Hypergraphs Transversals in Association Rule Mining

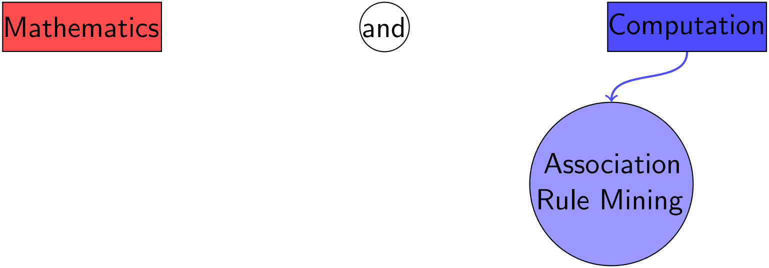
Cristina Tîrnăucă

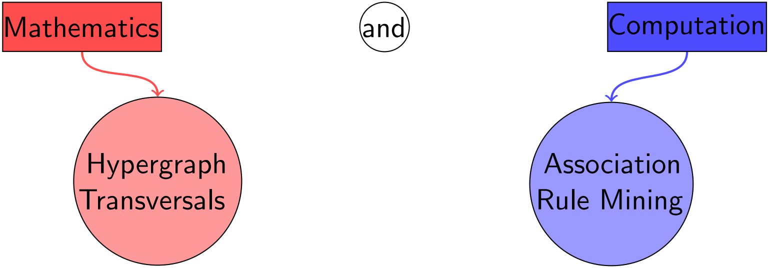
Departmento de Matemáticas, Estadística y Computación
University of Cantabria

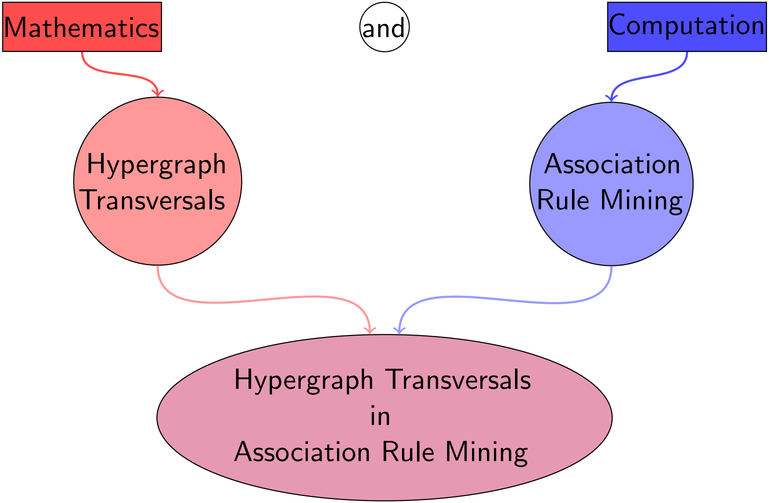
Congreso de Jóvenes Investigadores
Real Sociedad Matemática Española

Universidad de Sevilla, septiembre de 2013

Mathematics and Computation


Mathematics
and
Computation
Association Rule Mining


Mathematics
and
Computation
Hypergraph Transversals
Association Rule Mining


Mathematics
and
Computation
Hypergraph Transversals
Association Rule Mining
Hypergraph Transversals
in
Association Rule Mining

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.
- Random prediction hardly brings any advantage: we want to do it better than randomly

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.
- Random prediction hardly brings any advantage: we want to do it better than randomly
  - for this, we must rely on something

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.
- Random prediction hardly brings any advantage: we want to do it better than randomly
  - for this, we must rely on something
  - for example, on available data

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.
- Random prediction hardly brings any advantage: we want to do it better than randomly
  - for this, we must rely on something
  - for example, on available data
  - but, if we have all data, there is nothing to predict

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.
- Random prediction hardly brings any advantage: we want to do it better than randomly
  - for this, we must rely on something
  - for example, on available data
  - but, if we have all data, there is nothing to predict
- Essential ingredient: uncertainty.

Data Mining

Goal: obtaining an economic advantage (most of the times)

- The intention is to achieve it through successful predictions, at least partially.
- Random prediction hardly brings any advantage: we want to do it better than randomly
  - for this, we must rely on something
  - for example, on available data
  - but, if we have all data, there is nothing to predict
- Essential ingredient: uncertainty.
- One of the many ways to handle uncertain knowledge (the most relevant to data mining, but not the only one) is the statistic approach, based on the probabilities theory.

