i^{th} global rule by $v_i = \sum_{j=1}^h \alpha_{ij} v_{i_j}^j$ such that $\alpha_{ij} \geq 0$ and $\sum_{j=1}^h \alpha_{ij} = 1$. Intuitively, α_{ij} is the probability of local rule $v_{i_j}^j$ can really contribute to the global rule v_i . We denote this probability by $p(v_{i_j}^j)$.

Using a Bayesian equation, we can see that $p(v_{i_j}^j)P(S_j \mid v_{i_j}^j) = P(v_{i_j}^j \mid S_j)P(S_j)$, where S_j denote the j^{th} data source. Then we can get that:

$$\alpha_{ij} = p(v_{i_j}^j) = \frac{P(v_{i_j}^j \mid S_j)P(S_j)}{P(S_j \mid v_{i_j}^j)} .$$

In the above, the probabilities and conditional probabilities could be estimated as follows:

- $P(S_j)$ is the probability of data source S_j which could be evaluated by the number of data samples come from data source S_j over the number of all the data samples from all the data sources;
- $P(v_{i_j}^j \mid S_j)$ is the conditional probability of $v_{i_j}^j$ from data source S_j which is the importance of local rule $v_{i_j}^j$ among all the local rules from data source S_j . This could be evaluated traditionally by the energy function approach. In other words, since all the local rules from data source S_j are eigenvectors of its covariance matrix and each eigenvector correspond to an eigenvalue, mathematically, the corresponding eigenvalue of a given eigenvector over the trace of the covariance matrix is usually used to represent the importance of this eigenvector among all eigenvectors. However, we consider the streaming data problem in this paper, it is hard to evaluated and update

the eigenvalues each time. Thus we use a simple approach to approximate the conditional probability by Gaussian function here.

$$P(v_{i_j}^j \mid S_j) = \frac{\exp\{-\frac{(i_j)^2}{2}\}}{Const}$$

where i_j means $v_{i_j}^j$ is the i_j^{th} rule, i.e. i_j^{th} eigenvector of data source S_j and the *const* is used to normalize the probabilities such that they can satisfy $\sum P(v_{i_j}^j \mid S_j) = 1$. Note that when computation, we can ignore the *const* since $P(v_{i_j}^j \mid S_j) \propto \exp\{-\frac{(i_j)^2}{2}\}$.

$P(S_j \mid v_{i_j}^j)$ is the conditional probability of a data source S_j once a local rule is given. In other words, once given $v_{i_j}^j$, this is the probability of data source S_j which can generate this rule. Intuitively, this probability describes the generalization ability of $v_{i_j}^j$. Since $P(S_j \mid v_{i_j}^j) = 1 - \sum_{k \neq j} P(S_k \mid v_{i_j}^j)$, the larger the $P(S_j \mid v_{i_j}^j)$ is, the smaller the $\sum_{k \neq j} P(S_k \mid v_{i_j}^j)$ should be. This means that $v_{i_j}^j$ mostly can describe the data distribution of data source S_j and at the same time, it is hard to use it to describe the data distribution of other data sources. Thus to compute the global rule, the larger the value of $P(S_j \mid v_{i_j}^j)$ is, the smaller the weight should be assigned to $v_{i_j}^j$. The intuition behind this observation is that, if a rule very close to another rule in some data source while far away from all other rules by the distance measurement introduced in Section 5.1, its generalization ability should be weak; otherwise if it is close to many rules in different data sources, its generalization ability should be strong. This probability could be evaluated manually or calculated by the distance measurement introduced in section 5.1. For simplicity, we ignore this factor in the experiments of this paper. And then we can simply evaluate the weights by:

$$\alpha_{ij} \propto P(v_{i_j}^j \mid S_j) P(S_j)$$

Note that the integration procedure is easy to be updated by just counting. Though the clustering procedure can be time consuming for the streaming data problem, we do not need to re-cluster the rules each time. Instead, the rules only need to be re-clustered when some data source converge to different solutions by RARR. If the data distributions do not change rapidly, the clusters will not change. Thus the whole system can be efficient to run for multiple-source streaming-data ratio-rule-mining problems.

