Extracting decision models from data and text

Jan Vanthienen, Vedavyas Etikala, Alexandre Goossens
LIRIS, Research Center for Management Informatics
KU Leuven, Belgium

RuleML Webinar, 2021
Overview

• Decision modeling
• Extracting decision models from data/cases
• Extracting decision models from text
• Challenges and future research
• Conclusion
Decision Modeling with DMN

Decision Model (DMN)

Decision requirements level

Offer

Stock data
Customer loyalty
Customer data

Customer loyalty knowledge

Offer knowledge

Customer Loyalty

<table>
<thead>
<tr>
<th>U</th>
<th>Longtime Customer</th>
<th>Yearly Sales</th>
<th>Customer Loyalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>TRUE</td>
<td>&gt;= 20k$</td>
<td>OK</td>
</tr>
<tr>
<td>2</td>
<td>FALSE</td>
<td>&lt; 20k$</td>
<td>NOT_OK</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td></td>
<td>NOT_OK</td>
</tr>
</tbody>
</table>

Decision logic level

Business process model (BPMN)

Collected application data
Decide on offer

Accept
Make offer
Decline
Refuse offer
Decision Requirements Diagram
# Decision Logic

## Primary Decision Logic

## Input Conditions

## Result

### Decision Table

<table>
<thead>
<tr>
<th>BMI Level</th>
<th>Sex</th>
<th>Waist(cm)</th>
<th>Risk Level</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Overweight&quot;</td>
<td>&quot;Male&quot;</td>
<td>&gt;102</td>
<td>&quot;Increased&quot;</td>
</tr>
<tr>
<td>&quot;Overweight&quot;</td>
<td>&quot;Male&quot;</td>
<td>&gt;102</td>
<td>&quot;High&quot;</td>
</tr>
<tr>
<td>&quot;Overweight&quot;</td>
<td>&quot;Female&quot;</td>
<td>&lt;=88</td>
<td>&quot;Increased&quot;</td>
</tr>
<tr>
<td>&quot;Overweight&quot;</td>
<td>&quot;Female&quot;</td>
<td>&gt;88</td>
<td>&quot;High&quot;</td>
</tr>
<tr>
<td>&quot;Obese I&quot;</td>
<td>&quot;Male&quot;</td>
<td>&lt;= 102</td>
<td>&quot;High&quot;</td>
</tr>
<tr>
<td>&quot;Obese I&quot;</td>
<td>&quot;Male&quot;</td>
<td>&gt;102</td>
<td>&quot;Very High&quot;</td>
</tr>
<tr>
<td>&quot;Obese I&quot;</td>
<td>&quot;Female&quot;</td>
<td>&lt;=88</td>
<td>&quot;High&quot;</td>
</tr>
<tr>
<td>&quot;Obese II&quot;</td>
<td>&quot;Female&quot;</td>
<td>&gt;88</td>
<td>&quot;Very High&quot;</td>
</tr>
</tbody>
</table>

Each row is a decision rule.

---

5
Decision rules and table extraction

Existing approaches:

• Decision modeling methodology

• Extracting rules (and tables) from text

• Mining rules and tables from data (accuracy vs comprehensibility)

• Extracting rules from code
DRD (Decision requirements) extraction

- Decision modeling methodology
- Extracting dependencies from text\(^1\)
- Mining decisions and tables from data (process mining + data mining)
- Extracting DRD from process models

Full decision model extraction

- Decision modeling methodology

- Extracting dependencies + rules from text

- Mining decisions models from data¹
  (process mining + data mining)

- Extracting DMN from process models

Extracting decision models from data/cases

From data
- From case data to decision trees, rules or networks (analytics, rule learning)
- From case data to analytics models (ANN) and then decision table models (Baesens, Vanthienen et al., 2003)
- From case data to DMN DRD

From event logs and data
- From process event logs to process models (process mining)
- From process event logs + case data to process models + predictive models (decision mining) (e.g. Rozinat & van der Aalst, 2006)
- From process event logs + case data to integrated process & decision models (integrated mining) (e.g. De Smedt, Hasic, vanden Broucke & Vanthienen (2017).
Extracting decision models from text

