# SWISSANALYST

*Data Mining Without the Entry Ticket*

O. Povel[1] and C. Giraud-Carrier[2]
[1] *Portia SA, Lausanne, Switzerland;* [2] *ELCA Informatique SA, Lausanne, Switzerland*

Abstract: This paper introduces SwissAnalyst, a complete Data Mining environment powered by Weka. SwissAnalyst was developed as an intuitive process layer, which offers all necessary features to place it on par, in terms of functionality, with most major commercial Data Mining software packages. As GNU GPL software, SwissAnalyst offers a license-free platform to develop both proofs of concept for potential business users and complex process-driven solutions for researchers.

Key words: Data Mining, KDD Process, GNU software, Graphical User Interface

## 1. INTRODUCTION

Data Mining techniques, especially machine learning, have only recently come out of the research labs and the past few years have seen a proliferation of commercial Data Mining (DM) software packages. These packages, for the most part, essentially wrap a varying number of public domain (i.e., freely available) components or algorithms (sometimes re-implemented with rather small proprietary extensions) in an user-friendly graphical interface.

Although such tools seemingly make DM technology more readily available to non-expert end-users, they tend either to provide only limited functionality (e.g., few pre-processing facilities, only decision tree induction, etc.) or to come with a non-negligible price-tag, which is generally perceived as an expensive entry ticket for those "would-be adopters" who might just wish to first "peek through the fence" before committing more seriously to the technology.

The challenge with having high DM functionality at low cost, however, is in the expertise required to use and orchestrate the heterogeneous set of raw, public domain versions of algorithms. One of the most significant attempts at meeting this challenge is the recent development (and continuous improvement) of the Weka tool (Witten and Frank, 2000). Under its GNU General Public License (GPL), Weka makes freely available a large collection – probably the richest of its kind – of DM algorithms (pre-processing, classification, association, clustering and visualization). The success of Weka is evident in a recent survey, which suggested that it was the second most regularly used DM tool, after SPSS Clementine and before SAS Enterprise Miner (KDnuggets Poll, 2002a).

Satisfied with this popularity, we wanted to leverage Weka's rich functionality while extending it with a process-supporting tier to facilitate the development of more complex data mining projects, and offer an intuitive platform to develop real case studies or proofs of concept for potential industrial users, with no license fees.

The resulting system, SwissAnalyst – powered by Weka, thus addresses both the needs of business users who want to explore/demonstrate DM technology, as well as those of researchers wishing to have more support for the overall *process* of DM.

Although not currently as graphically-rich, SwissAnalyst includes most features necessary to offer as much functionality and process support as the best available commercial packages (Giraud-Carrier and Povel, 2003). As GNU GPL software written in Java, it is distributed freely and can easily be extended. This short paper details and illustrates the key features of SwissAnalyst.

## 2. RELATED WORK

Although most data mining algorithms belong to the public domain, there are ironically very few freeware (i.e., GNU GPL) DM tools or packages. To the best of our knowledge, there are only three (with the exclusion of Weka itself) at the time of this writing:

YALE (Ritthoff, 2001; Fischer, 2002)

XELOPES (Thess and Bolotnikov, 2003)

Orange (Zupan and Demsar, 2003)

YALE is similar in essence to SwissAnalyst, in that it was also designed to specify and execute (although with a somewhat different approach) complex learning chains for pre-processing as well as multi-strategy learning. In addition, the latest version of YALE comes with an "interface" to Weka thus allowing users to leverage all of Weka's algorithms. There are

two main differences, however, between YALE and SwissAnalyst. On the one hand, YALE lifts the requirement that all data fit in main memory, thus ensuring scalability, whilst SwissAnalyst inherits that limitation of Weka. On the other hand, YALE has no graphical user interface and all projects must be specified directly in XML, whereas SwissAnalyst produces XML from (semi) "graphical specifications," thus increasing usability.

XELOPES is a rich, PMML-compliant, open library for embedded data mining. It has a connector to Weka and is easily extensible with custom-defined classes. XELOPES does include a graphical user interface, but its purpose is "not to provide a professional Data Mining tool but to explore the usage of the XELOPES library." (Thess and Bolotnikov, 2003, p. 210). The interface organizes information around a static schema for creating and applying algorithms, and does not directly support the notions of process flow and learning streams, as does SwissAnalyst.