Rule Mining
Binary Data

Implications: Horn clauses with the same antecedent.
"(Rich $\Rightarrow$ Male) $\wedge$ (Rich $\Rightarrow$ White)" $=$ "(Rich $\Rightarrow$ Male, White)"

Rule Mining
Binary Data

Implications: Horn clauses with the same antecedent.
"(Rich $\Rightarrow$ Male) $\wedge$ (Rich $\Rightarrow$ White)" $=$ "(Rich $\Rightarrow$ Male, White)"

Properties: a, b, c, d;
Observations: t_1, t_2, t_3;

ID	a	b	c	d
t_1	1	1	0	1
t_2	0	1	1	1
t_3	0	1	0	1

transaction
$\{a, b, d\}$
$\{b, c, d\}$
$\{b, d\}$

$d \Rightarrow b$
$a, b \Rightarrow d$
$a \Rightarrow b, d$
$\ldots$

Rule Mining
Binary Data

Implications: Horn clauses with the same antecedent.
"(Rich $\Rightarrow$ Male) $\wedge$ (Rich $\Rightarrow$ White)" $=$ "(Rich $\Rightarrow$ Male, White)"

Properties: a, b, c, d;
Observations: t_1, t_2, t_3;

ID	a	b	c	d
t_1	1	1	0	1
t_2	0	1	1	1
t_3	0	1	0	1

transaction
$\{a, b, d\}$
$\{b, c, d\}$
$\{b, d\}$

$d \Rightarrow b$
$a, b \Rightarrow d$
$a \Rightarrow b, d$
$\ldots$

Relational case: a boolean attribute for each attribute-value pair

Examples

Implications in real data

ML Abstracts Data Set

Abstracts of research articles in Machine Learning:

support, margin $\Rightarrow$ vector

descent $\Rightarrow$ gradient

hilbert $\Rightarrow$ space

carlo $\Rightarrow$ monte

monte $\Rightarrow$ carlo

Examples

Implications in real data

ML Abstracts Data Set

Abstracts of research articles in Machine Learning:

support, margin $\Rightarrow$ vector

descent $\Rightarrow$ gradient

hilbert $\Rightarrow$ space

carlo $\Rightarrow$ monte

monte $\Rightarrow$ carlo

Adult Data Set (`http://archive.ics.uci.edu/ml/datasets/Adult`)

United States Census:

Exec-managerial, Husband $\Rightarrow$ Married-civ-spouse

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\longrightarrow$ United-States, White
- Husband $\longrightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\longrightarrow$ Husband, Male

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K
- Black $\rightarrow$ $\leq$ 50K, United-States

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow \leq$ 50K
- Black $\rightarrow \leq$ 50K, United-States
- Adm-clerical, Private $\rightarrow \leq$ 50K

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K
- Black $\rightarrow$ $\leq$ 50K, United-States
- Adm-clerical, Private $\rightarrow$ $\leq$ 50K
- Self-emp-not-inc $\rightarrow$ Male

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K
- Black $\rightarrow$ $\leq$ 50K, United-States
- Adm-clerical, Private $\rightarrow$ $\leq$ 50K
- Self-emp-not-inc $\rightarrow$ Male
- $\leq$ 50K , Sales $\rightarrow$ Private

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K
- Black $\rightarrow$ $\leq$ 50K, United-States
- Adm-clerical, Private $\rightarrow$ $\leq$ 50K
- Self-emp-not-inc $\rightarrow$ Male
- $\leq$ 50K , Sales $\rightarrow$ Private
- hours-per-week:50 $\rightarrow$ Male

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K
- Black $\rightarrow$ $\leq$ 50K, United-States
- Adm-clerical, Private $\rightarrow$ $\leq$ 50K
- Self-emp-not-inc $\rightarrow$ Male
- $\leq$ 50K , Sales $\rightarrow$ Private
- hours-per-week:50 $\rightarrow$ Male
- Female, Some-college $\rightarrow$ $\leq$ 50K

Association Rules

"Implications" that allow "exceptions"

In the United States Census, in more than 2/3 of all cases:

- $\rightarrow$ United-States, White
- Husband $\rightarrow$ Male, Married-civ-spouse
- Married-civ-spouse $\rightarrow$ Husband, Male
- Not-in-family $\rightarrow$ $\leq$ 50K
- Black $\rightarrow$ $\leq$ 50K, United-States
- Adm-clerical, Private $\rightarrow$ $\leq$ 50K
- Self-emp-not-inc $\rightarrow$ Male
- $\leq$ 50K , Sales $\rightarrow$ Private
- hours-per-week:50 $\rightarrow$ Male
- Female, Some-college $\rightarrow$ $\leq$ 50K
- Divorced $\rightarrow$ $\leq$ 50K

Implication Intensity

How to measure how few are the exceptions

Usual proposal: the confidence:

$$c(X \to Y) = \frac{s(XY)}{s(X)}$$

where $s(X)$ is the support of X: the number of observations in which the event X occurs.

