A Mass Assignment Approach to Granular Association Rules for Multiple Taxonomies

Trevor Martin, Yun Shen and Ben Azvine
AI Group
Engineering Mathematics
University of Bristol

Currently: Senior Research Fellow,
Intelligent Systems Research Lab, BT

URSW, November 12, 2007
Overview of presentation

- what is the problem?
- research background and the need for fuzziness
- mass assignment-based approach
- demonstrator application
- summary

The aim is to give a high level overview, not to describe the specific methods step-by-step.
For detail, see the paper or talk offline.
Digital obesity needs machine assistance

- Digital obesity (personal and corporate)
  - Britons carry an average of 20GB of data in their pockets (toshiba)
  - estimated 5 exabytes of new data produced globally in 2002, much as text (sims)
  - 15% of enterprise info is in a core database, the rest is on lap/desk tops (HP labs)
  - 80% of a typical “new” document is recycled from existing documents (HP labs)
  - there is a need for automatic assistance in managing this information

- Machine-based solutions must be understandable ...
  “never trust anything that thinks if you can’t see where it keeps its brain” *
  - google uses syntactic features (presence of words, links, ...)
  - computational linguistics / natural language processing aims to use grammar and deep structure / meaning
  - humans use natural language - machines cannot understand NL but can process it
  - engineering approach - be consistent with (fuzzy) humans

* Arthur Weasley, UK Ministry of Magic - Harry Potter and the Chamber of Secrets
Four Steps

Organisation / Granulation: iPHI

Summary: approximate relations (class - class)

Most films in the “space adventure” category are directed by Lucas or Spielberg

Fusion: SOFT

pre-processing: fuzzy grammars, entity tagging
Granulation and Hierarchies

- Hierarchical organisation is widespread

- Each category in the hierarchy is a granule
  - e.g. shapes, geography, wine, species, books, movies, documents, ...

- There is rarely a unique hierarchy, and categories are rarely crisp

- Leaf nodes are often categorical values in a database
Equivalence of hierarchies

- it is helpful to know how different hierarchies (views) are related
  - enables reuse of categorised information
  - enables combination of information from different sources

- "DL approach" based on binary logic, requires equivalence / strict subset ... \( \forall, \exists \)

- better approach - look for strong associations
  - "most current customers are satisfied customers"
Association Rules

- most prominent use - transaction analysis
  - how often do customers buy cheese or crackers when they buy wine or beer?
  - internet retailers: “customers who purchased A and B also purchased Y and Z”

<table>
<thead>
<tr>
<th></th>
<th>beer</th>
<th>wine</th>
<th>cheese</th>
<th>crackers</th>
<th>milk</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

$s = \{\text{beer, wine}\}$  \hspace{1cm}  $t = \{\text{cheese, crackers}\}$

$S, T \subseteq Tr$ are (multi-)sets of transactions containing items in $s, t$ respectively

$Support(S, T) = |S \cap T|$  \hspace{1cm}  $Conf(S, T) = \frac{|S \cap T|}{|S|}$
Fuzzy Categories

- words mean what we agree that they mean
  - wicked, cool, bling, chav, rubbish
  - light snowfall, bright colour, rock music
  - what makes a White Christmas in the UK? “That one flake of snow will fall on Met office monitoring stations over the 24 hr period of the 25th of December”
    www.mybetting.co.uk/white-christmas-betting.htm

- communication is made more efficient by use of loose definitions

- over-precision is not user-friendly- it needs soft definitions
  - most relations / categories / attributes / ... are not defined by unbreakable rules,
  - data can be missing, inconsistent, unreliable, ...
  - should we adapt our thinking to the computer or adapt the computer to our thinking
An “illogical” hierarchy - dairy-free spread

Bakery
Beers, Wines, Spirits
Beverages, Hot Drinks
Breakfast Cereals
Clothing
Confectionery, Biscuits, Cakes
Cooking/Baking Ingredients
Crisps, Nuts, Snacks
Dairy
Delicatessen
Easter Confectionery
...
Pickles, Preserves, Oils, Spreads
...

