An Extended Neural Gas Model for Efficient Data Mining Tasks

Jean-Charles Lamirel*, Shadi Al Shehabi**

CORTEX-INRIA research project
LORIA, Campus Scientifique, BP 239
54506 Vandoeuvre-lès-Nancy Cedex, France
*Jean-Charles.Lamirel@loria.fr, **shadialshehabi@gmail.com

Abstract
This paper presents a numerical association rule extraction method that is based on original quality measures which evaluate to what extent a numerical classification model behaves as a natural symbolic classifier such as a Galois lattice. The proposed method copes with the usual problems of the symbolic association rule extraction method that are computation time and rule selection.

Introduction
Symbolic association rule extraction models [1] suffer of very serious limitations. Rule generation is a highly time-consuming process that generates a huge number of rules, including a large ratio of redundant rules. Hence, this prohibits any kind of rule computation and selection as soon as data are numerous and they are represented by very high-dimensional description space. This latter situation is very often encountered with documentary data. In this paper we propose a new approach for knowledge extraction that consists in using a MultiGAS model as a front-end for unsupervised extraction of association rules. In our approach we exploit both the generalization and the intercommunication mechanisms of the model. We also make use of our original recall and precision measures that derive from the Galois lattice theory and from Information Retrieval (IR) domains.

Basic principles
The MultiGAS model is a neural network model that represents a viewpoint-oriented extension of the Neural Gas model. Its main principle is to be constituted by several gases that have been generated from the same data. Each gas is itself issued from a specific data description subspace (i.e. viewpoint). The relation between gases is established through the use of two main mechanisms: the inter-gas communication mechanism and the generalization mechanism. A detailed description of the model is given in [2].

The classical evaluation measures for the quality of classification are based on the intra-class inertia and the inter-class inertia (see [3]). These measures are often strongly biased because they depend both on the pre-processing and on the classification methods. Therefore, we have proposed to derive from the Galois lattice and Information Retrieval (IR) domains two new quality measures, Recall and Precision. The Precision and Recall measures are based on the properties of class members [3]. The Precision criterion measures in which proportion the content of the classes generated by a classification method is homogeneous. The greater the Precision, the nearer the intensions of the data belonging to the same classes will be one with respect to the other, and consequently, the more homogeneous will be the classes. In a complementary way, the Recall criterion measures the exhaustiveness of the content of said classes, evaluating to what extent single properties are associated with single classes. The Recall criterion should be considered as a specific application of the statistical concept of sensitivity (i.e. true positive rate) to class properties. The Recall (Rec) and Precision (Prec) measures for a given property \( p \) of the class \( c \) are expressed as:

\[
\text{Rec} \ (p) = \frac{\sum\xi_d^c \xi_d^p}{\sum\xi_d^c}, \ \text{Prec} \ (p) = \frac{\sum\xi_d^c \xi_d^p}{\sum\xi_d^c}
\]

such that, \( C \) is a set of classes issued from a classification method applied on a set of documents \( D \), \( c \in C \), and

\[ c = \{ d \in c, \ \xi_d^p > 0 \} \]

where \( \xi_d^p \) is the weight of the property \( p \) for the data \( d \).

We will rely on the class quality criteria for extracting rules from the set of numerical classes (i.e. clusters) generated through the MultiGAS model. For a given class \( c \), the general form of the extraction algorithm follows:

\[
\forall p_1, p_2 \in P^c, \ \text{If} \ (\text{Rec}(p_1) = \text{Rec}(p_2) = \text{Prec}(p_1) = \text{Prec}(p_2) = 1) \ \text{Then} \ p_1 \leftrightarrow p_2 \ (\text{equivalence rule})
\]

\[
\text{ElseIf} \ (\text{Rec}(p_1) = \text{Rec}(p_2) = \text{Prec}(p_2) = 1) \ \text{Then} \ p_1 \rightarrow p_2
\]

\[
\text{ElseIf} \ (\text{Rec}(p_1) = \text{Rec}(p_2) = 1) \ \text{Then} \]