In favor:

- it is relatively natural
- it is easy to explain to a non-expert user

Requires careful handling:

- High levels of confidence do not prevent against negative correlations.

Alternative Metrics

For selecting association rules

Implication intensity criteria:

- confidence
- lift (originally called interest) [1]
- collective strength [1]
- Gini Index
- leverage [2]
- all-confidence [3]
- conviction [1]
- ...

Standard Association Rules

- fix some minimal support and confidence thresholds
- compute "frequent itemsets": sets of items whose support is greater than the threshold
- for each frequent itemset, try all ways of taking one item as a consequent and the rest as antecedents, and filter those rules that do not reach the confidence threshold

(We loose all rules with more than one consequent.)

Standard Association Rules

- fix some minimal support and confidence thresholds
- compute "frequent itemsets": sets of items whose support is greater than the threshold
- for each frequent itemset, try all ways of taking one item as a consequent and the rest as antecedents, and filter those rules that do not reach the confidence threshold

(We loose all rules with more than one consequent.)

Alternative: for each frequent itemset X and each $Y \subseteq X$, calculate $c(X \setminus Y \rightarrow Y)$ and return only those rules that meet the confidence threshold.

Association Rules in Practice

Pros

- reasonably efficient algorithms (and open source).
- pre-processing data is not trivial but easy enough
- relatively little training is enough to understand the result

Cons

- We need to adjust support and confidence thresholds with the famous method K.T.: keep trying
- Which measure should we use for the intensity of implication? How do we explain it to the final user?
- But, this is not the worse...

Abundance of Association Rules

The major problem with this approach

You pre-process your data, run your associator, adjust your parameters . . . and when you finally guess values that give you sufficiently interesting rules . . .

Abundance of Association Rules

The major problem with this approach

You pre-process your data, run your associator, adjust your parameters . . . and when you finally guess values that give you sufficiently interesting rules . . . you get dozens of thousands of rules. And many of them you could easily spare...
First 56 rules shown for the adult dataset (Weka, default values):

- husband $\longrightarrow$ male
- married-civ-spouse, husband $\longrightarrow$ male
- husband $\longrightarrow$ married-civ-spouse
- husband, male $\longrightarrow$ married-civ-spouse
- husband $\longrightarrow$ married-civ-spouse, male
- married-civ-spouse, male $\longrightarrow$ husband
- . . .

Removing Redundancy in the Output

We need:

- precise notions of redundancy between association rules
- methods to find irredundant minimal "bases" (min-max approximate basis [1], Representative (or Essential) Rules [2, 2], Confidence Boost [3], . . .)
- and ways to discard rules that are not novel

Approaches addressing this issue can be classified into three main trends:

- provide mechanisms for filtering extracted association rules
- allow the analyst to define some templates in order to select rules according to his/her preferences
- use some formal measures of rule interestingness directly into the mining process so that the output would be smaller

What Can Cause Redundancy?

Non Minimal Antecedents

- $\longrightarrow$ United States, White
- Husband $\longrightarrow$ United-States, White
- Married-civ-spouse $\longrightarrow$ United-States, White
- ...

Non Maximal Consequents http://archive.ics.uci.edu/ml/datasets/Mushroom

- free gills $\longrightarrow$ edible
- free gills $\longrightarrow$ edible, partial veil
- free gills $\longrightarrow$ edible, white veil
- free gills $\longrightarrow$ edible, partial veil, white veil

Minimal Generators, Closed Sets

Closed Sets

A set X is called closed if for all $Y \supset X$, $s(Y) < s(X)$.

Computing closed sets is easy.

Minimal Generators

A set X is called minimal generator if for all $Y \subset X$, $s(Y) > s(X)$.

Computing minimal generators is not.

Hypergraph Transversals

Hypergraphs and Hypergraph Transversals

- A hypergraph is a family of subsets (hyperedges) of a finite set of vertices.
- A transversal, also called hitting set, is a subset of vertices that intersects each and every hyperedge of the hypergraph. It is minimal if none of its proper subsets is a transversal.

