Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Mining Unexpected Multidimensional Rules

M. Plantevit*, S. Goutier+, F. Guisnel+, A. Laurent* and M. Teisseire*

*: LIRMM, Université Montpellier II, France
+: EDF R&D, Clamart, France

DOLAP'07
Nov. 9, 2007, Lisbon, Portugal

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Agenda

1. Introduction
2. Related Work
3. Proposal
   - Specification of wild carded premisses
   - Difference in the conclusion
   - Unexpected Multidimensional Sequential Rules
   - Algorithm
4. Experiments
5. Conclusion

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

OLAP & KDD

User Navigation

- OLAP users are now decision makers.
- Users navigate in the aggregated datacube in order to discover knowledge.
  - ROLL UP, DRILL DOWN, ...

Our Goal:

Providing automatically knowledge thanks to data mining approaches

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

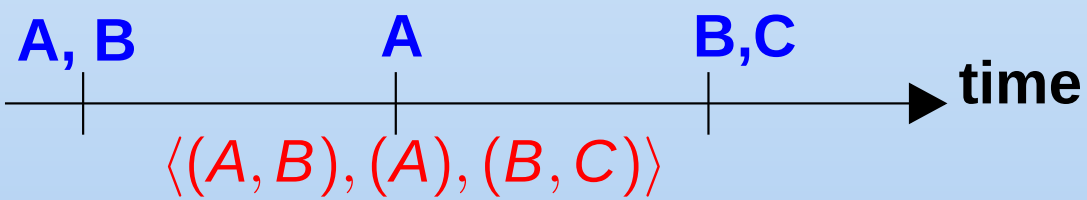
Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Sequential Patterns

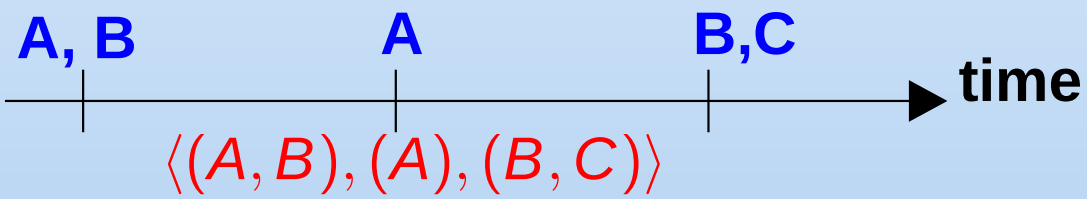
- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .



Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Sequential Patterns

- Well adapted for temporal data
- Discovering correlations between events through time.
- Several applications: marketing, decision making, protein sequence, network security, music, . . .

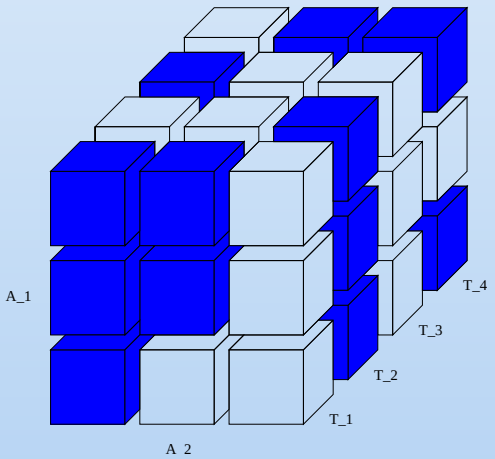


- ☹: Sequential patterns are quite poor (only one mined dimension)

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Data Cube

- Knowledge are mined among one dimension: *product* dimension.
- What about the other ones ?



Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Multidimensional Sequential Patterns

Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Multidimensional Sequential Patterns

Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions
- Classical item: c

Multidimensional Sequential Patterns

Combining Several Analysis Dimensions

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

- Items are not defined on one dimension, they are defined on several dimensions
- Classical item: c
- Multidimensional item: $(Young, c, 100), (Old, c, *)$

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Multidimensional Sequential Patterns

Combining Several Analysis Dimensions

- Items are not defined on one dimension, they are defined on several dimensions
- Classical item: c
- Multidimensional item: $(Young, c, 100), (Old, c, *)$
- Multidimensional sequence:

$$\langle \{(Young, c, 100), (Young, d, 54)\}\{(*, b, 2)\}\rangle$$

instead of

$$\langle (c, d), b\rangle$$

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Problem

Large number of rules

- unmanageable for any human user.

