Knowledge Acquisition through Machine Learning: Minimising Expert’s Effort

Ricardo Blanco-Vega
José Hernández-Orallo
María José Ramírez-Quintana
Agenda

- Introduction
- Trade-off Analysis
- Optimisation of the Size of the training set using a Modified MML
- Experimental Evaluation
- Application Procedure
- Conclusions y Future Work
Introduction

- One of the main problems in expert systems: knowledge acquisition bottleneck

- Many experts are not able to write down their knowledge:
  - Clear
  - Unambiguous rules.
Introduction

- Expert write down all their knowledge:
  - a high effort
  - can be very time-consuming
  - difficult to maintain and
  - sometimes the result isn’t a model fully automated.
Introduction

Minimising Expert’s Effort

- Training a model:
  - captures the expert’s knowledge
  - high accuracy
  - high comprehensibility
  - with a minimum number of queries.
Introduction

Applications:
- Diagnosis
- Estimation
- Detection,
- Selection, etc.

Cases are described by a fixed series of attributes and a dependent value.

- Expert’s model predicts the dependent value according to the rest of attributes.
- This model structure is similar to predictive models in machine learning
Introduction

Mimetic Technique

First Stage
- Training Data
- First Learning Algorithm
- First Classifier (Oracle)

Second Stage
- Unlabelled Random Data
- First Classifier (Oracle)
- Labelled Random Data
- Training Data + Labelled Random Data
- Second Learning Algorithm
- Second Classifier (Mimetic)
Introduction

Specialization Mimetic Technique

First Stage

Unlabelled Random Data

First Classifier (Oracle)

Training Data + Labelled Random Data

T + R

decision tree

Second Stage

Second Learning Algorithm

Second Classifier (Mimetic)

µ
Introduction

Mimetic process with expert oracle
Introduction

Expert

Set of examples

Labelled dataset

ML technique

Mimetic Model

invented
Introduction

- Method based on learning curves of mimetic method.
- To predict the number of cases.
- Trade-off between accuracy and comprehensibility of the models.
Trade-off Analysis

size of the invented data increases

accuracy increases

smaller invented datasets

fewer rules (greater comprehensibility)
Trade-off Analysis

The minimum message length (MML) principle

\[ \text{MsgLen}(H \cap D) = \text{MsgLen}(H) + \text{MsgLen}(D | H) \]
Optimisation of the Size

- This problem can be seen as an optimisation process.
- The objective is to maximise the accuracy and the comprehensibility of the model given some constraints.
- The constraints are obtained by the learning curves for the mimetic model.
Optimisation of the Size

Modified MML

Cost of mimetic model:

\[
\text{Cost}(\mathcal{M}) = \text{MsgLen}(\mathcal{M}) + \text{MsgLen}(D|M) + \text{Query}(D)
\]
Learning Curves for the Mimetic Models

Optimisation of the Size

$E = \alpha \ln(S) + \beta$

$E = -5.0688 \ln(S) + 50.587$

$R^2 = 0.997$

Error vs. size for the balance-scale dataset
Learning Curves for the Mimetic Models

Optimisation of the Size

Learning Curves for the Mimetic Models

\[ R = \delta S + \lambda \]

\[ R = 0.0576 S + 16.194 \]

\[ R^2 = 0.983 \]

Number of rules vs. size for the balance-scale dataset
Calculating

\[ \text{MsgLen}(M) \approx R^* cr \]
\[ \text{MsgLen}(D/M) \approx E^* ce \]
\[ \text{Query}(D) \approx |D|/cq \]
\[ \text{Cost(Model)} \approx R^* cr + E^* ce + |D|/cq \]
\[ \text{Cost(Model)} \approx R^* cr' + E^* ce \]

the second derivative

\[ S_{opt} = -K^* \alpha / \delta \]
\[ -K^* \alpha / S^2 \]