6 EXPERIMENTS

The running example throughout this paper is from a synthetic dataset. In the first experiment, we extend this running example by splitting the synthetic dataset into 4 independent clusters randomly and consider it as a multiple source data problem. In contrast to the global rule mined by traditional RR approached without outliers, the proposed system can achieve a rule which is very similar to the target global rule. (The Euclidian distance between this two rules is smaller than 0.003.)

In this section, we also give experiments on some real datasets to show the performance of RARR and the integration procedure. All the results of RARR are the solutions after our proposed algorithm converges. Since the convergence speed is fast and all the convergence pictures look like straight line just as Figure 5 has shown, we ignore them to save space. Note we also state the improvements of RARR are significant in contrast to traditional RR by statistical T-test.

6.1 Real Datasets

We ran our experiments on three real datasets to test our approach. Among them, two have been used by the experiments of traditional RR algorithm (Korn, Labrinidis et al. 1998). They are described as follows:

- 'NBA' [2]- this is a basketball statistics dataset from the data collected in 2003-2004 season, including Total minutes played, Field goals made and Field goals attempted etc. This dataset has 435 records and each record has 16 attributes. Note the NBA data used by RR is statistics of earlier seasons (Korn, Labrinidis et al. 1998).
- 'Abalone'[3] - this is a physical measurement of an invertebrate animal, including length, diameter, and weights etc. It has 4177 samples and each sample has 7 attributes.
- 'Adult'[3] - this is a census dataset collected from US adults and the goal of this dataset is to predict whether income exceeds $50,000. This database has 32561 data records and each data item has 15 attributes.

6.2 Experiments Setup

Key Steps of Experiments

Our experiments on these three datasets are both conducted following the steps below:

1) Compute the real ratio rules of the original dataset without outliers as a single source dataset. In this step, we use the same approach used by (Korn, Labrinidis et al. 1998).
2) Given a ratio, randomly generate outliers following this ratio and combine these outliers with the clean data to compose the new dataset with noise. Each outlier is generated by selecting a sample randomly from the clean dataset, and then randomly generates some positions from this selected sample to change. We change these positions by multiply a positive random value between 10 and 1000.
3) Compute the ratio rules of the new data generated with noise as a single data source problem by original RR.

[2] It is available at http://www.dougstats.com/03-04.HomeRD.txt

[3] It is available at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/

4) Split the dataset into 3 blocks as 3 data sources with overlap 0%, 25%, 50% respectively.

5) From one point of view, look the multiple source data as a single source data stream and compute the ratio rules of the data generated with noise adaptively by RARR. The parameters are chosen C=4 and $\varepsilon = 0.05$ in our experiments following Theorem 4. The number of ratio rules *k* is determined by the approach given by (Korn, Labrinidis et al. 1998). We choose the energy threshold 90% in this paper.

6) From the other point of view, use RARR to mining local rules from each data source and integrate them using the algorithms proposed in section 5.

7) Compare and evaluate the results in contrast to the target rule mined in step 1).

Performance Evaluation

To evaluate the performance of our approach, the performance of single-source RARR is the most important since firstly, multiple source data could be considered as a single data stream to mine the global rules directly. Secondly, if a dataset is split into several data sources, some regular data points maybe outliers in some local data source. Thus if the local mining approach is not robust, it is impossible to obtain the high quality the global rules. In this section, we use an inner product and *t*-test to show and evaluate the performances of RARR.

As demonstrated at the end of Section 4.3, $\|m - m'\| = 2(1 - m \cdot m')$, and $m = m'$ iff $< m, m' >= 1$, the correlation between two unit eigenvectors is represented by their inner product, and the larger the inner product is, the more similar the two eigenvectors are. If an algorithm is robust, the inner product between the ratio rule vector mined by itself on data generated with noise and ratio rule vector mined by traditional ratio-rules (RR) on clean data should be close to one.

T-test is used for testing if exist the significance of difference between RR and RARR on data generated with noise. It is a method of examining the accuracy measures commonly used in data mining and information retrieval experiments (Buckley and Voorhees 2000). The T-test gives the probability that the difference between the two approaches is caused by chance. It is customary to say that if this probability is less than 0.05 that the difference is 'significant', i.e., it is not caused by chance, we wish to use T-test to see whether the difference between RR and RARR on data generated with noise is significant. T-test is performed using Microsoft Excel 2002 by us in this paper through the T-test function.

