Genotype Determination of Danish Jersey Cattle

Assumptions about parents:
- risk about misstatement

- genotype mother
- genotype father

- genotype child, 6 possible values

- 4 lysis values measured by photometer

- Reliability of databases
- Inheritance rules
- Blood group determination
Qualitative Knowledge

- Parental error
  - Dam correct
    - Phenogroup 1: Stated dam
    - Phenogroup 2: True dam
    - Phenogroup 1: Offspring
  - Sire correct
    - Phenogroup 1: Stated sire
    - Phenogroup 2: True sire
    - Phenogroup 2: Offspring

- Genotype offspring
  - Factor 40 (F1)
    - Lysis 40
  - Factor 41 (F2)
    - Lysis 41
  - Factor 42 (V1)
    - Lysis 42
  - Factor 43 (V2)
    - Lysis 43
Example: Genotype Determination of Jersey Cattle

variables: 22, state space $6 \cdot 10^{13}$, parameters: 324

Graphical Model

- node
  - random variable
- edges
  - conditional dependencies
- decomposition
  - $P(X_1, \ldots, X_{22}) = \prod_{i=1}^{22} P(X_i \mid \text{parents}(X_i))$
- diagnosis
  - $P(\cdot \mid \text{knowledge})$
Learning Graphical Models

data + prior information

Inducer

→

A

B

C

+ local models
Genotype Determination of Danish Jersey Cattle: Database of Cases

747 cases
22 entries per case

Case 657:

\[ (0, 6, 5, 5, f_1, v_2, f_1, v_1, f_1v_1, f_1 \text{ or } v_2, f_1, f_1, v_2, f_1, v_1, \ldots) \]

- Lysis factors from 0 to 7 (photometer)
- Observed phenogroups of parents
- Genotype offspring
- Phenogroup offspring 1
- Phenogroup offspring 2
- True phenogroups of parents

- ESPRIT Project DRUMS 2, BR 6156

Problems:
- How to reduce complexity problems?
- How to handle imprecise (fuzzy, vague, ...) data?
# The Learning Problem

<table>
<thead>
<tr>
<th>known structure</th>
<th>unknown structure</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
<td><img src="https://via.placeholder.com/150" alt="Diagram" /></td>
</tr>
</tbody>
</table>

**Complete data**

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_4$</td>
<td>$b_3$</td>
<td>$c_1$</td>
</tr>
<tr>
<td>$a_3$</td>
<td>$b_2$</td>
<td>$c_4$</td>
</tr>
</tbody>
</table>

**Statistical Parametric Estimation** (closed from eq.):
- statistical parameter fitting,
- ML Estimation,
- Bayesian Inference, ...

**Discrete Optimization over Structures** (discrete search):
- likelihood scores,
- MDL

**Problem:** search complexity ➔ heuristics

**Incomplete data** (missing values, hidden variables,...)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>$a_4$</td>
<td>?</td>
<td>$c_1$</td>
</tr>
<tr>
<td>$a_3$</td>
<td>$b_2$</td>
<td>?</td>
</tr>
</tbody>
</table>

**Parametric Optimization:**
- EM,
- gradient descent, ...

**Combined Methods:**
- structured EM
- only few approaches

**Problems:**
- criterion for fit?
- new variables?
- local maxima?
- fuzzy values?
Genotype Determination

Directed dependency network

Rule → conditional dependency

Hypergraph representation

Rule → constraint
Application of Induction of Graphical Models in the Automotive Industry

Daimler-Chrysler Research and Technology Ulm, „Data Mining“ Project

Fields of Application

- Improvement of Product Quality by Finding Weaknesses
  - Learn dependency network for vehicle properties and faults
  - Look for unusual conditional fault frequencies
  - Find causes for these unusual frequencies
  - Improve construction of vehicle

- Improvement of Error Diagnosis in Garages
  - Learn dependency network for vehicle properties and faults
  - Record properties of new faulty vehicle
  - Test for the most probable faults
Analysis of Daimler/Chrysler Database

- Database: ~ 18,500 passenger cars
  > 100 attributes per car

- Analysis of dependencies between special equipment and faults.

- Results used as a starting point for technical experts looking for causes.
General Structure of (most) Learning Algorithms for Graphical Models

- Use a criterion to measure the degree to which a network structure fits the data and the prior knowledge
  (model selection, goodness of hypergraph)

- Use a search algorithm to find a model that receives a high score by the criterion
  (optimal spanning tree, K2: greedy selection of parents, ...)
Measuring the Deviation from an Independent Distribution

Probability- and Information-based Measures

- information gain *
  identical with mutual information
- information gain ratio *
- g-function (Cooper and Herskovits)
- minimum description length
- gini index *

Possibilistic Measures

- expected nonspecificity
- specificity gain
- specificity gain ratio

(Measures marked with * originated from decision tree learning)
Data Mining Tool Clementine
Fictitious example:
There are significantly more **faulty batteries**, if both **air conditioning** and **electrical roof top** are built into the car.
### Example Subnet

Influence of special equipment on battery faults:

<table>
<thead>
<tr>
<th>(fictitious) frequency of battery faults</th>
<th>air conditioning</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>with</td>
</tr>
<tr>
<td>electrical sliding roof</td>
<td>with</td>
</tr>
<tr>
<td></td>
<td>without</td>
</tr>
</tbody>
</table>

- significant deviation from independent distribution
- hints to possible causes and improvements
- here: larger battery may be required, if an air conditioning system and an electrical sliding roof are built in

(The dependencies and frequencies of this example are fictitious, true numbers are confidential.)
Data Mining Tool “Information Miner”