Ratio Rules: A New Paradigm for Fast, Quantifiable Data Mining

Flip Korn, Alexandros Labrinidis, Yannis Kotidis
University of Maryland
Christos Faloutsos
Carnegie Mellon University

Proceedings of VLDB Conference, 1998.

1999.12.2

Mining Quantitative Association Rules in Large Relational Tables

Ramakrishnan Srikant, Rakesh Agrawal
IBM Almaden Research Center

Proceedings of ACM SIGMOD Conference, 1996.

Outline

- **Introduction**
- **Ratio Rule Discovery**
- **Prediction of Missing Values**
- **Measurement of the Goodness**
- **Experiments**
- **Discussion**

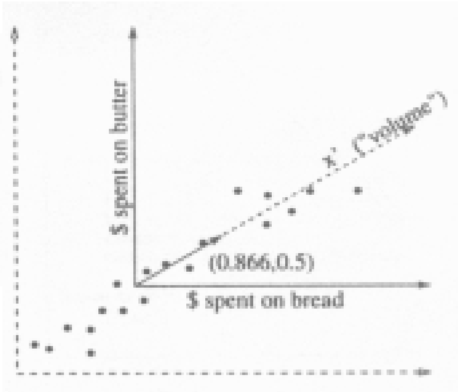
Introduction

- **Paradigms**
  - boolean association rule
    - **{bread,milk} ⇒ butter (90%)**
  - quantitative association rule
    - **bread:[3-5] and milk:[1-2] ⇒ butter:[1.5-2]**
  - ratio rule
    - **bread:milk:butter=1:2:5**
    - **applications**
      - *data cleaning, forecasting, decision support, outlier detection, visualization*

Introduction

- **An example**

customer	*bread* ($)	*butter* ($)
Billie	.89	.49
Charlie	3.34	1.85
Ella	5.00	3.09
...	...	...
John	1.78	.99
Miles	4.02	2.61



P.2

Introduction

- **Innovations**
  - ■ a single-pass algorithm for ratio rule discovery
    - ◆**eigensystem analysis**
  - ■ a method to predict missing/hidden values from the ratio rules
    - ◆**linear algebra**
  - ■ a measure of the goodness for a set of rules
    - ◆**guessing error**

P.3

Introduction

- **Notations**

symbol	definition
N	number of records
M	number of attributes
k	cutoff (number of Ratio Rules retained)
h	number of holes
$\mathcal{H}$	set of cells which have holes
$\mathcal{R}$	set of rules
GE_1	guessing error over each hole
GE_h	guessing error over h holes
$\times$	matrix multiplication
$\mathbf{X}$	the $N \times M$ data matrix
$\mathbf{X}_c$	the centered version of $\mathbf{X}$
$\mathbf{X}^t$	the transpose of $\mathbf{X}$
$x_{i,j}$	value at row i, column j of the matrix $\mathbf{X}$
$\hat{x}_{i,j}$	reconstructed (approximate) value at row i and column j
$\bar{x}$	the mean cell value of $\mathbf{X}$
$\mathbf{C}$	the $M \times M$ covariance matrix ($\mathbf{X}_c^t \times \mathbf{X}_c$)
$\mathbf{V}$	the $M \times k$ RR matrix

P.4

Ratio Rule Discovery

- **Eigensystem analysis**
  - compute the eigenvalues and eigenvectors of the covariance matrix for the given data points
  - identify the axes of greatest variance
  - reduce the dimensionality of a data set while retaining as much as variation as possible
    - ◆**only the eigenvectors of the k largest eigenvalues are used as the ratio rules**

$$\frac{\sum_{i=1}^{k} \lambda_i}{\sum_{j=1}^{M} \lambda_j} \approx 85\%$$

P.5

Ratio Rule Discovery

- **Proposed Method**
  - zero-mean the input matrix to derive X_c
  - compute C $C \equiv X_c^t \times X_c$
  - compute the eigenvalues and eigenvectors of C
- **An example**
  - N=4, M=2, k=1 (column averages:[2 3])

$$X = \begin{bmatrix} 2 & 4 \\ 3 & 3 \\ 3 & 4 \\ 0 & 1 \end{bmatrix} \Rightarrow X_c = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ -2 & -2 \end{bmatrix} \Rightarrow C = \begin{bmatrix} 6 & 5 \\ 5 & 6 \end{bmatrix}$$