To evaluate the performance of our whole system, we report the inner product between the target global rule and the rule mined by our system (RARR, cluster, integration) on noised data;

4.3 Experimental Results

For the three datasets used in our experiments, the energy ratio (Korn, Labrinidis et al. 1998) of the first ratio rule over all the ratio rules are 94% , 92.65% and 99.51% respectively. Thus the experimental results below focus on the primary ratio rule only.

To ensure that our algorithm be convincing, at the end of this section we give some results for some additional ratio rules with a time complexity analysis.

Firstly, we give tables which list the ratio rules generated by steps 1), 3), 5) and 6) respectively for intuition. Tables 4 and 5 are first ratio rules on 'NBA' and 'Abalone' dataset.

Table 4. First ratio rule on NBA data

Attribute	RR	RR(N)	WHOLE	RARR
Total minutes played	0.81913	-0.0035	0.68423	0.77432
Field goals made	0.14306	0.000807	0.22154	0.15726
Field goals attempted	0.31534	0.010281	0.28155	0.33299
Threes made	0.022247	0.00116	0.00033	0.02358
Threes attempted	0.062488	0.00147	0.10024	0.05509
Free throws made	0.079663	0.004008	0.08089	0.10857
Free throws attempted	0.10246	0.002213	0.10001	0.12011
Offensive rebounds	0.037597	0.000631	0.00121	0.04503
Total rebounds	0.14543	-0.00037	0.22445	0.15038
Assists	0.086909	0.001996	0.00321	0.06562
Steals	0.027519	-6.32E-05	0.0102	0.02061
Turnovers	0.05079	0.001485	0.0203	0.05051
Blocks	0.017732	-2.15E-05	0.01752	0.00705
Personal fouls	0.054868	0.000943	0.00023	0.0569
Total points	0.38804	-0.99992	0.22450	0.44667
Games started	0.022648	2.33E-05	0.10321	0.01663

Table 5. First ratio rule on Abalone data

Attribute	RR	RR(N)	WHOLE	RARR
Sex	-0.09614	-0.00131	0.10233	0.11204
Length	0.021181	-0.00021	0.09425	0.03928
Diameter	0.018047	-0.00046	0.04242	0.03107
Height	0.007352	-0.00053	0.00039	0.00915
Whole weight	0.084289	1	0.04172	0.04304
Shucked weight	0.030085	-0.00042	0.03267	0.01782
Viscera weight	0.017607	-0.00048	0.021541	0.00875
Shell weight	0.027538	-0.00036	0.10231	0.01296
Rings	0.99038	1.99E-06	0.8733	0.99118

The results of 'Adult' are similar, and are omitted for brevity. In the tables, "RR" denotes the solution of traditional RR algorithm on the clean data, i.e. our target global rule, "RR(N)" denotes the solution of traditional RR algorithm on the data generated with noise and "RARR" is the solution of our RARR on the data generated with noise. “WHOLE” means the rule mined by our whole system includes RARR, clustering and integration on noised data. It can be seen that the solution of RARR on data generated with noise is very similar to the solution of RR on clean data. In other words, RARR is robust. On the other hand, RRs are sensitive to outliers; i.e. RRs are not robust. Moreover, the whole system on multiple-source ratio-rule mining can achieve similar rule quality with RARR on a single data stream, which is a satisfactory result.

Figures 6, 7 and 8 are used to show the performance of RARR against levels of noise, through the inner products with the noise-ratio pictures. The x-axis is the ratio of outliers among the datasets. The y-axis is the inner product. The noise ratios used by us are 1%, 2%, 5%, 8%, 10% and 20% respectively. Since the generation of outliers is random, we re-generate outliers following each ratio for ten times to show the performance or RARR in contrast to RR. The solutions showed in these pictures are the average performance among ten runs.

