Hierarchical Classification of Data Streams

Antonio Rafael Sabino PARMENZAN, Gustavo E. A. P. A. BATISTA
Laboratory of Computational Intelligence – LABIC
Institute of Mathematics and Computer Science – ICIM, University of São Paulo – USP, São Carlos - SP

Introduction

Hierarchical Classification: Data Representation and Modelling Approaches

<table>
<thead>
<tr>
<th>Attributes</th>
<th>Examples</th>
<th>Flat Classification Approach</th>
<th>Local Classifier Approaches</th>
<th>Global Classifier Approach</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$X_1$, $X_2$, $X_3$, ..., $X_M$, Class (Y)</td>
<td>$R \rightarrow R.1 \rightarrow R.1.1$</td>
<td>$R \rightarrow R.1 \rightarrow R.1.2$</td>
<td>$R \rightarrow R.2 \rightarrow R.2.2$</td>
</tr>
<tr>
<td>$E_1$</td>
<td>$X_{11}$, $X_{12}$, $X_{13}$, ..., $X_{1M}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_2$</td>
<td>$X_{21}$, $X_{22}$, $X_{23}$, ..., $X_{2M}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_3$</td>
<td>$X_{31}$, $X_{32}$, $X_{33}$, ..., $X_{3M}$</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$E_N$</td>
<td>$X_{N1}$, $X_{N2}$, $X_{N3}$, ..., $X_{NM}$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

(a) Tree
(b) Direct Acyclic Graph

Our Goal

Our aim is to investigate, propose and evaluate incremental Machine Learning methods for the task of Hierarchical Classification of Nonstationary Data Streams and with restrictions concerning the availability of labels. Besides seeking to advance the topics of Hierarchical Classification and Data Stream Mining, this study also intends to collaborate with a relevant and promising application for automatic insect identification.

Data Stream Mining

1. Input
   - Requirement 1: Process an example at a time, and inspect it only once (at most)
   - Requirement 2: Use a limited amount of memory
   - Requirement 3: Work in a limited amount of time
   - Requirement 4: Be ready to predict at any time

2. Learning
   - Training Examples
     - Test Examples
     - Predictions

3. Model
   - Requirement 4

Proposed Method

- Acquisition of initial data
- Induction of the hierarchical classifier
- Classification
- Verification and inclusion of the example in the training dataset
- Classifier update

Initial Data Collection

Preliminary Results and Future Work

1. The high effort required to include the class hierarchy into the process of labeling Data Streams is compensated by the flexibility and robustness provided to the classification process
2. We intend to compare the proposed methods with state-of-the-art classifiers in Data Streams, which disregard the eventual hierarchical organization of the classes
3. We would also like to address the problems of imbalanced classes and open-set classification