Minimum Variance Associations — Discovering Relationships in Numerical Data

Szymon Jaroszewicz

National Institute of Telecommunications
Warsaw, Poland

PAKDD 2008

Frequent pattern mining

Frequent itemset mining

Given a binary table find all sets of attributes such that

$$\text{supp}(I) = \frac{|\{t \in \mathcal{D} : t[I] = (1, 1, \ldots, 1)\}|}{|\mathcal{D}|} \geq \text{min}_{\text{supp}}$$

Frequent pattern mining

Frequent itemset mining

Given a binary table find all sets of attributes such that

$$\mathrm{supp}(I) = \frac{|\{t \in \mathcal{D} : t[I] = (1, 1, \ldots, 1)\}|}{|\mathcal{D}|} \geq \mathsf{min}_{\mathrm{supp}}$$

- Defined for binary datasets
- Easy extension to categorical attributes
- Applicable to: trees, graphs, etc.
- but... how to do it for numerical attributes?

How to do it for numerical attributes?

- Discretization
  - information loss
  - rules split among many intervals

How to do it for numerical attributes?

- Discretization
  - information loss
  - rules split among many intervals
- Recently a few other approaches:
  - definitions of support for numeric data [Steinbach]
  - using ranks [Calders, Goethals, Jaroszewicz]
  - using polynomials [Jaroszewicz, Korzeń]
  - equations discovery [Langley, Dzeroski, Todorovski]

This paper: Minimum Variance Associations

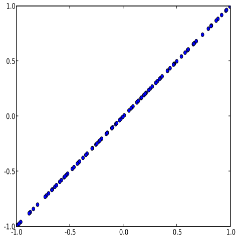
A framework for pattern mining analogous to association rules

- Handles numeric data directly
- Able to discover arbitrary nonlinear relationships

Minimum Variance Associations: main idea

Trivial examples:

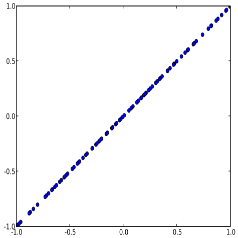
$$x = y$$



Minimum Variance Associations: main idea

Trivial examples:

$$x = y$$

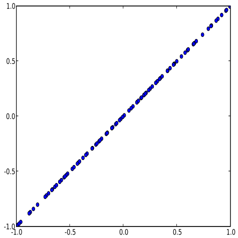


Pattern:

$$F(x, y) = x - y$$
$$= 0 \text{ for all transactions}$$

Trivial examples:

$$x = y$$

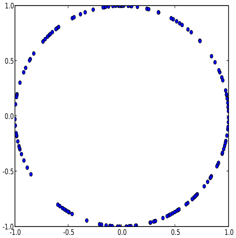


Pattern:

$$F(x, y) = x - y$$

$$= 0 \text{ for all transactions}$$

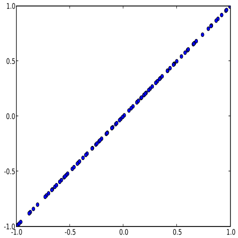
$$x^2 + y^2 = 1$$



Minimum Variance Associations: main idea

Trivial examples:

$$x = y$$

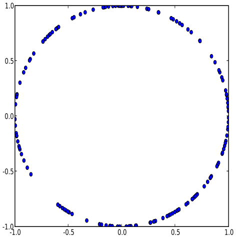


Pattern:

$F(x, y) = x - y$

$= 0$ for all transactions

$$x^2 + y^2 = 1$$



Pattern:

$F(x, y) = x^2 + y^2 - 1$

$= 0$ for all transactions

Minimum Variance Itemsets

Attributes $x_1 x_2 \dots x_n$ are related if there exists a function $F(x_1 \dots x_n)$ which has low variance

$$\sum_{t \in \mathcal{D}} F^2(t[x_1 \dots x_n]) \approx 0$$

These are our itemsets

Minimum Variance Itemsets

Attributes $x_1 x_2 \ldots x_n$ are related if there exists a function $F(x_1 \ldots x_n)$ which has low variance

$$\sum_{t \in \mathcal{D}} F^2(t[x_1 \ldots x_n]) \approx 0$$

Problem

$F \equiv 0$ trivially satisfies all cases.