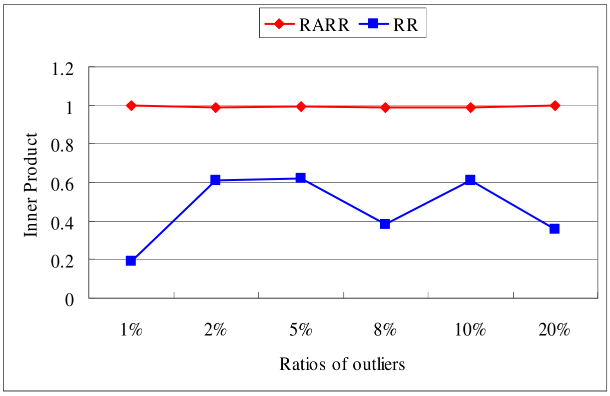


Figure 6. Inner product by different ratio of noise on NBA dataset

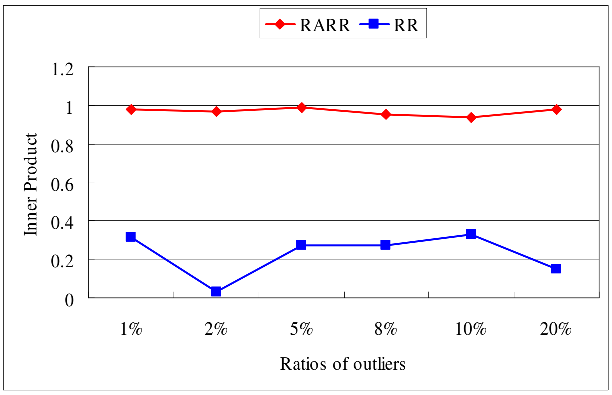


Figure 7. Inner product by different ratio of noise on Abalone dataset

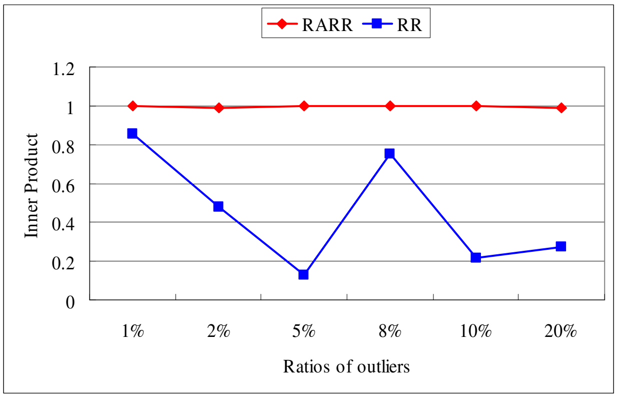


Figure 8. Inner product by different ratio of noise on Adult dataset

The top line with diamonds represents the performance of RARR on data generated with noise, i.e. the inner product between RARR and RR(N). It closes to one consistently. The line with square represents the performance of RR on data generated with noise, i.e. inner product between RR and RR(N). It can be seen again that RARR is robust while RR is not. To show the experimental results are not happen to be so by chance. We compute the probability that the difference between the two approaches is caused by chance. In other words, we compute the T-test values *T-value*=0.000137<<0.05 (NBA), *T-value*=2.4E-5<<0.05 (Abalone) and *T-value*=3. E-08<<0.05 (Adult), as a conclusion, the good performance of RARR is not happened by chance.

We select the "Adult" dataset as an example and analyze the time complexity of RARR as shown in Figure 9, in which the *x-axis* is the ratio of samples used for the ratio-rule mining; the y-axis is the total number of seconds spent.