An analogy: process model generation from natural language text (Friedrich et al., 2011)

- Deriving rules from text
- Deriving tables from text
- Deriving DRDs and decision logic from text
Three stages to extract decision models

1. Which part of the text: classification
2. Extracting dependencies
3. Extracting decision logic
Stage 1: Text Classification

Text sample: Medical Guidelines for Obesity

• “The health risk level of a patient should be assessed from the obesity level, waist circumference and the sex of the patient. Furthermore, the degree of obesity should be determined from the BMI value and sex of the patient. Patient’s height and weight are considered to calculate his BMI value.”
• “When the patient’s sex is a male and his BMI value is in between 25 and 29.9, then his obesity level is normal.”
• “If patient’s sex is female and BMI value is above 25.0 and less than 30, then obesity level is overweight. Where as, If BMI value is 30.0 or higher, obesity level falls within the obese I range.”

*Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults
Techniques considered for Sentence Classification

• Deep Learning Classifiers
  o BERT for Sequence Classification
    (Bidirectional Encoder Representations from Transformers) classify sentences into irrelevant, decision logic and decision dependency.
  o Neural Network with GloVe as an Embedding Layer

• Non-Deep Learning Models
  o multinomial logistic regression, Naive Bayes and support vector machines

Alexandre Goossens, Charlotte Parthoens, Michelle Claessens and Jan Vanthienen, Deep learning for the extraction of decision modelling components, In preparation, 2021.
Results

- The training set consists of 400 sentences and the test set contains 149 sentences. Both sets have a balanced distribution of the classes.
- BERT is able to retrieve all sentences labeled as dependency (Recall = 1.00) and is good at identifying the sentences labeled as logic (Recall = 0.86).

<table>
<thead>
<tr>
<th>Table 2: Overview of results</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Deep learning models</strong></td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>GloVe+MLP</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>GloVe + CNN</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>BERT for sequence classification</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td><strong>Non-deep learning models</strong></td>
</tr>
<tr>
<td>Model</td>
</tr>
<tr>
<td>BoW + Logistic Regression</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>BoW + Naive Bayes</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>BoW + SVM</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>TF-IDF + Logistic Regression</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>TF-IDF + Naive Bayes</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>TF-IDF + SVM</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
Stage 2: Extracting dependencies

1. Pattern based approach with NLP
2. Deep learning approach
Stage 2-1: A pattern based approach

Fig: Three Stage Methodology of Tex2Dec

### Sentence patterns

<table>
<thead>
<tr>
<th>Dependency Pattern</th>
<th>Example</th>
<th>Base Concept (A)</th>
<th>Derived Concept (B)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dec. Active A =&gt; B</td>
<td>Patient's <em>height</em> determines his <em>BMI value</em>.</td>
<td>height</td>
<td><em>BMI value</em></td>
</tr>
<tr>
<td>Dec. Passive B &lt;= A</td>
<td>Patient's <em>BMI value</em> is determined from his <em>height</em>.</td>
<td>height</td>
<td><em>BMI value</em></td>
</tr>
<tr>
<td>Conditional A =&gt; B</td>
<td><em>Unless the season</em> is summer, <em>do not plan a barbeque.</em></td>
<td><em>season</em></td>
<td><em>plan a barbeque</em></td>
</tr>
<tr>
<td>Conditional B &lt;= A</td>
<td><em>A customer is loyal, if his annual sales are high.</em></td>
<td><em>annual sales</em></td>
<td><em>customer</em></td>
</tr>
</tbody>
</table>

Patterns considered to extract dependencies.
Stage 2: NLP Pipeline.

- Preprocessing
- Coreference Resolution
- Anaphora Resolution
- Concept Recognition
- Dependency Parsing
- Dependency Extraction

Selected Text → Semantic Analysis → Tuples

Dependencies

Syntactic Analysis
The health risk level of a patient should be assessed from the obesity level, waist circumference and the sex of the patient. Furthermore, the degree of obesity should be determined from the BMI value and sex of the patient. Patient's height and weight are considered to calculate his BMI value.

