Exploratory Data Analysis in Dynamic Environments

Rudolf Kruse

Computational Intelligence Working Group

http://fuzzy.cs.uni-magdeburg.de

Aveiro, 26.06.2012
Research Topics: Information Mining with CI-Methods

Characteristics of the CI Research Group

- Books with Ph.D. students

- Software Tools NEFCLASS, Information Miner, ...

- Projects with Industrial Partners
Outline

Item Planning at Volkswagen
Change Mining at British Telecom
Temporal Pattern Mining at Daimler
Emerging Trends: Spatio–Temporal Data Analysis
Conclusions and Outlook
## Item Planning at VW

<table>
<thead>
<tr>
<th>Marketing strategy</th>
<th>prefer individual vehicle specifications by customers</th>
<th>bestseller-oriented vehicle specifications by car maker</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complexity</td>
<td>very large number of possible variants</td>
<td>low number of possible variants</td>
</tr>
</tbody>
</table>

### Vehicle specification

<table>
<thead>
<tr>
<th>Item</th>
<th>short back</th>
<th>2,8L 150 kW spark</th>
<th>Type alpha</th>
<th>4</th>
<th>leather, Type L3</th>
<th>yes</th>
<th>......</th>
</tr>
</thead>
<tbody>
<tr>
<td>Item family</td>
<td>body variant</td>
<td>engine</td>
<td>radio</td>
<td>door layouts</td>
<td>seat covering</td>
<td>vanity mirror</td>
<td>......</td>
</tr>
</tbody>
</table>

- Item: Planning at VW
- Complexity: Very large number of possible variants
- Item family: Body variant
Item Planning at VW

- approximately 200 item families (variables)
- from 2 to 100 items in each family
- i.e. more than \(2^{200}\) possible vehicle specifications
- choice of valid specifications is restricted by RULE SYSTEMS
  (10,000 technical rules, even more marketing- and production-oriented)

Example (technical rules that restrict validity of item combinations)

\[
\begin{align*}
\text{if } & \quad \text{engine} = e_1 \quad \text{then} \quad \text{transmission} = t_3 \\
\text{if } & \quad \text{engine} = e_4 \quad \text{and} \quad \text{auxiliary heater} = h_1 \\
\text{then } & \quad \text{generator} \in \{g_3, g_4, g_5\}
\end{align*}
\]

How to predict installation rates of item combinations?
Problem Presentation

**Historical Data**
Sample of produced *vehicle specifications* (representative choice, context-dependent, e.g. Golf)

(Golf, short back 2.8l 150kW, spark engine, radio alpha, ...)

**System of Rules**
*Rules* for the validity of item combinations (specified for a vehicle class and a planning interval)

If engine=e1 and auxiliary heater=h2 then generator in {g3,g4,g5}

**Prediction & Planning**
Predicted / assigned *planning data* (production program, demands, installation rates, capacity restrictions, ...)

Scientific Topics

Handling rules: **Modelling Constraints**
Handling historical data: **Machine Learning**
Combining the different sources: **Fusion of Models**
Supporting planners: **Belief Change**

200-dimensional space

Our recommendation: **Decision Support System** for the planners that is based on a **Decomposable Model**.
Decomposable Models

- Second Edition in 2009
- Graphical Model = Decomposition + Local Model + Operations
- Decomposition: (Un)directed Graphs, Hypergraph, Clique Trees
- Local Models: Relational Constraints, Possibilistic Constraints, Probabilistic Constraints
- Belief Change Operations
High Dimensional Probability Space used for VW Jetta

186 variables
174 cliques
max. 9 dimensions
Belief Change Operations (Gärdenfors)

**Focussing** : Calculation of part demands
Compute the installation rate of item combination

**Simulation**
Analyze customers’ preferences with respect to those persons who buy a *navigation system* in a *VW Polo*.

Calculate the conditional probability of the given item combination (via propagation algorithm).