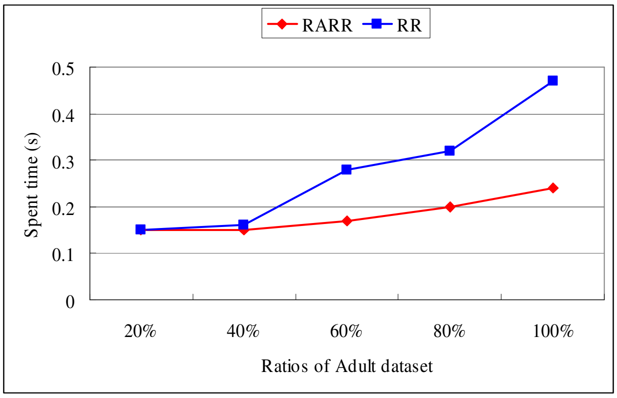


Figure 9. Time complexity analysis on Adult dataset

We use Matlab 6.5 as our tool to record the CPU runtime. It can be seen that RARR on data stream is much more efficient than the traditional batch RR. This is due to the reason that the iterative procedure does not need to re-compute the covariance matrix and do not need to perform SVD at each step. Moreover, the larger the data scale is, the faster RARR should be in contrast to RR since the time complexity of RR is exponentially increasing with the data size while RARR is linearly increasing with the data sizes.

Though the performance of RARR is very important to the whole system, we still need to give a performance overview of the whole system by a table of inner-product. Table 6 shows the performance of the system on the three data sets respectively. We use the two leading rules of each data source and the result of our clustering procedure is that the same order of eigenvectors has always been clustered into the same clusters.

Table 6. Inner products between ratio rules

	<RARR, RR>	<WHOLE, RR>	<RR(N), RR>
NBA, rule1	1	0.9123	0.7477
NBA, rule2	0.9872	0.8762	0.6247
Abalone, rule1	1	0.8999	0.5322
Abalone, rule2	0.9653	0.8528	0.3210
Adult, rule1	1	0.9215	0.8322
Adult, rule2	0.9337	0.8866	0.7631

To make our algorithm more convincing, we list the inner products of the solutions by other ratio rules on NBA data in Table 7. It can be seen that RARR is robust for other rules and the whole system can obtain acceptable rules. Note in this experiment, all the first rules of each data source are clustered into cluster one, all the second rules are clustered into cluster two, and so on.

Table 7. Inner products between ratio rules on NBA data

	<RARR, RR>	<WHOLE, RR>	<RR(N), RR>
Second Rule	0.9872	0.8762	0.6247
Third Rule	0.9536	0.8463	0.6018
Fourth Rule	0.9494	0.8156	0.5823

5 CONCLUSION AND FUTURE WORK

Multiple-source data and streaming data present two of the most important problems for data mining research. A new model called Ratio Rules (RR) has been proposed and has attracted much attention in the past decades. Ratio rules can capture the additional quantitative association knowledge in contrast to the traditional association rules. We address the challenging issue of mining ratio rules on multiple-source data streams in this paper. We solve the problem into two parts: 1) mining local rules from each data stream source; 2) integrating the local rules into global rules. Since the multiple-source ratio-rule mining problem requires the mining algorithm to be robust on each data source which cannot be offered by the traditional approaches, we additionally design an adaptive ratio-rule mining procedure for the streaming data problem. We give an implied energy function of the traditional ratio rule mining procedure and transform this energy function by steady criteria to make it robust. We then solve this optimization problem adaptively by stochastic approximation. We also give the convergence proof of the algorithm and give the mathematical analysis of robustness. Finally, we propose a simple integration technique to combine all the local rules. Note that from the global-rule point of view, we can also look the multiple-source data as a single data stream, which integrates all the data sources previously and mining the global ratio rules just by the proposed RARR algorithm without integration at the end. However, our multiple-source RARR system can adaptively and continuously mine the rules as the data streams arrive. This gives us a more practical solution.

In the future, we plan to conduct our experiments on other forms of data such as the Web data to analyze the behavior of Web users to improve the performance of the search engines. The integration procedure is a challenging problem mathematically when evaluating the eigenvector of a matrix from the eigenvectors of sub-parts of this matrix. To solve this problem, we only give an intuitive approach which can work well in this paper. In the future, we want to design a more effective information integration approach with theoretical proofs for integrating the local rules.

References

Agrawal, R., T. Imielinski, et al. (1993). Mining Association Rules between Sets of Items in Large Databases. ACM SIGMOD Conference, Washington D.C. USA.

Aumann, Y. and Y. Lindell (1999). A statistical theory for quantitative association rules. ACM SIGKDD Conference, San Diego, California, USA, ACM Press New York, NY, USA.