P.6

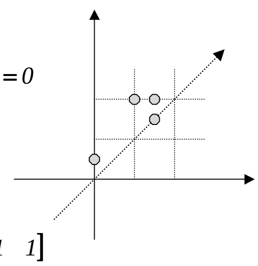
Ratio Rule Discovery

- **Results**
  - the largest eigenvalues: λ_1=11
  - the first ratio rule (RR): [1 1]

$$det\left(\begin{bmatrix} 6-\lambda & 5 \\ 5 & 6-\lambda \end{bmatrix}\right) = 0 \Rightarrow (6-\lambda)(6-\lambda) - 25 = 0$$

$$\Rightarrow \lambda^2 - 12\lambda + 11 = 0 \Rightarrow \lambda = 11 \vee 1$$

$$\begin{bmatrix} 6-11 & 5 \\ 5 & 6-11 \end{bmatrix}\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = 0 \Rightarrow \nu_1 = [y_1 \quad y_2] = [1 \quad 1]$$



P.7

Ratio Rule Discovery

- **Single-pass algorithm**

```
/* input: training set X on disk */
/* output: covariance matrix C */
for j := 1 to M do
    colavgs[j] ← 0;
    for l := 1 to M do
        C[j][l] ← 0;
for i := 1 to N do
    Read ith row of X from disk (X[i][1],...,X[i][M]);
    for j := 1 to M do
        colavgs[j] += X[i][j];
        for l := 1 to M do
            C[j][l] += X[i][j]*X[i][l];
for j := 1 to M do
    colavgs[j] /= N;
for j := 1 to M do
    for l := 1 to M do
        C[j][l] -= N * colavgs[j] * colavgs[l];
```

P.8

Ratio Rule Discovery

- **Eigensystem computation**

```
input:
    covariance matrix C in main memory

output:
    eigenvectors v_1, ..., v_k (i.e., the RRs)

compute eigensystem:
    {v_1, ..., v_M} ← eigenvectors(C);
    {λ_1, ..., λ_M} ← eigenvalues(C);
    sort v_j according to the eigenvalues;
    choose k based on Eq. 1;
    return the k largest eigenvectors;

complexity:
    O(M^3)
```

P.9

Prediction of Missing Values

- **Definitions**
  - h-hole row vector b_H
    - ◆**$b_{\{2,4\}}$=[x_1 ? x_3 ? x_5] (h=2)**
  - (M-h)×M elimination matrix E_H
    - ◆**$E_{\{2,4\}}$=** $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$
- **Basic idea**
  - fill the unknowns by the ratio rules in $E_H \times V$ and the partial knowledge in $E_H \times b_H^t$

P.10

Prediction of Missing Values

- **Pseudo code**

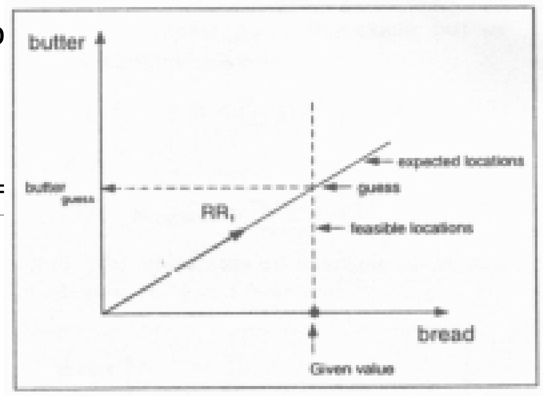
```
/* input:  b_H, a 1 × M row vector with holes */
/* output: b̂, a 1 × M row vector with holes filled */
1.  V' ← E_H × V;                          /* ''RR-hyperplane'' */
2.  b' ← E_H × b_H^t;                      /* ''feasible sol'n space'' */
3.  solve V' × x_concept = b' for x_concept  /* solution in k-space */
4.  d ← V × x_concept;                      /* solution in M-space */
5.  b̂ ← b × [E_H^c]^t + d × [E_H]^t;
```

- **2-D example**
  - $b_{\{2\}}$=[1 ?], $E_{\{2\}}$=[1 0], V=[1 1]^t
  - V'=[1], b'=[1], $x_{concept}$=[1], d=[1 1]

