Modified Apriori algorithm

Iva Tomanová, Jiří Kupka

Centre of Excellence IT4Innovations
Division of the University of Ostrava
IRAFM
iva.tomanova@osu.cz, jiri.kupka@osu.cz

May 12, 2012

Outline

IRAFM

Motivation

[Novák et. al., 2008]

models of evaluative linguistic expressions and mining of linguistic associations via GUHA method were introduced

[Kupka, Tomanová, 2010]

other mathematical models (based on fuzzy partitions, resp. coverings) were elaborated

[Kupka, Tomanová, 2012]

some properties of fuzzy confirmation measures were studied

Outline

IRAFM

Introduction

We have numerical (real-valued) data in the form

$$\begin{array}{c|cccc} & X_1 & X_2 & \dots & X_m \\ \hline o_1 & f_{11} & f_{12} & \dots & f_{1m} \\ o_2 & f_{21} & f_{22} & \dots & f_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ o_n & f_{n1} & f_{n2} & \dots & f_{nm}. \end{array}$$

- Where o_i are **objects**, $\mathcal{D}_o$ set of objects, X_j are **attributes** and f_{ij} represent values of jth attribute measured on ith object.
- For each attribute X_j we have to specify its **context** $w_j := [a_j, b_j] \subseteq \mathbb{R}$ and its linguistic description.

IRAFM

Linguistic association

Linguistic association

$$\underbrace{A(\{Y_l\}_{l=1}^{p})}_{\text{antecedent}} \Rightarrow \underbrace{B(\{Z_k\}_{k=1}^{q})}_{\text{succedent}}$$

where A, B are conjunctive evaluative linguistic predications.

One of possible forms:

Example: “IF the area of a base of a cylinder is **big** AND the height of this cylinder is **big but not extremely big** THEN the volume of this cylinder is **more or less big**.”

Example of linguistic description

Consider the attribute X on its context (i.e., $[c, d]$) whose covering contains 9 fuzzy sets $\{S_k\}$,

$S_1 \sim$ *Ve Sm*,

$S_2 \sim$ *Sm but not Ve Sm*,

$S_3 \sim$ *ML Sm but not Sm*,

$S_4 \sim$ *Ve Me*,

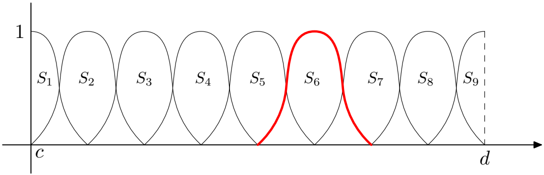
$S_5 \sim$ *Me but not Ve Me*,

$S_6 \sim$ *ML Me but not Me*,

$S_7 \sim$ *Ve Bi*,

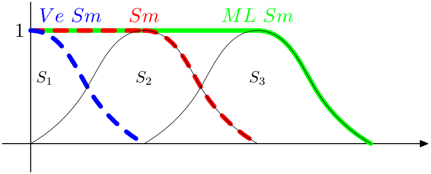
$S_8 \sim$ *Bi but not Ve Bi*,

$S_9 \sim$ *ML Bi but not Bi*.



For “small values” we have

$$S_1 \text{ OR } S_2 \sim Sm,$$
$$S_1 \text{ OR } S_2 \text{ OR } S_3 \sim Sm \text{ OR } S_3 \sim ML\ Sm.$$



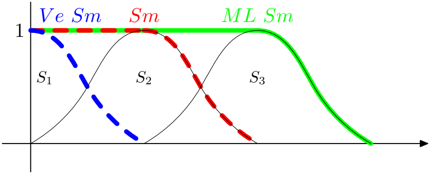
$$S_1 \sim Ve\ Sm,$$
$$S_2 \sim Sm \text{ but not } Ve\ Sm,$$
$$S_3 \sim ML\ Sm \text{ but not } Sm.$$

Specificity ordering of fuzzy sets

- an ordering interpreting evaluative linguistic predications,
- for each $x \in X$: $S'(x) \leq S(x)$,
- denoted by $S' \preceq S$.

Example: Let S, S' denote fuzzy sets from the previous mathematical model.