Animal and Vegetable Fats
Artificial Sweeteners
Colouring and Decoration
Custard and Cornflour
Dried Fruit
Flour - Other
...
Cheese - American
Cheese - Canadian
...
Cheese - Snacking
...
Spreads (Butter, Margarine, etc.)
...
Yogurt - Twin Pots

Dairy Free Spread 500g
dairy-free classified as dairy.
Not logic!
But it works.

No luck …
Fuzzy Sets in Information Systems

• Logic approaches are too “black and white”
  ✷ data must be known precisely or not at all (confusion over NULL values)
  ✷ the real world is rarely as neat and tidy as this
  ✷ uncertainty in data values -
    e.g. John is quite heavy, John’s car is a small hatchback
  ✷ uncertainty in relations -
    e.g. John hates Microsoft Word a lot
  ✷ both
    e.g. driving in a small car at high speed is very uncomfortable
  ✷ uncertainty in deduction
    a car usually has high running costs if its list price is expensive

• how can the logical model be extended?
  ✷ incorporation of uncertainty in attribute values, relations, rules, inference, and queries
  ✷ fuzzy - lack of adequate definition e.g. what is meant by large number
  ✷ probabilistic - lack of information e.g. will dice score be 6?
Possibility Distributions vs Monadic Fuzzy Relations

- speed of car is fast $\mu_{fast}(70)$: speed of car is 70 mph OR
  $\mu_{fast}(71)$: speed of car is 71 mph OR ...

- John is speaking L $\mu_L$ (spanish): John is speaking spanish OR
  $\mu_L$ (portugese): John is speaking portugese OR ...
  - single value, not known precisely

- legal speed is about 55 $\chi_{a55}(55)$: legal speed is 55 mph AND
  $\chi_{a55}(56)$: legal speed is 56 mph AND ...

- John speaks L fluently $\chi_L$ (spanish): John speaks spanish AND
  $\chi_L$ (portugese): John speaks portugese AND ...
  - multiple values, all satisfy predicate to some degree
Fuzzy Association Rules

\[ \text{Support}(S, T) = |S \cap T| \quad \quad \text{Conf}(S, T) = \frac{|S \cap T|}{|S|} \]

- straightforward fuzzification
  - allow \( S, T \) to be fuzzy (multi-)sets (monadic relations)
  - use t-norm (min) for intersection and sigma-count for cardinality

\[ \text{Conf}(S, T) = \frac{\sum_{x \in X} \mu_{S \cap T}(x)}{\sum_{x \in X} \mu_S(x)} \]

\begin{tabular}{|c|c|c|}
\hline
name & sales & salary \\
\hline
a & 100 & 1000 \\
b & 80 & 400 \\
c & 50 & 800 \\
d & 20 & 700 \\
\hline
\end{tabular}

\( S = \text{goodSales} = [a/1, b/0.8, c/0.5, d/0.2] \)
\( T = \text{highSalary} = [a/1, b/0.4, c/0.8, d/0.7] \)

\[ \text{Conf}(S, T) = \frac{1 + 0.4 + 0.5 + 0.2}{1 + 0.8 + 0.5 + 0.2} = 0.84 \]
Mass assignment-based approach

Fuzzy set \( A = \{a/1, b/0.8, c/0.3, d/0.2\} \)

\[ \Rightarrow \text{alpha cuts } \{a\} /1, \{a, b\} /0.8, \{a, b, c\} / 0.3, \{a, b, c, d\} /0.2 \]

\[ \Rightarrow \text{mass assignment} \]

\[ M(A) = \{(a) : 0.2, \{a, b\} : 0.5, \{a, b, c\} : 0.1, \{a, b, c, d\} : 0.2\} \]