P.11

Prediction of Missing Values

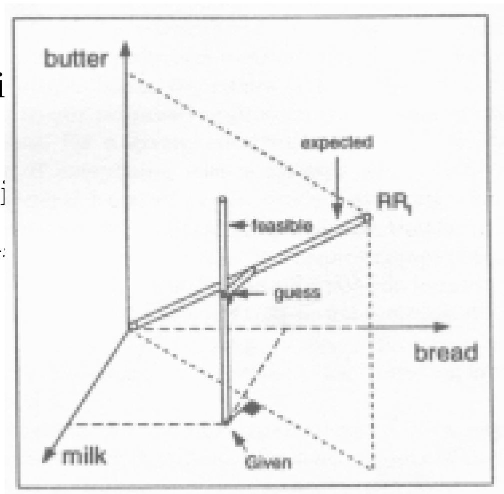
- **Case 1**
  - exactly-sp
  - M-h=k
  - one exact
    - ◆**M=2, h=**



P.12

Prediction of Missing Values

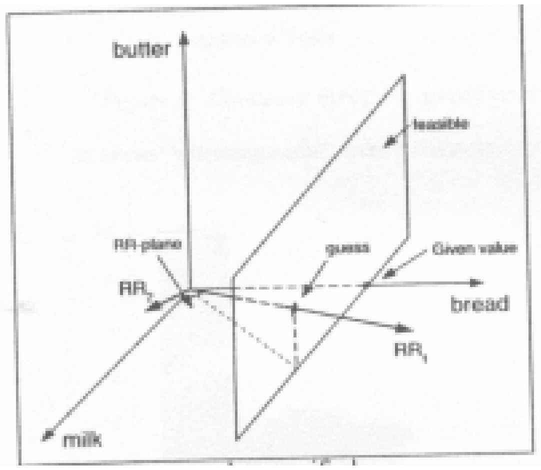
- **Case 2**
  - over-specifi
  - M-h>k
  - no intersecti
    - ◆**M=3, h=1**



P.13

Prediction of Missing Values

- **Case 3**
  - under-s
  - M-h<k
  - infinite
    - ◆**M=3,**



P.14

Measurement of the Goodness

- **Guessing error**

$$GE_1 = \sqrt{\frac{1}{NM}\sum_{i}^{N}\sum_{j}^{M}\left(\hat{x}_{ij} - x_{ij}\right)^2} \qquad GE_h = \sqrt{\frac{1}{Nh|H_h|}\sum_{i}^{N}\sum_{H \in H_h}\sum_{j \in H}\left(\hat{x}_{ij} - x_{ij}\right)^2}$$

- **Examples**

$$X = \begin{bmatrix} 2 & 4 \\ 3 & 3 \\ 3 & 4 \\ 0 & 1 \end{bmatrix} \Rightarrow \hat{X} = \begin{bmatrix} 4 & 2 \\ 3 & 3 \\ 4 & 3 \\ 1 & 0 \end{bmatrix} \Rightarrow \begin{aligned} &GE_1(X) \\ &= \sqrt{2^2 + 2^2 + 0 + 0 + 1^2 + 1^2 + 1^2 + 1^2} \\ &= 2\sqrt{3} \end{aligned}$$

$$X_1 = \begin{bmatrix} 1 & 1 \\ 2 & 0 \end{bmatrix} \Rightarrow \hat{X}_1 = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix} \Rightarrow \begin{aligned} &GE_1(X_1) = \sqrt{0 + 0 + 2^2 + 2^2} \\ &= 2\sqrt{2} \end{aligned}$$

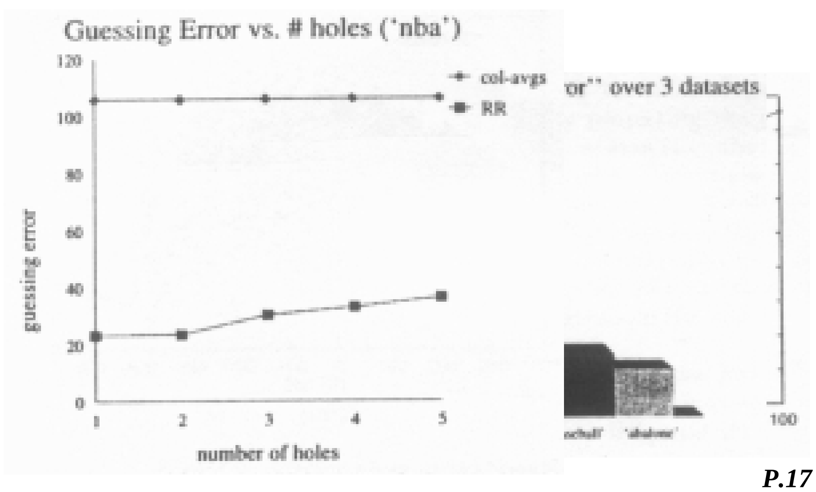
P.15

Experiments

- **Data sets**
  - NBA (452×12)
    - **minutes played, field goals, rebounds, fouls**
  - baseball (1574×17)
    - **batting average, at-bats, hits, home runs**
  - abalone (4177×7)
    - **length, diameter, weights**
- **Competitor**
  - column averages