Type of knowledge

- Frequent and Confident Rules
- useless for unexpected relations

10%-90% Rule

- The occurrence of a frequent event carries less information than the occurrence of a rare or hidden event.

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Our goal

Mining Unexpected Rules

- Multidimensional framework;
- Time

Example

- CR: *If a young customer has recently received his car license then he will buy a little car.*
- UR: *If a young customer who likes surfing has received his car license then he will buy a van.*

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Agenda

1 Introduction

2 Related Work

3 Proposal
- Specification of wild carded premisses
- Difference in the conclusion
- Unexpected Multidimensional Sequential Rules
- Algorithm

4 Experiments

5 Conclusion

Unexpected Knowledge

A frequent event carries less information than a rare or hidden event.

- Cohen et al., 2001.
- Suzuki et al., 2005.
- Yang et al., 2004.

Two philosophies

- user-driven or subjective approaches;
- data-driven or objective approaches.

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Data-Driven Approaches

Two subfields

Use of interestingness measures

- different from usual support and confidence.

Try to discover unexpected knowledge which are not extracted by classical algorithms

- unfrequent patterns but highly correlated;
- rare itemsets;
- exceptions;
- anomalous rules.

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

No approach for mining unexpected rules

- Multidimensional framework,
- Ordered relation (e.g. time)

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Agenda

1. Introduction
2. Related Work
3. Proposal
   - Specification of wild carded premisses
   - Difference in the conclusion
   - Unexpected Multidimensional Sequential Rules
   - Algorithm
4. Experiments
5. Conclusion

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Overview

Goal

- Extract unexpected rules which are often hidden by common rules with high support.

Main Idea

- to use wild card values in the if-part of a C.R.
- detect a different and unexpected conclusion

- CR (frequent, confident): $P \rightarrow Q$
- UR (unfrequent, confident) : $P_{specialized} \rightarrow Q'$

Multidimensional Sequential Rules

$$\alpha = \langle i_1, i_2, \ldots i_k, i_{k+1}, \ldots, i_n \rangle$$

Multidimensional Sequential Rule

An implication:

$$R : \langle i_1, i_2, \ldots i_k \rangle \rightarrow \langle i_{k+1}, \ldots, i_n \rangle$$

Support

$$support(R) = support(\langle i_1, i_2, \ldots i_k, i_{k+1}, \ldots, i_n \rangle)$$

Confidence

$$Conf(R) = \frac{support(\langle i_1, i_2, \ldots i_k, i_{k+1}, \ldots, i_n \rangle)}{support(\langle i_1, i_2, \ldots i_k \rangle)}$$

Functions

- $*_\lambda(x)$ such that $*_\lambda(x) = x$
- a_{i_λ} such that

$$a_{i_\lambda}(x) = \begin{cases} a_i & \text{if} \quad x = a_i \\ \emptyset & \text{otherwise} \end{cases}$$

Application to item

- Item $C = (t_1, t_2, \ldots, t_m)$,
- we can construct the function C_λ:

$$X = (x_1, \ldots, x_m) \mapsto C_\lambda(X) = (t_{1_\lambda}(x_1), t_{2_\lambda}(x_2), \ldots, t_{m_\lambda}(x_m))$$

we can construct from an item $C = (a, *, c)$ the function $C_\lambda = (a_\lambda, *_\lambda, c_\lambda)$.

Instance

Let C and X be two multidimensional items, X is said to be an instance of C if $C_\lambda(X) = X$.

Example

- $X = (a, b, c)$ is an instance of $C = (a, *, c)$ since $C_\lambda(X) = X$.
- We can denote that X is an instance of itself ($X_\lambda(X) = X$).

