Theory of Dependence Values

Marc Glisse and Pierre Senellart,
based on an article by Rosa Meo

31st May 2001

Introduction to data mining

Study of recurrent phenomena in a data basis

- Association rules: $x \longrightarrow y$
- Support
- Confidence

Introduction to data mining (2)

Applications:

- Well-suited for the market basket problem
- Not for other applications

Solution: dependence rules

Association rules

- Basket of items

- Absurd example

Dependence rules

Definition 1 *The events* $x_1, \ldots, x_n$ *are independent iff:*

$$P(x_{i_1} \wedge \ldots \wedge x_{i_n}) = P(x_{i_1}) \cdot \ldots \cdot P(x_{i_k}) \text{ for } i_1 < \ldots < i_k$$

Definition 2 *The variables* $A_1, \ldots, A_n$ *are independent iff the events* $(A_i = v_i)_{1 \leq i \leq n}$ *are independent for all values* $v_i \in \{true, false\}$

Theorem 1 *A superset of a dependent set is dependent.*

Notations

- $I_1, I_2, \ldots, I_k$ are k Boolean variables.

- For all j, i_j and $\overline{i_j}$ denote respectively the events "$I_j = true$" and "$I_j = false$".

- Thus, $P(i_1, \overline{i_3}, i_7)$ is the probability of the event:

"$I_1 = true$ and $I_3 = false$ and $I_7 = true$"

Unicity of the value

Theorem 2 *All the kth-order joint probabilities can be calculated as functions of the joint probabilities of the orders less than k and a single kth-order joint probability.*

Definition 3 *The* maximum independence estimate *is the value of $P(i_1, \ldots, i_k)$ maximizing the entropy $\mathcal{E}$ defined as following:*

$$\mathcal{E} = - \sum_{I=(i_1,i_2 \ldots i_k),(\overline{i_1},i_2 \ldots i_k),\ldots,(\overline{i_1},\overline{i_2} \ldots \overline{i_k})} (P(I) \log(P(I)))$$

It is well defined because of the latest theorem.

It is denoted: $P(i_1, \ldots, i_k)_{MI}$

Dependence Value

Definition 4 *The difference:*

$$\Delta = P(i_1, \ldots, i_k) - P(i_1, \ldots, i_k)_{MI}$$

is defined as the dependence value of $(I_1, ...I_k)$.

Theorem 3 *If the joint probabilities up to the order* $k-1$ *are known, the knowledge of the dependence value* Δ *is sufficient to describe all the* k*th order joint probabilities.*

Dependence State

If $|\Delta|$ exceeds a given thresold, $(I_1, \ldots, I_k)$ is said to be *connected by a dependence of order k*. The dependence is either *positive* or *negative*, with the sign of Δ. We denote:

$$D_k(I_1, \ldots, I_k) \gg 0$$

$$D_k(I_1, \ldots, I_k) \ll 0$$

$$D_k(I_1, \ldots, I_k) \sim 0$$

if the dependence is negative, positive or negligible.

Dependence Function

Definition 5 *The* dependence function *of* $(I_1, \ldots, I_k)$ *is the Boolean function of variables* $I_1, \ldots, I_k$ *which takes the value* $true$ *on* $(v_1, \ldots, v_n)$ *if and only if* $P(i_1, \ldots, i_k) > P(i_1, \ldots, i_k)_{MI}$.

Theorem 4 *If* $D_k(I_1, \ldots, I_k) \gg 0$, *the dependence function of* $(I_1, \ldots, I_k)$ *is the parity function with even parity.*

If $D_k(I_1, \ldots, I_k) \ll 0$, *the dependence function of* $(I_1, \ldots, I_k)$ *is the parity function with odd parity.*

Justification for the Maximum Independence definition

2 variables

$\mathcal{E}(A,B) = \mathcal{E}(A) + \mathcal{E}(B|A) = \mathcal{E}(B) + \mathcal{E}(A|B)$

Maximization **3 variables**

Impossible independence

$$\begin{aligned} \mathcal{E}(A,B,C) &= \mathcal{E}(A,B) + \mathcal{E}(C|A,B) \\ &= \mathcal{E}(A,C) + \mathcal{E}(B|A,C) \\ &= \mathcal{E}(B,C) + \mathcal{E}(A|B,C) \end{aligned}$$

Maximum amount of information

An entropy based approach

$$\mathcal{I}(A;B) = \mathcal{E}(A) - \mathcal{E}(A|B)$$

$$\mathcal{E}(A|B) = -\sum_{a^*,b^*} P(a^*,b^*) \cdot \log P(b^*|a^*)$$

Average information content carried by B on A.

More variables?

Algorithm

- k-plets with sufficient support
- Joint "positive" probabilities
- Maximum independence estimates
- Dependence values