1. If $S' \sim \mathit{Ve\ Sm}$ and $S \sim \mathit{Sm}$ then $S' \preceq S$,
2. If $S' \sim \mathit{Sm\ but\ not\ Ve\ Sm}$ and $S \sim \mathit{Sm}$ then $S' \preceq S$.



k-itemset S

A set of ordered pairs (l, D_l) where D_l is a fuzzy set from a covering of the context of X_l ($l \in \{1, 2, \dots, m\}$).

There exists a one-to-one correspondence between a conjunction of k linguistic predications and k-itemsets.

Example:
2–itemset: $T = \{(2, D_2), (5, D_5)\}$,
where $D_2 \sim Sm$ and $D_5 \sim Bi\ but\ not\ Ve\ Bi$
"X_2 is *small* AND X_5 is *big but not very big*"

Ordering of itemsets

For a p-itemset $S = \{(i, D_i)\}_{i \in I}$ and q-itemset $T = \{(j, E_j)\}_{j \in J}$ we denote $S \preceq T$ if $I \subseteq J$ and $D_i \preceq E_i$ for any $i \in I$.

Example

If

- 2–itemset: $T = \{(2, D_2), (5, D_5)\}$,
  where $D_2 \sim Sm$ and $D_5 \sim Bi\ but\ not\ Ve\ Bi$
  "X_2 is *small* AND X_5 is *big but not very big*"

- 1–itemset: $S = \{(2, D_2)\}$,
  where $D_2 \sim Ve\ Sm$
  "X_2 is *very small*"

then $S \preceq T$.

Outline

Fuzzy confirmation measures [Dubois et. al., 2006]

Support measures

- *t-norm-based support measure*
  $supp_t(A \Rightarrow B) := \sum_{o \in \mathcal{D}_o} A(o) \otimes B(o)$,
- *minimum-based support measure*
  $supp_m(A \Rightarrow B) := \sum_{o \in \mathcal{D}_o} \min\{A(o), B(o)\}$,
- *implication-based support measure*
  $supp_c(A \Rightarrow B) := \sum_{o \in \mathcal{D}_o} A(o) \cdot (A(o) \rightarrow B(o))$,

Confidence measures

- *confidence measures*
  $conf_p(A \Rightarrow B) := \frac{supp_p(A \Rightarrow B)}{\sum_{o \in \mathcal{D}_o} A(o)}, \ p = t, m, c.$

Outline

1. Introduction
2. Fuzzy confirmation measures
3. **Properties induced by fuzzy confirmation measures**
4. Implementation into Apriori algorithm

Properties induced by fuzzy confirmation measures

Let A, B, B', C, D are evaluative linguistic predications

- $(A \Rightarrow B) \vdash (A \Rightarrow (B \text{ OR } C))$,
- $(A \text{ AND } B) \Rightarrow (C \text{ AND } D) \vdash (A \text{ AND } B \text{ AND } D) \Rightarrow C$.

Background knowledge

- prior information or experience about a given data set
- can be specified by the user
- denoted by $B \Rightarrow^* C$

Using of background knowledge

- $A \Rightarrow B,\ B \Rightarrow^* C \vdash A \Rightarrow C$.

Outline

1. Introduction
2. Fuzzy confirmation measures
3. Properties induced by fuzzy confirmation measures
4. Implementation into Apriori algorithm

Apriori algorithm

- construct a set of all 1–itemsets C_1
- check a cardinality of each $s \in C_1$
- construct a set of frequency 1–itemsets L_1
- form C_2 from L_1 (see Example)
- check cardinality of each $s \in C_2$
- form L_2
- ...until $L_r = \emptyset$
- research combinations of associations (see Example)

Example

$$\begin{array}{c} L_2 \\ \hline \{s_1, s_2\} \\ \{s_1, s_3\} \\ \{s_1, s_4\} \\ \{s_2, s_3\} \end{array} \rightarrow \begin{array}{c} C_3 \\ \hline \{s_1, s_2, s_3\} \end{array}$$

Then $\{s_1, s_2, s_4\} \notin C_3$, because $\{s_2, s_4\} \notin L_2$.