**Updating** : Technical modification
Change of a rule from invalid to valid

**Revision** : Marketing stipulation, Logistic Restrictions
Changes of marginal/conditional probabilities
Planning Model based on Belief Change

context: vehicle class, planning interval

Historical data
context-dependent sample of produced vehicle specifications

System of rules
context-dependent rules for the validity of item combinations

Estimate prior distribution of installation rates

Quantitative Learning
PGM (Markov network) having the structure of the relational network

Modify representation
Transformation into a relational network with hypertree structure

Use cond. independencies (Composition)

Revision
Adaptation of installation rates of item combinations that change from valid to invalid

Fusion

Planning Model
fused consistent Markov network for item planning Conditioning

Updating
Find referential for item combinations that change from invalid to valid
Some Project Details

- Project Leader: Jörg Gebhardt
- Client-Server System
- Server on 6-8 Machines (16 GB each)
- 4-Processor AMD Opteron system
- Terabyte storage device
- Linux, JAVA, Oracle
- WebSphere Application Developer, Eclipse

- 5000 Networks are in daily use, worldwide
Current Scientific Issues

Improvements
Learning Methods
Rule Inconsistency Management

New Topics (proposed by the planners)
- How do the networks evolve over time?
  Change Mining
- How to improve man-machine-interaction?
  Visual Support
Outline

- Item Planning at Volkswagen
- Change Mining at British Telecom
- Temporal Pattern Mining at Daimler
- Emerging Trends: Spatio–Temporal Data Analysis
- Conclusions and Outlook
Mining Changing Customer Segments in Dynamic Markets

- **How do patterns evolve over time?**
  - world changes, so does the customer, so does the data, so do the patterns contained in it
  - changes mean risk (e.g. losing customers) or chance (e.g. satisfying upcoming needs first)
  - patterns which never change are in general well known

- **Change Mining** is a data mining paradigm for the study of *time-associated data*. Its objective is the discovery, modelling, monitoring, prediction and interpretation of changes in the models that describe an evolving population.
Rule Change Mining

- **Formulate/Induce a Rule of Interest**
  If \( \text{product} = \text{broadband} \) and \( \text{customer} = \text{male} \) then \( \text{InternetUsage} = \text{low} \)

- **Partition Data in time intervals, Calculate Rule Measures**
  \[
  \text{confidence}(\text{if } X \text{ then } y) = P(y|X), \quad \text{relsupport}(\text{if } X \text{ then } y) = P(X,y), \quad \ldots
  \]

- **Analyze Time Series of Rule Measures for a given Rule**

- **Generate “Knowledge”**
  Confidence of this rule increased towards winter, and decrease after a recall was effective
Rule Change Mining

Histories
- $H_{\text{supp}}(r) : \supp(r, T_1) \supp(r, T_2) \supp(r, T_3) \ldots \supp(r, T_n)$
- $H_{\text{conf}}(r) : \conf(r, T_1) \conf(r, T_2) \conf(r, T_3) \ldots \conf(r, T_n)$

Rules
- \[ \uparrow \]
- \[ \uparrow \]
- \[ \uparrow \]

Time Periods
- $T_1$
- $T_2$
- $T_3$
- $\ldots$
- $T_n$

**Rule Change**: Change of a rule’s statistical properties – not its symbolic representation

**Rule Change Mining**
- Discovery of interesting regularities (e.g. trends) within rule histories
- Rule interestingness is determined by features of its histories
Why are Rule Histories Important?

- Prevalent patterns are rarely of interest
- A user is interested in information which is novel or changes
- Change has an intrinsic interestingness
- Not the actual value of a traditional interestingness measure is of interest but how it changes.

Second Order Data Mining
Architecture for Rule Change Mining in Customer Segmentation

Association rules have to be discovered, stored and managed.

Change patterns have to be reliably detected.

Histories with change patterns have to be analysed for redundancies and evaluated.