Pseudo-instance of an itemset

Let $i = \{e_1, e_2, \ldots, e_m\}$ and $i' = \{e'_1, e'_2, \ldots, e'_{m'}\}$ be two itemsets such that $m \leq m'$, i is said to be a pseudo-instance of i' if there exist some integers $1 \leq k_1 \leq k_2 \leq \ldots \leq k_m \leq m'$ such that $\forall e_j \in i$, $e'_{k_j \lambda}(e_j) = e_j$.

Why “pseudo" ?

- itemset i can be smaller than i'
- if they have the same size, we can use the term *instance*

Example

- $\{(a, b, c)\}$ is a pseudo-instance of $\{(a, *, c), (*, b, b)\}$
- $\{(a, b, c), (*, b, b)\}$ is an instance of $\{(a, *, c), (*, b, b)\}$ since they have the same size.

Pseudo-instance of a sequence

Let $s = \langle i_1, i_2, \ldots i_m \rangle$ and $s' = \langle i'_1, i'_2, \ldots i'_{m'} \rangle$ be two sequences such that $m \leq m'$, s is said to be a pseudo-instance of s' if there exist some integers $1 \leq k_1 \leq k_2 \leq \ldots \leq k_m \leq m'$ such that $\forall i_j \in s$, i_j is a pseudo instance of i'_{k_j}.

Example

$\langle \{(a,b,c)(a,*,d)\}\{(a,b,*)(b,*,*)\} \rangle$ is a pseudo-instance of $\langle \{(a,b,c)(a,*,d),(d,e,f)\}\{(a,*,*)(b,*,*)\} \rangle$

Instantiation

Let s' and s be two sequences such that s is a pseudo-instance of s', the function $\iota(s', s)$ is the substitution of at least one item e'_i in s' for an instance of e'_i in s.

$$\iota : \textit{sequence} \times \textit{sequence} \rightarrow \textit{set of sequences}$$
$$\iota(s', s) \mapsto \{s'' \text{ such that the following conditions hold}\}$$

- s'' is an instance of s'.
- $\exists$ *items* $e''_i \in s'', e'_i \in s'$ *and* $e_i \in s$ *such that* e_i is an instance of e'_i *and* $e''_i = [e_i/e'_i]$ where $[e_i/e'_i]$ is the substitution of e'_i for e_i.

Example

$\iota(\langle\{(a, *, c), (e, f, *)\}\{(*, c, d)\}\rangle, \langle\{(a, b, c)\}\rangle)$ is equal to the sequence $\langle\{(a, b, c), (e, f, *)\}\{(*, c, d)\}\rangle$

Difference

Let $s=\langle i_1, i_2, \ldots, i_l \rangle$ and $s' = \langle i'_1, i'_2, \ldots, i'_{l'} \rangle$ be two sequences, s and s' are said to be different ($s \neq s'$) if $s \not\subseteq s'$ and $s' \not\subseteq s$.