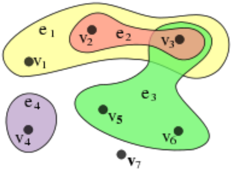
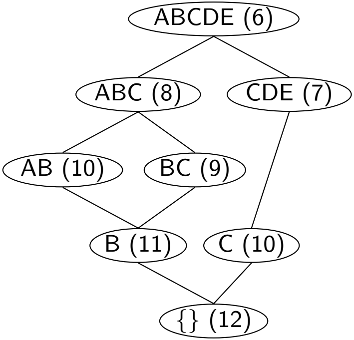


Figure: Hypergraph Example (Source: Wikipedia)

Minimal Generators as Minimal Transversals

It turns out that minimal generators are minimal transversals of a certain simple[1] hypergraph (in which its hyperedges are formed by taking the complements with respect to a given closed set of its maximal closed subsets).

t_1	$\{A, B, C, D, E\}$
t_2	$\{A, B, C, D, E\}$
t_3	$\{A, B, C, D, E\}$
t_4	$\{A, B, C, D, E\}$
t_5	$\{A, B, C, D, E\}$
t_6	$\{A, B, C, D, E\}$
t_7	$\{A, B, C\}$
t_8	$\{A, B, C\}$
t_9	$\{C, D, E\}$
t_{10}	$\{A, B\}$
t_{11}	$\{A, B\}$
t_{12}	$\{B, C\}$

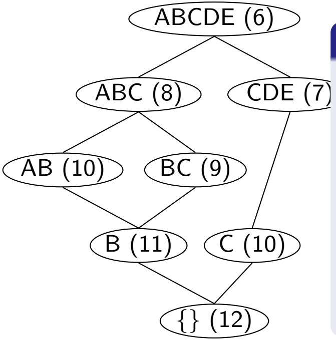


[1]a hypergraph is simple if no hyperedge is strictly included in another one

Minimal Generators as Minimal Transversals

It turns out that minimal generators are minimal transversals of a certain simple[1] hypergraph (in which its hyperedges are formed by taking the complements with respect to a given closed set of its maximal closed subsets).

t_1	$\{A, B, C, D, E\}$
t_2	$\{A, B, C, D, E\}$
t_3	$\{A, B, C, D, E\}$
t_4	$\{A, B, C, D, E\}$
t_5	$\{A, B, C, D, E\}$
t_6	$\{A, B, C, D, E\}$
t_7	$\{A, B, C\}$
t_8	$\{A, B, C\}$
t_9	$\{C, D, E\}$
t_{10}	$\{A, B\}$
t_{11}	$\{A, B\}$
t_{12}	$\{B, C\}$



Ex: MinGen(ABCDE)

- maximal closed subsets: ABC & CDE
- take complements: DE & AB
- construct hypergraph: {A,B},{D,E}
- compute minimal transversals: AD, AE, BD, BE

[1] a hypergraph is simple if no hyperedge is strictly included in another one

The Bad News . . .

It is an open problem whether all minimal transversals can be computed in output-polynomial time (i.e., in time polynomial in the combined sizes of the input and the output).

Moreover, for the related problem of deciding whether a given hypergraph $\mathcal{G}$ is the transversal hypergraph of $\mathcal{H}$ (known to be in co-NP) there is no proof of co-NP completeness.

Conclusion

By applying data mining techniques to our records from the Congress of Young Researchers organized by the Spanish Royal Mathematical Society

Speaker	Length "long"	Length "short"	Subject "boring"	Subject "interesting"	No questions asked
John	1	0	1	0	1
Mary	0	1	0	1	0
Pete	0	1	0	1	1
Ron	0	1	1	0	0
Elisabeth	1	0	1	0	1
...					

we can state, with high confidence, that

Length = "long", Subject = "boring" $\rightarrow$ No questions asked

References

C. C. Aggarwal and P. S. Yu.
A new framework for itemset generation.
In A. O. Mendelzon and J. Paredaens, editors, *PODS*, pages 18–24. ACM Press, 1998.

C. C. Aggarwal and P. S. Yu.
A new approach to online generation of association rules.
IEEE Trans. Knowl. Data Eng., 13(4):527–540, 2001.

J. L. Balcázar.
Formal and computational properties of the confidence boost in association rules.
To appear in ACM Transactions on Knowledge Discovery from Data.

References, II

S. Brin, R. Motwani, J. D. Ullman, and S. Tsur.
Dynamic itemset counting and implication rules for market basket data.
In *SIGMOD Conference*, pages 255–264, 1997.

M. Kryszkiewicz.
Representative association rules.
In X. Wu, K. Ramamohanarao, and K. B. Korb, editors, *PAKDD*, volume 1394 of *Lecture Notes in Computer Science*, pages 198–209. Springer, 1998.

E. Omiecinski.
Alternative interest measures for mining associations in databases.
IEEE Trans. Knowl. Data Eng., 15(1):57–69, 2003.

References, III

N. Pasquier, R. Taouil, Y. Bastide, G. Stumme, and L. Lakhal.
Generating a condensed representation for association rules.
J. Intell. Inf. Syst., 24(1):29–60, 2005.

G. Piatetsky-Shapiro.
Discovery, analysis, and presentation of strong rules.
In *Knowledge Discovery in Databases*, pages 229–248.
AAAI/MIT Press, 1991.