Structural Analyser
- Association Rule Mining System
- Rule History Database

Change Analyser
- Noise Filter
  - Double exponential smoothing
- Trend Detection
  - Mann-Kendall Test
  - Cox-Stuart Test
- Stability Detection
  - $\chi^2$ Test

Interestingness Evaluator
- Temporal Redundancy Filter
- Interestingness Measures
- Interactive Filter Refinements
- Visual Exploration
- Matching user specified linguistic concepts
Outline

Item Planning at Volkswagen
Change Mining at British Telecom
Temporal Pattern Mining at Daimler
  - Visual Exploration
  - Mining Temporal Relations
Emerging Trends: Spatio–Temporal Data Analysis
Conclusions and Outlook
Quality Management

Which attribute values, have what kind of impact on the failure?

Complete network after dependency analysis at Daimler research plant.
Exploration: Interactive Mining Tool

Ratio between (marginal) Aircondition sale rate and sale rate given the Country

Airconditions of type 1 fail much more often in Egypt and Oman

Marginal Country distribution

Relative frequencies of Engine given Country
Visual Exploration: Local Pattern Discovery

Subnet of the entire network

- Partitioning of entire vehicle set according to network's potential tables
- Induction of rules
- Encoding of rules and their measures (graphically)
Patterns (describing failed vehicles) do not arise out of a sudden.
Rather: Evolvement as time progresses
Management decisions: Again, it takes time to see an effect.
Therefore: Consider the temporal changes of pattern properties.
Example: Before Template Matching

All rule trajectories over time.
Example: After Template Matching

Return rule trajectories over time meeting the linguistic concept “approximately unchanged lift and slightly increasing support”
Outline

Item Planning at Volkswagen
Change Mining at British Telecom
Temporal Pattern Mining at Daimler
  ▪ Visual Exploration
  ▪ Mining Temporal Relations
Emerging Trends: Spatio–Temporal Data Analysis
Conclusions and Outlook
Quality Management

- Until now: Classical if–then rule whose statistical properties were tracked over time.

- Now: Represent temporal relations directly with the rule itself

Aux. Heating, Automatic, ...

Rep. 1 on 2009/4/1, Rep. 2 on 2009/11/11, ...

Aux. Heating

Automatic

Rep. 1  Rep. 2  Rep. 3

time
Quality Monitoring

- **Aux. Heating**
  - Automatic
  - Rep. 1
  - Rep. 2
  - Rep. 3

- **Estate Car**
  - Automatic
  - Rep. 3

- **Diesel**
  - Automatic
  - Rep. 1
  - Rep. 3
Quality Monitoring

- **Aux. Heating**
  - Rep. 1
  - Rep. 2
  - Rep. 3

- **Estate Car**
  - Rep. 3

- **Diesel**
  - Rep. 1
  - Rep. 3

- **Automatic**
  - Rep. 1
  - Rep. 3

Confidence: 0.66
Mining Task

“Find all frequent temporal patterns for a given set of interval sequences and a minimal support threshold.”

Requirement:

Multiple occurrences of a temporal pattern in an interval sequence have to be considered.

Solution:

New definition of support (number of “minimal” occurrences)
What is a Temporal Relation?

**Relation a to b**
- **a before b**
- **a meets b**
- **a overlaps b**
- **a is-finished-by b**
- **a contains b**
- **a is-started-by b**
- **a equals b**

**Inverse Relation b to a**
- **b after a**
- **b is-met-by a**
- **b is-overlaped-by a**
- **b finishes a**
- **b during a**
- **b starts a**
- **b equals a**
What is a Temporal Pattern?