**Result:**

Health risk level of a patient should be assessed from obesity level, waist circumference and sex of patient. Degree of obesity should be determined from BMI value and sex of patient. Patient's height and weight are considered to calculate his BMI value.
Semantic Analysis

• Step B: Coreference resolution – fix cross referred concepts

• Result:
Semantic Analysis

• Step C: Anaphora resolution – Fix pronoun references and ownerships

• Result:

Patient’s risk level should be assessed from his obesity level, waist circumference and sex.

patient’s risk level should be assessed from patient’s obesity level, patient’s waist circumference and patient’s sex.
Syntactic Analysis

• Step D: Concept Recognition – Identifying nouns and noun phrases

noun phrase

health risk level should be assessed from obesity level, waist circumference and sex

noun phrase

noun phrase

• Result:

health risk level should be assessed from obesity level, waist circumference and sex
Syntactic Analysis

- **Step E: Dependency parsing – Identify the root action verb**

  - **Step F: Dependency extraction based on action verbs**
S1: The health risk level of a patient should be assessed from the obesity level, waist circumference and the sex of the patient.

S2: Furthermore, the degree of obesity should be determined from the BMI value and sex of the patient.

S3: Patient's height and weight are considered to calculate his BMI value.

S4: If the weight of the patient given in kgs and height of patient given in meters, then the BMI value is weight/(height*height).
Stage 2-2: a deep learning approach

• Using inside-outside-beginning (IOB) tagging format for base concepts, derived concepts and action verbs

• Two techniques were investigated to extract tags
  o BERT  
    (Bidirectional Encoder Representations from Transformers) classify sentences into irrelevant, decision logic and decision dependency.
  o Bi-LSTM-CRF  
    (Bi-directional-Long Short-Term Memory- Conditional Random Field)

Alexandre Goossens, Charlotte Parthoens, Michelle Claessens and Jan Vanthienen, Extracting decision dependencies and conditional clauses using deep learning, In preparation, 2021.
Results

- The training set consists of 195 explicit and 245 conditional dependency sentences, manually tagged.
- The test set contains 60 and 82 sentences.

<table>
<thead>
<tr>
<th></th>
<th>Explicit Dependency sentences</th>
<th>Conditional sentences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level</td>
<td>BERT base-uncased without stopword removal</td>
<td>BI-LSTM-CRF without preprocessing</td>
</tr>
<tr>
<td></td>
<td>B- DER</td>
<td>I- DER</td>
</tr>
<tr>
<td>Precision</td>
<td>0.79 ± 0.02</td>
<td>0.84 ± 0.05</td>
</tr>
<tr>
<td>Recall</td>
<td>0.87 ± 0.03</td>
<td>0.86 ± 0.03</td>
</tr>
<tr>
<td>F1-score</td>
<td>0.83 ± 0.02</td>
<td>0.85 ± 0.02</td>
</tr>
<tr>
<td></td>
<td>0.61 ± 0.06</td>
<td>0.78 ± 0.02</td>
</tr>
<tr>
<td></td>
<td>0.62 ± 0.03</td>
<td>0.63 ± 0.06</td>
</tr>
<tr>
<td></td>
<td>0.62 ± 0.02</td>
<td>0.70 ± 0.03</td>
</tr>
</tbody>
</table>

Fig. 3: Results for dependency tag extraction
Stage 3: extracting decision logic

• “When the patient’s sex is a male and his BMI value is in between 25 and 29.9, then his obesity level is normal.”

Extracted Rule: \[ \text{IF BMI value in [25, 29.9] AND sex = male THEN obesity level = normal} \]
## 3-1: Sentence patterns

<table>
<thead>
<tr>
<th>Sentence Pattern</th>
<th>Example</th>
<th>Condition</th>
<th>Consequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit IF - THEN</td>
<td><em>If patient’ BMI value is above 25.0 and less than 30, then obesity level is overweight</em></td>
<td>BMI value in [25.0, 30]</td>
<td>obesity level = overweight</td>
</tr>
<tr>
<td>Synonym IF- THEN</td>
<td><em>Unless the season is summer, do not plan a barbeque.</em></td>
<td>Season = summer</td>
<td>plan a barbeque = true</td>
</tr>
<tr>
<td>Implicit IF-THEN</td>
<td><em>Any customer with high annual sales is loyal.</em></td>
<td>annual sales = high</td>
<td>customer = loyal</td>
</tr>
</tbody>
</table>

Patterns considered to extract logical rules.
3-2: Deep learning approach

- The training set consists of 264 conditional sentences, manually tagged.
- The test set contains 82 sentences.
- Separate condition part and consequence part.