Apriori algorithm

- construct a set of all 1–itemsets C_1
- check a cardinality of each $s \in C_1$
- construct a set of frequency 1–itemsets L_1
- form C_2 from L_1 (see Example)
- check cardinality of each $s \in C_2$
- form L_2
- …until $L_r = \emptyset$
- research combinations of associations (see Example)

Example

If $\{s_1, s_2, s_3\} \in C_3$ then we have to verify these associations

$$\begin{array}{llllllll} s_1 & \Rightarrow s_2 & \texttt{AND} & s_3, & s_2 & \texttt{AND} & s_3 & \Rightarrow s_1, \\ s_2 & \Rightarrow s_1 & \texttt{AND} & s_3, & s_1 & \texttt{AND} & s_3 & \Rightarrow s_2, \\ s_3 & \Rightarrow s_1 & \texttt{AND} & s_2, & s_1 & \texttt{AND} & s_2 & \Rightarrow s_3. \end{array}$$

Modified Apriori algorithm

C_r - sets of candidate r-itemsets

- we start with $C_r = \emptyset$
- construct a set of r-itemsets

$$C_r := \{\{(i, S_{ik})\} \mid S_{ik} \in P(X_i),\ i = 1, 2, \ldots, m\}.$$

For each $s \in C_r$,

$$count(s) = \mathrm{AND}_{l=1}^{m} D_l([o_i]_l),$$

where AND is a relevant t-norm.

IRAFM

L_r - sets of large r-itemsets

Check the $count(s)$ of each $s \in C_r$

(a) If $count(s) \geq \alpha$ then put s into L_r

(b) If $count(s) < \alpha$, then we have to consider "wider" linguistic expressions s' in every attribute satisfying $s \preceq s'$.
If $count(s') \geq \alpha$ then $s' \in L_r$.

(c) For any $s \in L_r$ we may assume that elements of s are ordered by their cardinalities.

Consider $s \in L_r$. Start with 1–itemset in an antecedent.
If $conf_p(s) \geq \gamma$ then

- **(a)** put association s into set $\mathcal{A}$.
- **(b)** $(A \text{ AND } B) \Rightarrow (C \text{ AND } D) \vdash (A \text{ AND } B \text{ AND } D) \Rightarrow C$,
- **(c)** $(A \Rightarrow B) \vdash (A \Rightarrow (B \text{ OR } C))$ that all associations s' are valid if succedent of $s \preceq$ succedent of s',
- **(d)** $A \Rightarrow B,\ B \Rightarrow^* C \vdash A \Rightarrow C$.

(b) $(A \text{ AND } B) \Rightarrow (C \text{ AND } D) \vdash (A \text{ AND } B \text{ AND } D) \Rightarrow C$

If

$$s_1 \Rightarrow s_2 \text{ AND } s_3 \in \mathcal{A}$$

then

$$\begin{aligned} s_1 \text{ AND } s_3 &\Rightarrow s_2 \in \tilde{\mathcal{A}}, \\ s_1 \text{ AND } s_2 &\Rightarrow s_3 \in \tilde{\mathcal{A}}. \end{aligned}$$

(c) $(A \Rightarrow B) \vdash (A \Rightarrow (B \text{ OR } C))$

If

"X_1 is *Sm* AND X_2 is *Me* $\Rightarrow$ X_3 is *Ve Sm*"$\in \mathcal{A}$

then

"X_1 is *Sm* AND X_2 is *Me* $\Rightarrow$ X_3 is *Sm*"$\in \tilde{\mathcal{A}}$,
"X_1 is *Sm* AND X_2 is *Me* $\Rightarrow$ X_3 is *ML Sm*"$\in \tilde{\mathcal{A}}$

(d) $A \Rightarrow B,\ B \Rightarrow^* C \vdash A \Rightarrow C.$

If

"X_1 is *Sm* AND X_2 is *Me* $\Rightarrow$ X_3 is *Ve Sm*"$\in \mathcal{A}$

and

"X_3 is *Ve Sm* $\Rightarrow^*$ X_4 is *Bi but not Ve Bi*"

then

"X_1 is *Sm* AND X_2 is *Me* $\Rightarrow$ X_4 is *Bi but not Ve Bi*"$\in \tilde{\mathcal{A}}$.

IRAFM

Advantages of the proposed algorithm

- more flexible model
- results are interpretable in natural language

Our future work

- to study further properties induced by fuzzy confirmation measures
- to reduce mined linguistic associations in a reasonable way
- to estimate complexity of the proposed algorithm

Thank you for your attention.