Buckley, C. and E. M. Voorhees (2000). Evaluating evaluation measure stability. ACM SIGIR Conference, Athens, Greece, ACM Press.

Chaudhuri, S., K. Ganjam, et al. (2003). Robust and efficient fuzzy match for online data cleaning. ACM SIGKDD Conference, San Diego, California.

Duda, R. O., P. E. Hart, et al. (2000). Pattern Classification, Wiley Interscience.

Golub, G. H. and L. C. F. Van (1996). Matrix Computations. Maryland, Johns Hopkins University Press.

H., R. and M. S. (1951). "A Stochasitc Approximation Methods." Ann. Math. Stat. 22: 400-407.

Hamilton, A. G. (1990). Linear Algebra, Cambridge University Press.

Han, J. and Y. Fu (1995). Discovery of Multiple-Level Association Rules from Large Databases. VLDB Conference, Zurich, Switzerland.

Huber, P. J. (2003). Robust Statistics, Wiley-IEEE.

Jolliffe, I. (2002). Principal Component Analysis. New York, Springer.

Kallenberg, O. (2002). Foundations of Modern Probability. New York, Springer-Verlag.

Korn, F., A. Labrinidis, et al. (1998). Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining. VLDB Conference, New York City, USA.

Korn, F., A. Labrinidis, et al. (2000). "Quantifiable Data Mining Using Principal Component Analysis." VLDB Journal: Very Large Data Bases 8: 254-266.

Kushner, H. J. and D. S. Clark (1978). Stochastic Approximation Methods for Constrained and Unconstrained Systems. Berlin, Springer-Verlag.

Li, Y. (2004). "On incremental and robust subspace learning." Pattern Recognition 7(37): 1509-1518.

Liano, K. (1996). "Robust Error Measure for Supervised Neural Network Learning with Outliers." IEEE Transactions on Knowledge and Data Engineering 7(1).

Ljung, L. (1977). "Analysis of Recusive Stochasitic Algorithms." IEEE Transactions on Automatic Control AC-22(4).

Oja, E. and J. Karhunen (1985). "On Stochastic Approximation of the Eigenvectors and Eigenvalues of the Expectation of a Random Matrix." J. Math. Anal. Appl. 106: 69-84.

Srikant, R. and R. Agrawal (1996). Mining Quantitative Association rules in Large Relational Tabels. ACM SIGMOD Conference, Montreal, Quebec, Canada.

Wang, R. Y., V. C. Storey, et al. (1995). "A framework for analysis of data quality research." IEEE Transactions on Knowledge and Data Engineering 7(4): 623-640.

Weng, J., Y. Zhang, et al. (2003). "Candid covariance-free incremental principal component analysis." IEEE Transactions on Pattern Analysis and Machine Intelligence 25(8): 1034-1040.

Wu, X., C. Zhang, et al. (2005). "Database Classification for Multi-Database Mining." Information Systems 30(2005): 71-88.

Xu, L. (1993). "Least Mean Square Error Reconstruction Principle for Self-Organizing Neural-Nets." Neural Networks 6: 627-648.

Xu, L. (1995). "Robust Principal Analysis by Self-Organizing Rules Based on Statistical Phsics Approach." IEEE Transactions Neural Networks 6(1): 131-143.

Yan, J., B. Zhang, et al. (2004). IMMC: incremental maximum margin criterion. ACM SIGKDD Conference, Seattle, WA, USA.

Zhang, S., X. Wu, et al. (2003). "Multi-Database Mining." IEEE Computational Intelligence Bulletin 2(1): 5-13.

Appendix

To prove Theorem 2, we must give a lemma firstly. To save space, we ignore the proof of this lemma in this paper.

Lemma 1: Suppose X satisfies H1 and H2, $\rho(\cdot)$ is a SCF. If $J(M) = E\{\rho(d(X,M))\}$ exists for any $m \in R^d$ then $m^* = \arg\min J(m)$ must belong to set $\{m \in R^d \mid \|m\| = 1\}$.