the \( S_{opt} \) value corresponds to a minimum
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>anneal</td>
<td>6</td>
<td>32</td>
<td>6</td>
<td>898</td>
</tr>
<tr>
<td>2</td>
<td>audiology</td>
<td>0</td>
<td>69</td>
<td>24</td>
<td>226</td>
</tr>
<tr>
<td>3</td>
<td>balance-scale</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>625</td>
</tr>
<tr>
<td>4</td>
<td>breast-cancer</td>
<td>0</td>
<td>9</td>
<td>2</td>
<td>286</td>
</tr>
<tr>
<td>5</td>
<td>cmc</td>
<td>2</td>
<td>7</td>
<td>3</td>
<td>1,473</td>
</tr>
<tr>
<td>6</td>
<td>colic</td>
<td>7</td>
<td>15</td>
<td>2</td>
<td>368</td>
</tr>
<tr>
<td>7</td>
<td>diabetes</td>
<td>8</td>
<td>0</td>
<td>2</td>
<td>768</td>
</tr>
<tr>
<td>8</td>
<td>hayes-roth</td>
<td>0</td>
<td>4</td>
<td>3</td>
<td>132</td>
</tr>
<tr>
<td>9</td>
<td>hepatitis</td>
<td>6</td>
<td>13</td>
<td>2</td>
<td>155</td>
</tr>
<tr>
<td>10</td>
<td>iris</td>
<td>4</td>
<td>0</td>
<td>3</td>
<td>150</td>
</tr>
<tr>
<td>11</td>
<td>monks1</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>556</td>
</tr>
<tr>
<td>12</td>
<td>monks2</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>601</td>
</tr>
<tr>
<td>13</td>
<td>monks3</td>
<td>0</td>
<td>6</td>
<td>2</td>
<td>554</td>
</tr>
<tr>
<td>14</td>
<td>sick</td>
<td>7</td>
<td>22</td>
<td>2</td>
<td>3,772</td>
</tr>
<tr>
<td>15</td>
<td>vote</td>
<td>0</td>
<td>16</td>
<td>2</td>
<td>435</td>
</tr>
<tr>
<td>16</td>
<td>vowel</td>
<td>10</td>
<td>3</td>
<td>11</td>
<td>990</td>
</tr>
<tr>
<td>17</td>
<td>waveform-5000</td>
<td>40</td>
<td>0</td>
<td>3</td>
<td>5,000</td>
</tr>
<tr>
<td>18</td>
<td>zoo</td>
<td>1</td>
<td>16</td>
<td>7</td>
<td>101</td>
</tr>
</tbody>
</table>
# Experimental Evaluation

## Parameters and determination coefficients ($R^2$) for the learning curves with three points ($n=3$)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Error vs Size</th>
<th>Rules vs Size</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$\alpha$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>1</td>
<td>-4.5711</td>
<td>0.97</td>
</tr>
<tr>
<td>2</td>
<td>-7.9902</td>
<td>0.99</td>
</tr>
<tr>
<td>3</td>
<td>-5.4376</td>
<td>1.00</td>
</tr>
<tr>
<td>4</td>
<td>-0.4968</td>
<td>0.74</td>
</tr>
<tr>
<td>5</td>
<td>-1.5936</td>
<td>0.93</td>
</tr>
<tr>
<td>6</td>
<td>-2.66</td>
<td>0.95</td>
</tr>
<tr>
<td>7</td>
<td>-1.2611</td>
<td>0.98</td>
</tr>
<tr>
<td>8</td>
<td>-7.197</td>
<td>0.96</td>
</tr>
<tr>
<td>9</td>
<td>-3.628</td>
<td>0.98</td>
</tr>
<tr>
<td>10</td>
<td>-8.9288</td>
<td>0.97</td>
</tr>
<tr>
<td>11</td>
<td>-7.6452</td>
<td>0.95</td>
</tr>
<tr>
<td>12</td>
<td>-7.5818</td>
<td>0.95</td>
</tr>
<tr>
<td>13</td>
<td>-4.1454</td>
<td>0.94</td>
</tr>
<tr>
<td>14</td>
<td>-0.2817</td>
<td>1.00</td>
</tr>
<tr>
<td>15</td>
<td>-1.5628</td>
<td>0.99</td>
</tr>
<tr>
<td>16</td>
<td>-4.9703</td>
<td>1.00</td>
</tr>
<tr>
<td>17</td>
<td>-3.1991</td>
<td>0.99</td>
</tr>
<tr>
<td>18</td>
<td>-10.865</td>
<td>0.99</td>
</tr>
<tr>
<td>Avg</td>
<td></td>
<td>0.97</td>
</tr>
</tbody>
</table>
Application Procedure

```c
size_set = {10, 20}; // initial number of examples
margin = 0.1; // percentage of error wrt. the optimum size.
i = 20;
while(true) {
    Ask_Expert_Until(i);
    opt = Estimate_Opt_Value(size_set);
    if ((opt < i) || (i/opt > 1 - margin))
        break;
    else {
        i = opt;
        size_set = size_set ∪ { i };
    }
}
```
## Application Procedure

### An example of the trace

<table>
<thead>
<tr>
<th>Iteration</th>
<th>$i$</th>
<th>$opt$</th>
<th>$i/opt$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>20</td>
<td>207</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>207</td>
<td>340</td>
<td>0.6</td>
</tr>
<tr>
<td>3</td>
<td>340</td>
<td>353</td>
<td>0.97 (STOP)</td>
</tr>
</tbody>
</table>
Conclusions

- We have analysed a scenario where knowledge acquisition is made through simple queries to one or more experts.
- Our approach is:
  - practical,
  - easy-to-implement and
  - general (in many situations).
Conclusions

We need:

- The expert,
- some unlabelled data and
- any machine learning technique.

We propose a methodology to estimate the number of cases needed to obtain the “optimal” model.
Conclusions

- A step forward in making knowledge acquisition through machine learning much more practical and easy.
- Which can help to solve the knowledge acquisition bottleneck.
Future work

- Grouping similar cases by clustering techniques and then ask the expert to label the clusters.

- Applied for other machine learning methods and other machine learning tasks.