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>B</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>e</td>
<td>o</td>
<td>b</td>
</tr>
<tr>
<td>B</td>
<td>io</td>
<td>e</td>
<td>m</td>
</tr>
<tr>
<td>A</td>
<td>a</td>
<td>im</td>
<td>e</td>
</tr>
</tbody>
</table>
How to derive Frequent Temporal Rules from Frequent Temporal Patterns?

\[
\begin{array}{c|cc}
A & A & B \\
A & e & o \\
B & io & e
\end{array} 
\Rightarrow 
\begin{array}{c|ccc}
A & B & A \\
A & e & o & b \\
B & io & e & m \\
A & a & im & e
\end{array}
\]
Temporal Rules vs. Association Rules

**Temporal Rule:**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>e</td>
</tr>
<tr>
<td>B</td>
<td>io</td>
</tr>
</tbody>
</table>

Premise Pattern

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>e</td>
<td>o</td>
</tr>
<tr>
<td>B</td>
<td>io</td>
<td>e</td>
</tr>
</tbody>
</table>

Conclusion Pattern

| A, B, C | D |

**Association Rule:**

| A, B, C | D |

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>A</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>e</td>
<td>o</td>
</tr>
<tr>
<td>B</td>
<td>io</td>
<td>e</td>
</tr>
<tr>
<td>A</td>
<td>a</td>
<td>im</td>
</tr>
</tbody>
</table>

**Same transaction**
Assessing Rules (Contingency Table)

\[ X \implies Y \]

<table>
<thead>
<tr>
<th>Y</th>
<th>( \overline{Y} )</th>
<th>( \sum ) = N</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>(</td>
<td>XY</td>
</tr>
<tr>
<td>( \overline{X} )</td>
<td>(</td>
<td>\overline{XY}</td>
</tr>
</tbody>
</table>

We count multiple occurrences of a temporal pattern in an interval sequence. Hence, for Temporal rules Support(not \( X \)) = N – support (\( X \)) is not valid.

- Rule measures that do not work: Goodman-Kruskal, Odds ratio, Yule's Q, Yule's Y, Kappa, J-Measure, Gini-Index, Conviction, Collective strength, ...

- Rule measure that do work: Confidence, Average, Lift
Outline

Item Planning at Volkswagen
Change Mining at British Telecom
Temporal Pattern Mining at Daimler

Emerging Trend: Spatio–Temporal Data Analysis
- Online Marketing in Virtual Environments
- Estimation of automobile driver properties
- Analysis of aircraft movements on airport ground

Conclusions and Outlook
Online Marketing in Virtual Environments

- 3D Online Communities
- Virtual copies of real cities: Berlin, Frankfurt

**Goal**: What do have certain users in common?

**Challenge**: 450000 Avatar Profiles, Sensors track visiting users in their vicinity.

**Approach**: Analysis of spatio-temporal activities. Analysis of sequence of cooccurrence Graphs (by week, month, etc.). Are there subgraphs satisfying certain user-specified concepts, such as “subgraphs becoming less dense and more unbalanced”
Goal: Driver-specific reaction of assistant systems
Sporty drivers do not want to see/feel emergency break warning too early

Challenges:
- How to characterize drivers only based on car measurements?
- What effects have location and time?

Approach: Case Based Reasoning + Evolutionary Algorithm
Goal: Find groups of similar aircraft tracks (based on radar)
For shown airport, 1 million tracks in database
Each track: 1000–3000 points
Noise, e.g. observation gaps

Challenge:
- Similarity: based on interesting locations (blue crosses)
- high-dimensional dataset
Novel clustering algorithms needed

Approach: Fuzzy clustering with modifications
Future Trends in Exploratory Data Analysis

Main Challenge: Handling the vast amount of data
With mass storage devices, even high-resolution data in spatial and temporal dimensions can be easily stored
Problem: Usually, spatio-temporal relations are not analyzed
Needed:
  • Tools to explore data sets for better understanding data (especially over time!!)
  • Scalability of existing algorithms (most demand i.i.d. data)
How can background knowledge be fused into analysis?
Main focus in future: Real-world applications where interpretability, explanation, summarization and visualization are crucial for success