Proof of Theorem 2: suppose $X \sim N(c(n),0)$, where $c(n) = \Phi diag(\lambda_1,\ldots,\lambda_d)\Phi^T, \Phi = (\phi_1,\ldots,\phi_d)$ is orthogonal matrix, $\lambda_1 > \lambda_2 > \ldots > \lambda_n > 0$, $\rho_e(t) = 1 - e^{-ct^2}$.

From lemma 1, we know that $\arg\min J(M)$ belong to set $\{M \in R^{d\times k} \big| MM^T = I_d\}$. Then if $MM^T = I_d$, denote $c(n)$ by Σ we have,

$$J(M) = \frac{1}{(2\pi)^{\frac{d}{2}} |\Sigma|^{\frac{d}{2}}} \int_{R^n} (1 - e^{-c\|x - MM^T x\|^2}) e^{-1/2 x^T \Sigma^{-1} x} dx = 1 - \frac{|\tilde{\Sigma}|^{\frac{1}{2}}}{|\Sigma|^{\frac{1}{2}}}$$ where

$$\tilde{\Sigma} = \{\Sigma^{-1} + 2cl - 2cMM^T\}^{-1}$$
$$= \Phi\{diag(1/\lambda_1 + 2c, 1/\lambda_2 + 2c, \ldots, 1/\lambda_d + 2c) - 2cWW^T\}^{-1}\Phi^T$$

and $W = \Phi^T M$.

Let $\Sigma_* = diag(1/\lambda_1 + 2c, 1/\lambda_2 + 2c, \ldots, 1/\lambda_d + 2c) - 2cWW^T$, then $\min(J(M)) \Leftrightarrow \min|\Sigma_*|$. On the other hand, since,

$$-\frac{\partial J_e(M)}{\partial M}$$
$$= E\{2w_e(d(X,M))[2XX^T M - XX^T MM^T M - MM^T XX^T M]\}$$

we know that the $MM^T = I_d$ should satisfy $\tilde{\Sigma}M - MM^T\tilde{\Sigma}M = 0$.

It is easy to prove that $\tilde{\Sigma}$ is nonsingular and $\tilde{\Sigma}M - MM^T\tilde{\Sigma}M = 0$

$$\Rightarrow \tilde{\Sigma} span\{m_1, m_2, \ldots, m_k\} = span\{m_1, m_2, \ldots, m_k\}$$
$$\Rightarrow \Sigma_*^{-1} span\{w_1, w_2, \ldots, w_k\} = span\{w_1, w_2, \ldots, w_k\}$$
$$\Rightarrow \Sigma_* span\{w_1, w_2, \ldots, w_k\} = span\{w_1, w_2, \ldots, w_k\}$$
$$\Rightarrow diag(1/\lambda_1 + 2c, 1/\lambda_2 + 2c, \ldots, 1/\lambda_d + 2c) span\{w_1, w_2, \ldots, w_k\}$$
$$= span\{w_1, w_2, \ldots, w_k\}$$

Suppose $\mu_i = 1/\lambda_i + 2c$ .Because the rank of W is k, there must be k row vectors of W are linearly independent. Suppose the first k rows are linearly independent. We rewrite W into block:

$W = \begin{bmatrix} W_k \\ W_p \end{bmatrix}$, where W_k is a $k \times k$ nonsingular matrix. *p*=*d*-*k*. It is obvious that: $\overline{W} = WW_k^{-1} = \begin{bmatrix} I_k \\ V \end{bmatrix}$, $V = W_p W_k^{-1}$. let $\overline{W} = (\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_k)$, from derivation above, we have:

$diag(u_1, u_2, \ldots, u_d) span\{\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_k\} = span\{\overline{w}_1, \overline{w}_2, \ldots, \overline{w}_k\}$.