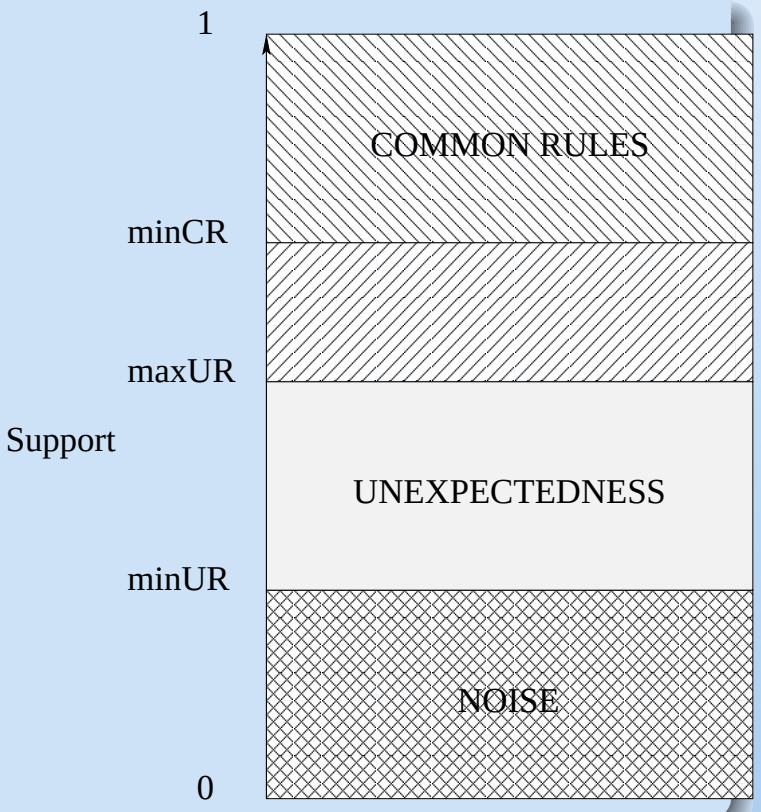
Example

- $s_1 = \langle \{(a_1, b_1)\}\{(a_2, *)\} \rangle$
- $s_2 = \langle \{(a_1, b_1)\} \rangle$
- $s_3 = \langle \{(a_2, b_2)\}\{(a_1, b_2)(a_2, b_1)\} \rangle$
- s_1 and s_2 are not different since $s_2 \subset s_1$.
- s_1 and s_3 are different since there is no relation between s_1 and s_3.

Thresholds

We consider the following user-defined thresholds:

- *minCR*
- *maxUR*
- *minUR*
- *minConf*



Common Rule *CR*

$CR : s_\alpha \rightarrow s_\beta$ s.t.:

- $supp(CR) > minCR$
- $conf(RC) > minConf$

Instance of *CR*

s' *s.t.* $supp(s') > minCR$ *and* s' *is an instance of* s_α

Unexpected Rule *UR*

$UR : \iota(s_\alpha, s') \rightarrow s_c$ *s.t.*

- $minUR \leq supp(UR) \leq maxUR$
- $s_c \neq s_\beta$

Verification

$VR : s_\alpha \rightarrow s_c$ *s.t.*

- $conf(VR) < minConf$ *and/or* $supp(RV) < maxUR$

Algorithm

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Algorithm: 3 steps

1. Multidimensional sequences are mined and stored in a prefix tree.
2. Common rules are discovered in the prefix tree.
3. Searching for unexpected rules.

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Agenda

1. Introduction
2. Related Work
3. Proposal
   - Specification of wild carded premisses
   - Difference in the conclusion
   - Unexpected Multidimensional Sequential Rules
   - Algorithm
4. Experiments
5. Conclusion

- Real data from EDF (Electricité de France) marketing context,
- Data cube describing marketing activity on a very large EDF customer database (about 30 millions of residential customers).
- 7 dimensions.

#Offer	Offer	State_Of_Offer	Heating	Geography	Support_Of_Offer	Time
1	BIEN	Flux	Electrical	Bordeaux	Phoning	Jan 2003
2	BIEN	Flux	Gas	Montpellier	Mailing	Mar 2005
3	RENO	Stock	Electrical	Lyon	Phoning	Nov 2004
4	DCS	Flux	Gas	Lyon	Phoning	Feb 2003
5	RENO	Stock	Fuel	Paris	Mailing	Jan 2004
…	…	…	…	…	…	…