Minimum Variance Itemsets

Attributes $x_1 x_2 \ldots x_n$ are related if there exists a function $F(x_1 \ldots x_n)$ which has low variance

$$\sum_{t \in \mathcal{D}} F^2(t[x_1 \ldots x_n]) \approx 0$$

subject to additional constraint:

If $x_1, x_2, \ldots, x_n$ were statistically independent then

$$\sum_{t \in \mathcal{D}} F^2(t[x_1 \ldots x_n]) = 1$$

How to find F with minimum variance?

1. Assume F is a polynomial:
   $F(x, y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2$

How to find F with minimum variance?

1. Assume F is a polynomial:
   $F(x, y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2$
2. Compute two matrices $\mathbf{S}_{Data}$ and $\mathbf{S}_{Indep}$:

$$\mathbf{S}_{Data}[1,3] = \sum_{\mathcal{D}} x \cdot xy$$

$$\mathbf{S}_{Indep}[1,3] = \sum_{\mathcal{D}} x \cdot \sum_{\mathcal{D}} xy$$

How to find F with minimum variance?

1. Assume F is a polynomial:
   $F(x, y) = c_0 + c_1 x + c_2 y + c_3 xy + c_4 x^2 + c_5 y^2$
2. Compute two matrices $\mathbf{S}_{Data}$ and $\mathbf{S}_{Indep}$:

$$\begin{aligned} \mathbf{S}_{Data}[1,3] &= \sum_{\mathcal{D}} x \cdot xy \\ \mathbf{S}_{Indep}[1,3] &= \sum_{\mathcal{D}} x \cdot \sum_{\mathcal{D}} xy \end{aligned}$$

3. The coefficient vector $\mathbf{c} = [c_0, \ldots, c_n]$ is a solution of the Generalized Eigenvalue Problem

$$\mathbf{S}_{Data} \cdot \mathbf{c} = \lambda \mathbf{S}_{Indep} \cdot \mathbf{c}$$

Mining algorithm

Monotonicity property

Adding attributes decreases the minimum variance

If an itemset is good, all its supersets are also good.

Mining algorithm

Monotonicity property

Adding attributes decreases the minimum variance

If an itemset is good, all its supersets are also good.

Solution

Find smallest itemsets with given minimum variance.

Simple modification of standard itemset mining algorithms.

Rules

Two types of rules:

Regression rules: $y = F(X)$

Equality rules: $F(X) = G(Y)$

Rules

Two types of rules:

Regression rules: $y = F(X)$

Equality rules: $F(X) = G(Y)$

Variance of rule $F(X) = G(Y)$ is higher than of itemset $X \cup Y$

Rules

Two types of rules:

Regression rules: $y = F(X)$

Equality rules: $F(X) = G(Y)$

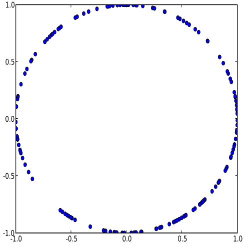
Variance of rule $F(X) = G(Y)$ is higher than of itemset $X \cup Y$

Like standard association rules:

1. First mine itemsets
2. Find rules for each itemset

Simple example

$$x^2 + y^2 = 1$$



itemset: $-1.99 + 1.99x^2 + 1.99y^2$

equality rule: $1.99y^2 = 1.99 - 1.99x^2$

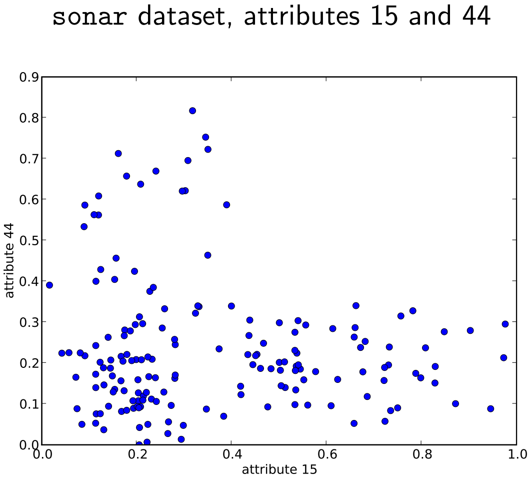
regression rule: none

Examples: extrasolar planets

Itemsets and rules corresponding to:

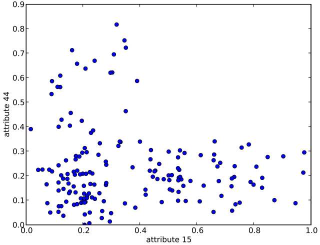
- Trigonometric identity between distance and angular distance
- Kepler's law

Examples: more interesting relationships



Examples: more interesting relationships

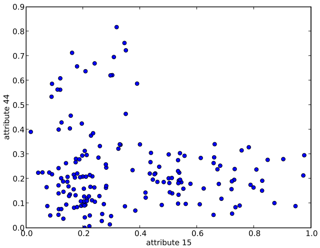
sonar dataset, attributes 15 and 44



- Correlation coefficient $= -0.114$
- No good regression rule
- No good equality rule

Examples: more interesting relationships

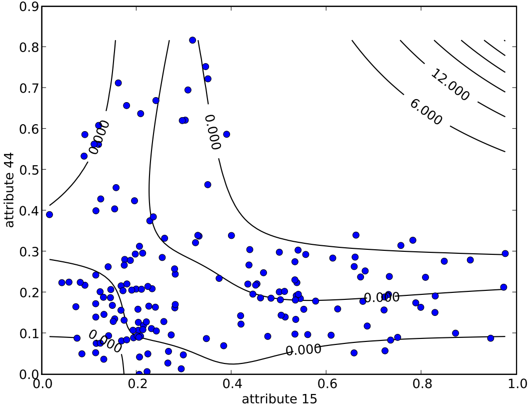
sonar dataset, attributes 15 and 44



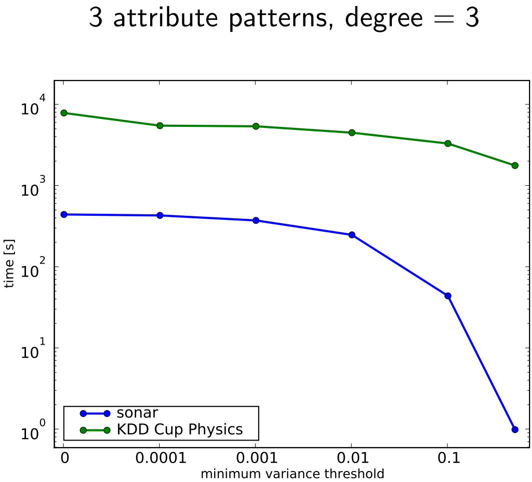
- Correlation coefficient $= -0.114$
- No good regression rule
- No good equality rule
- A minimum variance itemset with variance 0.0001

`sonar` dataset, attributes 15 and 44

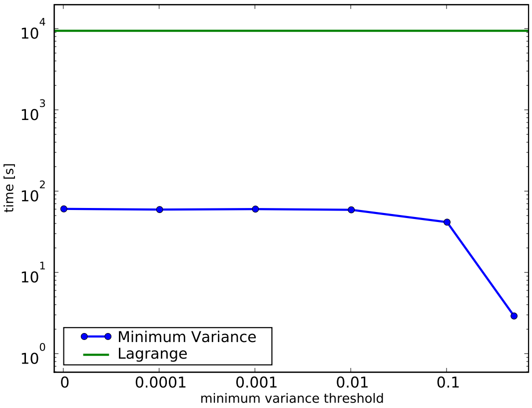
A minimum variance itemset with variance 0.0001



Performance



Performance: comparison with Lagrange equation discoverer



Conclusions:

- Association rule-like framework for numerical data
- Arbitrary non-linear relationships can be discovered efficiently

Future research:

- Background knowledge
- Combine with equation discovery