If $V \neq 0$, then $\exists k < i \leq d, 1 \leq j \leq k$, such that $\overline{w}_{ij} \neq 0$. So

$$\bar{w} = diag(u_1 u_{2,} \ldots, u_d)\bar{w}_j - u_j \bar{w}_j \in span(\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_k) .$$

It is obvious that the k leading elements of $\bar{w}$ are zeros and the i^{th} element is not zero, so we have: $\bar{w} \notin span(\bar{w}_1, \bar{w}_2, \ldots, \bar{w}_p)$. It is conflict between two conclusions. So $V = 0$. From $V = 0$, we know $W = 0$. So W_k is an orthogonal matrix. Let $O_k = W_k$, then $M = (\phi_1, \phi_2, \ldots, \phi_k) O_k$. When the k leading rows of W are not linearly independent, there exists a permutation matrix P with rank d such that the k leading rows of PW are linearly independent. Then we can get,

$Pdiag(u_1 u_{2,} \ldots, u_d) P^T Pspan\{w_1, w_2, \ldots, w_k\} = Pspan\{w_1, w_2, \ldots, w_k\}$ Similar to above, we can proof that $\exists\ O_k$, such that:

$$PW = \begin{bmatrix} O_k \\ 0 \end{bmatrix}, \text{ and so } M = \Phi P^T \begin{bmatrix} O_k \\ 0 \end{bmatrix}$$

Assume $\Phi P^T = (\phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_d})$, where $\{i_1, \ldots, i_d\}$ is one of the permutations of $\{1, \ldots, d\}$. Then $M = (\phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_k}) O_k$. It is easy to see that *J(M)* take its minima when $M = (\phi_{i_1}, \phi_{i_2}, \ldots, \phi_{i_k}) O_k$. □

Proof of Theorem 3: suppose $\rho(t) \le \kappa, \kappa > 1$. $\forall h \in F(R^d),$ $\varepsilon \in [0,1)$, let $\tilde{f} = (1-\varepsilon) f + \varepsilon h$. From f is nonsingular, we can know that $\tilde{f}$ is singular. Moreover, because $\rho(\bullet)$ is bounded and critical monotone increasing, we can proof that $J_{\rho,f}(m)$ and $J_{\rho,\tilde{f}}(m)$ exist and continuous for any $m \in R^d$, and they all take the minimum in the set C_1 . So $D_f \subseteq C_1, D_{\tilde{f}} \subseteq C_1$, D_f and $D_{\tilde{f}}$ are no-empty compact set.

Since D_f is compact set and $D_f \ne C_1$, we can get $D > 0$. $\forall r \in [0, D]$, Define: $S_{\rho,f}(r) = \inf\limits_{\substack{m \in C_1 \\ d(m, D_f) \ge r}} J_{\rho,f}(m)$, it is easy to know that $S_{\rho,f}(r)$ monotone increasing on $[0, D]$. Apparently, $\forall r \in [0, D]$, we must have $S_{\rho,f}(r) > S_{\rho,f}(0)$ .Take $\sigma = \min(1/3, (S_{\rho,f}(d) - S_{\rho,f}(0))/(3\kappa))$. $\forall h \in F(R^n)$ and $\varepsilon < \sigma$, let $\tilde{f} = (1-\varepsilon) f + \varepsilon h$ and $\tilde{m} \in D_{\tilde{f}}$.

Suppose $d(\tilde{m}, D_f) \ge d$. Since D_f is compact set, we can find $m' \in D_f$ such that $d(\tilde{m}, D_f) = \|\tilde{m} - m'\|$. Then

$$\begin{aligned} J_{\rho,\tilde{f}}(\tilde{m}) &= (1-\varepsilon)\int_{R^n} \rho(d(x,\tilde{m}))df + \varepsilon \int_{R^n} \rho(d(x,\tilde{m}))dh \\ &\ge (1-\varepsilon)\int_{R^n} \rho(d(x,\tilde{m}))df \\ &\ge (1-\varepsilon)\int_{R^n} \rho(d(x,m'))df + (1-\varepsilon)(S_{\rho,f}(d) - S_{\rho,f}(0)) \\ &\ge J_{\rho,\tilde{f}}(m') + (1-\varepsilon)(S_{\rho,f}(d) - S_{\rho,f}(0)) - 2\kappa\varepsilon \\ &> J_{\rho,\tilde{f}}(m') \end{aligned}$$

This comes into conflict with $\tilde{m} \in D_{\tilde{f}}$. $D_f = \{\pm\phi_1\}, D = \sqrt{2}, \kappa = 1$, so $\sigma = \min\{1/3, (S_{\rho,f}(d) - S_{\rho,f}(0))/3\}$

□