P.16

Experiments



P.17

Discussion

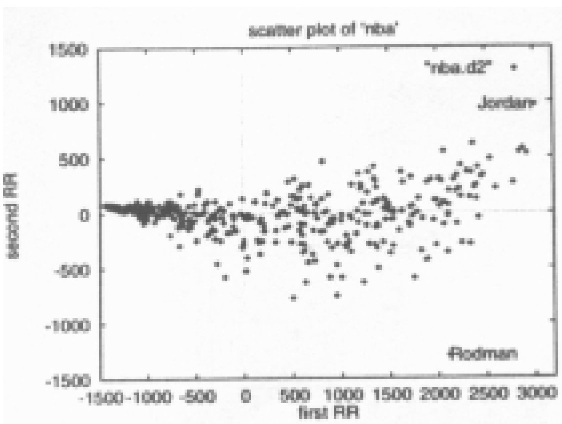
- **Interpretation**
  - RR_1
    - **court action**
  - RR_2
    - **field position**
  - RR_3
    - **height**

field	RR_1	RR_2	RR_3
minutes played	.808	−.4	
field goals			
goal attempts			
free throws			
throws attempted			
blocked shots			
fouls			
points	.406	.199	
offensive rebounds			
total rebounds		−.489	.602
assists			−.486
steals			−.07

P.18

Discussion

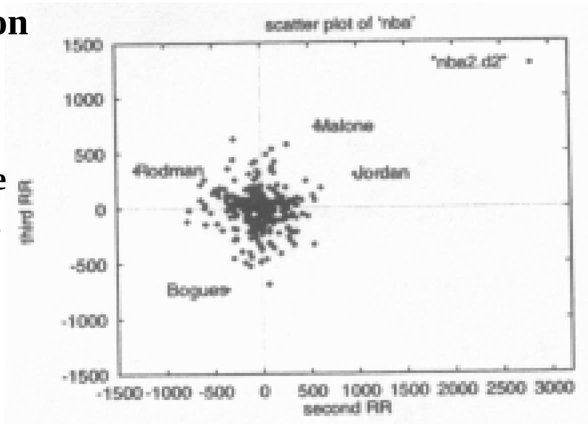
- **Visualization**
  - RR_1
    - **Jordan**
  - RR_2
    - **Jordan**
    - **Rodman**



P.19

Discussion

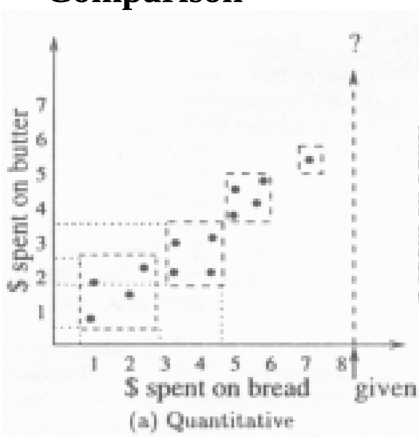
- **Visualization**
  - RR_2
  - RR_3
    - ◆**Malone**
    - ◆**Bogues**

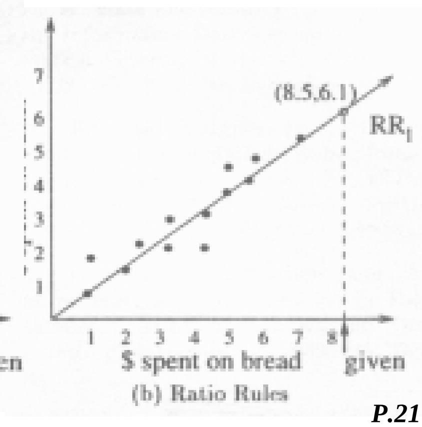


P.20

Discussion

- **Comparison**





P.21

Discussion

- **Advantages of ratio rules**
  - achievement of more compact descriptions if the data points are linearly correlated
  - prediction of one or more unknown values when a new data record is given
  - measure of the guessing error, which can quantify how good a given set of rules is
  - easy to implement
  - only a single pass over the data set is required

P.22