<table>
<thead>
<tr>
<th>Logic Extraction</th>
<th>BERT base-uncased without stopword removal</th>
<th>BI-LSTM-CRF without preprocessing</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conditional sentences</td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>B-CONS</td>
<td>0.88 ± 0.01</td>
<td>0.89 ± 0.02</td>
</tr>
<tr>
<td>I-CONS</td>
<td>0.91 ± 0.01</td>
<td>0.94 ± 0.02</td>
</tr>
<tr>
<td>B-COND</td>
<td>0.87 ± 0.00</td>
<td>0.94 ± 0.01</td>
</tr>
<tr>
<td>I-COND</td>
<td>0.93 ± 0.03</td>
<td>0.89 ± 0.01</td>
</tr>
<tr>
<td>B-ELSE</td>
<td>0.91 ± 0.00</td>
<td>0.87 ± 0.05</td>
</tr>
<tr>
<td>I-ELSE</td>
<td>0.95 ± 0.05</td>
<td>0.97 ± 0.02</td>
</tr>
<tr>
<td>B-EXCE</td>
<td>1.00 ± 0.00</td>
<td>0.80 ± 0.00</td>
</tr>
<tr>
<td>I-EXCE</td>
<td>1.00 ± 0.00</td>
<td>0.65 ± 0.14</td>
</tr>
<tr>
<td>AVG_MICRO</td>
<td>0.92 ± 0.02</td>
<td>0.92 ± 0.02</td>
</tr>
<tr>
<td>AVG_MACRO</td>
<td>0.93 ± 0.00</td>
<td>0.87 ± 0.00</td>
</tr>
</tbody>
</table>

Fig. 4: Results for logic tag extraction

Alexandre Goossens, Charlotte Parthoens, Michelle Claessens and Jan Vanthienen, Extracting decision dependencies and conditional clauses using deep learning, In preparation, 2021.
Early results

Eligibility for smallpox vaccine is depended on risk and outbreak.

If you are at risk or there is an outbreak then eligibility for smallpox vaccine is true.

Risk is determined from contact with a similar virus, labworker or smallpox virus exposure.

You are at risk if contact with a similar virus is true and you are a labworker.
You are also at risk if there is smallpox virus exposure.

LIST OF CONCEPTS =
outbreak
smallpox virus exposure
eligibility for smallpox vaccine
contact with a similar virus
risk
labworker
contact with a virus

Your set of dependencies can further help you to identify incorrect concepts.

LIST OF DEPENDENCIES =
('determined', 'labworker', 'risk')
('depended', 'risk', 'eligibility for smallpox vaccine')
('determined', 'contact with a similar virus', 'risk')
('determined', 'smallpox virus exposure', 'risk')
('depended', 'outbreak', 'eligibility for smallpox vaccine')
('None', 'contact with a virus', 'risk')
Challenges and Future research

• Linguistic Challenges from NLP
  • Ambiguities
  • Incompleteness

• Full automation of model extraction requires:
  • Understanding concepts and values
  • Order of the rules
  • Table hit policies, decompositions

• Huge potential for further study
  • Comparing pattern based approaches and deep learning
  • Digital Automation with DMN
Conclusion

• While data science and data analytics are doing just fine on their own, integrating it with DMN can not only add explainability but actually improve accuracy.

• Knowledge based systems rely mostly on textual guidelines, policies or regulations as their knowledge sources. Automatic extraction provides an insight about how a textual resource (e.g. a clinical guideline document) could be made interpretable not only to domain experts but also to computers systems. The same document/model to applications and users.

• Cuts down the modeling time notably.
Some references

Thank you 😊