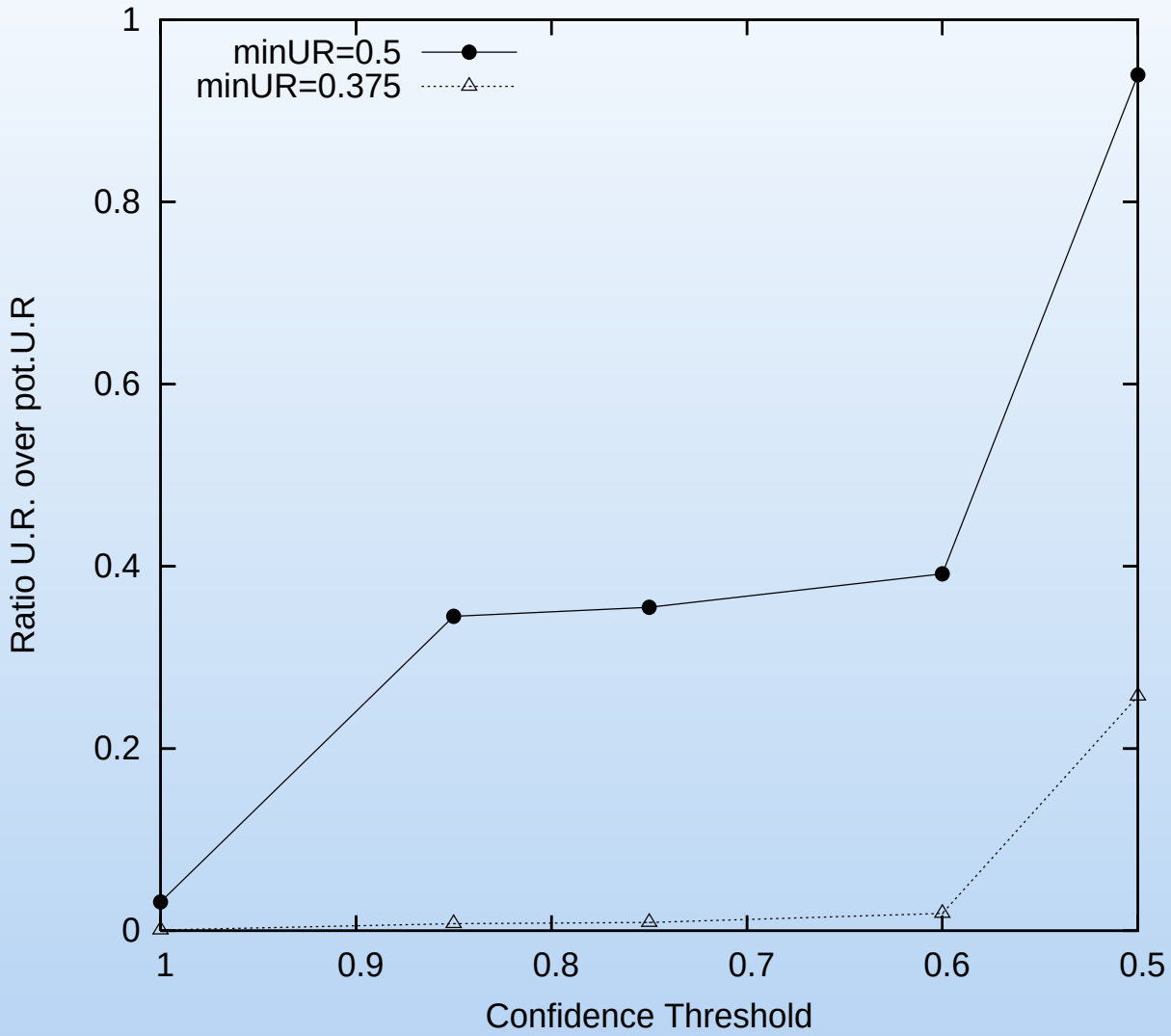


Figure: Percentage of Unexpected Rules over *minConf* ($maxUR = 0.8$)

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules
Algorithm
Experiments
Conclusion

Agenda

1. Introduction
2. Related Work
3. Proposal
   - Specification of wild carded premisses
   - Difference in the conclusion
   - Unexpected Multidimensional Sequential Rules
   - Algorithm
4. Experiments
5. Conclusion

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

Conclusion

Unexpected Multidimensional Sequential Rules

- Combining multidimensionality, time and unexpectedness.
- Experiments on real data cube

Future Avenues

Using the specificities of OLAP

- Hierarchies
- Aggregated values to compute the support

Reduce the space search

- Top *k* unexpected rules
- Condensed representation of rules

Generalization of Suzuki's Exceptions

- Time dimension to construct union.

Introduction
Related Work
Proposal
Premisses
Then-parts
Unexpected Multidimensional Sequential Rules Algorithm
Experiments
Conclusion

THANK YOU
Q&A