Copyright © 2007, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.
If (Extent(p₁) ⊆ Extent(p₂)) Then: p₁ → p₂
If (Extent(p₂) ⊆ Extent(p₁)) Then: p₂ → p₁
If (Extent(p₁) ≡ Extent(p₂)) Then: p₁ ↔ p₂

∀ p₁ ∈ Pₖ₊₁, p₂ ∈ Pₖ − Pₖ₊₁
4) If (Rec(p₁) = 1) If (Extent(p₁) ⊆ Extent(p₂)) Then: p₁ → p₂(*)

where Prec and Rec respectively represent the local Precision and Recall measures, Extent(p) represents the extension of the property p (i.e. the list of data to which the property p is associated), and Pₖ represent the set of peculiar properties of the class c.

Experimental data

Our test database is a database of 1000 patents [4]. For the viewpoint-oriented approach the structure of the patents has been parsed in order to extract four different subfields corresponding to four different viewpoints: Use, Advantages, Patentees and Titles. As it is full text, the content of the textual fields of the patents associated with the different viewpoints is parsed by a lexicographic analyzer in order to extract viewpoint specific indexes. The obtained indexes are then normalized by an expert of the patent domain.

Results

A global summary of the results is given in table 1. Said table includes a comparison of our extraction algorithm with a standard symbolic rule extraction method as regards to the amount of extracted rules. In single viewpoint experiment, when our extraction algorithm is used with its optional step, it is able to extract a significant ratio of the rules that can be extracted by a classical symbolic model basically using a combinatorial approach. In some case, such as the Patentees viewpoint, all the rules of 100% confidence can be extracted from a single level of the gas. Alternatively, as in the case of the Use viewpoint, the combination of gas levels of the same viewpoint can be used for extracting all the rules of 100% confidence (see table 1). The worse extraction performance is obtained with the Titles viewpoint. This relatively low performance (58% of rules of 100% confidence extracted using all the gas levels) can be explained both by the higher sparseness and by the higher sparseness of the data related to this viewpoint. Nevertheless, it is compensated by the much better extraction efficiency, as compared to the symbolic model. Moreover, in the case of this viewpoint, the extracted rules have an average support which is higher than the average support of the overall rule set (see again table 1).

Table 1: Summary of results.

<table>
<thead>
<tr>
<th></th>
<th>Patentes</th>
<th>Titles</th>
<th>Use</th>
<th>Advantages</th>
<th>Use ∩ Advantages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total rule count</strong></td>
<td>12</td>
<td>2526</td>
<td>536</td>
<td>404</td>
<td>649</td>
</tr>
<tr>
<td><strong>Average confidence</strong></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Average support</strong></td>
<td>3.583</td>
<td>1.049</td>
<td>1.139</td>
<td>1.042</td>
<td>1.081</td>
</tr>
<tr>
<td><strong>Global rule count</strong></td>
<td>26</td>
<td>4912</td>
<td>2238</td>
<td>1436</td>
<td>2822</td>
</tr>
<tr>
<td><strong>Average confidence</strong></td>
<td>53%</td>
<td>59%</td>
<td>59%</td>
<td>44%</td>
<td>45%</td>
</tr>
<tr>
<td><strong>Peculiar rule count</strong></td>
<td>12</td>
<td>422</td>
<td>251</td>
<td>287</td>
<td>250</td>
</tr>
<tr>
<td><strong>Average confidence</strong></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>Extended rule count</strong></td>
<td>12</td>
<td>1338</td>
<td>536</td>
<td>319</td>
<td>642</td>
</tr>
<tr>
<td><strong>Average confidence</strong></td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
<td>100%</td>
</tr>
<tr>
<td><strong>% of symbolic total</strong></td>
<td>100%</td>
<td>58%</td>
<td>100%</td>
<td>79%</td>
<td>99%</td>
</tr>
<tr>
<td><strong>Average support</strong></td>
<td>3.583</td>
<td>1.081</td>
<td>1.139</td>
<td>1.050</td>
<td>1.073</td>
</tr>
</tbody>
</table>

References