- interpretation (as a possibility distribution)
  - value is in \( \{a\} \) with probability mass 0.2
  - value is in \( \{a, b\} \) with probability mass 0.5
  - etc

Mass can be distributed between elements of a set (restriction) - may not correspond to fuzzy set

- e.g. \( \{a, b, c, d\} : 0.2 \Rightarrow \{a, b, c\} : 0.1 \text{ and } \{a, c\} : 0.1 \)

\[ M_R(A) = \{(a) : 0.2, \{a, b\} : 0.5, \{a, b, c\} : 0.2, \{a, c\} : 0.1\} \]

- Combination of 2 mass assignments - re-distribute mass in a way that is consistent with both original mass assignments

- cardinality \( p(|A| = n) = \sum_{A_i \subseteq A \atop |A_i| = n} m_A(A_i) \)

\[ p(|A| = 1) = 0.2, \ p(|A| = 2) = 0.5, \text{ etc} \]

Least prejudiced distribution (LPD)

- split mass equally between elements

\( \{a, b, c, d\} : 0.2 \Rightarrow \{a\} : 0.05, \{b\} : 0.05 \text{ etc} \)
Mass Assignment Association Rules

\[ S = \text{goodSales} = [a/1, b/0.8, c/0.5, d/0.2] \quad T = \text{highSalary} = [a/1, b/0.4, c/0.8, d/0.7] \]

<table>
<thead>
<tr>
<th></th>
<th>0.2</th>
<th>0.1</th>
<th>0.3</th>
<th>0.4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>a</td>
<td>a</td>
<td>ac</td>
<td>a</td>
</tr>
<tr>
<td>0.2</td>
<td>0.2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>a</td>
<td></td>
<td></td>
<td>ac</td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ab</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>0.1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ab</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.3</td>
<td></td>
<td>0.3</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>abc</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.2</td>
<td></td>
<td></td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>abd</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Assign mass in a way consistent with both components

Many possibilities!

Calculate support, confidence in standard (crisp) way

min and max give interval \([0.4, 1]\) in this case (potentially expensive computation)

Alternative - use least prejudiced distribution - very fast (see paper)
Demonstrator Application

Sample questions:
“who are the most active terrorist groups in countries near to Iran?”
“who are the main targets, and has that answer changed recently?”
Simple iPHI hierarchy

source data uses a given hierarchical categorisation - iPHI transforms to a more convenient (and fuzzy) subdivision

incidents in all regions

middle east

south asia

etc

incidents in all regions

near Iran

near Pakistan

etc.

Iran, Iraq, ... Afghanistan, Pakistan, India, ...

JAXB converts XML representation of category nodes and instances into executable code (java)
Demo

SQUAD | Visualisation

Top 10 Associations

- Eurasia  Secular P.: 0.257
- South Asia  Secular P.: 0.203
- Near Israel  Islamic Ext.: 0.16
- Near Indo... Secular P.: 0.129
- North Am... Environn...: 0.125
- North Am... Secular P.: 0.125
- Near Sout... Secular P.: 0.124
- Near Israel  Secular P.: 0.1
- Near Sour... Islamic Ext.: 0.062
- Eurasia  Islamic Ext.: 0.057
Summary

- soft relation = extended form of association rule
  - transaction analysis - e.g. do customers generally buy crisps or nuts when beer, lager or wine are bought?
  - rule $S \Rightarrow T$ (e.g. $S = \text{beer/lager/wine}$ $T = \text{crisps/nuts}$), rule confidence is the conditional probability of $T$ given $S$
  - extended to cope with fuzzy categories $S$, $T$ e.g. do customers generally buy salty snacks when alcoholic drinks and soft drinks are bought?
  - fast calculation if we use $lpd$
  - now looking at changes in association strength over time

```
all incidents
  └── high casualty
  └── low casualty
      └── no casualty

all incidents
  └── near iran
      └── not near iran
```
Thank you for your attention