Much like XELOPES, Orange is component-based, making available for integration in other applications a number of data mining techniques. Although currently far more restricted in functionality than SwissAnalyst, the work on Orange, with its Widgets, is another attempt at offering explicit DM process support.

Finally, we note that concurrent to, but independent of, the development of SwissAnalyst, the Weka team has also been improving its software (Version 3.4.1 at the time of writing), particularly with respect to visualization capabilities and an attractive "Knowledge Flow GUI" for graphical programming. This latter addition to Weka is very much in the same spirit as SwissAnalyst.

## 3. KEY FEATURES OF SWISSANALYST

This section gives a short account of the most practitioner-relevant features of SwissAnalyst:

- Process-oriented view
- Support for multiple streams
- XML save/reload
- Enhanced data exploration
- Improved results section
- Model pre-selection
- Enhanced pre-processing

Wherever applicable, illustrative screenshots are included. Familiarity with Weka is assumed.

## 3.1 Process-oriented View

Since Data Mining is a *process,* it seems only natural that DM software should explicitly support that process. Several process models, such as CRISP-DM (SPSS, 2000) and SEMMA (SAS, 1998), have recently been developed. Although each sheds a slightly different light on the DM process, their basic tenets are the same. A recent survey suggests that the most widely used methodology is CRISP-DM (KDnuggets Poll, 2002b).

SwissAnalyst organizes information on the screen in such a way that the entire process, inspired by CRISP-DM, is set out clearly and that it supports naturally the flow of activities within this process, as shown in Figure 1.

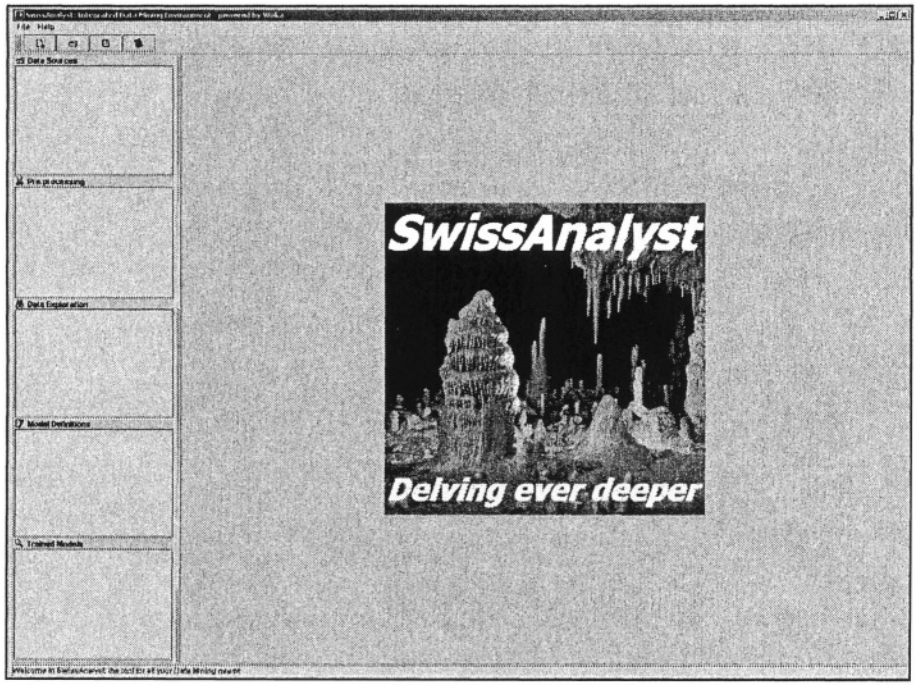


*Figure 1.* SwissAnalyst's Screen Layout

The five panels on the left-hand side correspond roughly to the standard activities of the DM process comprised between Business Understanding and Deployment:

- Data Sources: locating, loading and browsing datasets
- Pre-processing: defining and applying various transformations to the data
- Data Exploration: statistics, warnings and data visualization
- Model Definition: defining and executing various mining models

- Trained Models: evaluating models and applying them to new data

All menus are context-sensitive and accessible through right-mouse clicks over any object in the interface. For each section or object selected in a section, right-clicking presents the list of operations available for that section or object, as illustrated in Figures 2 and 3.

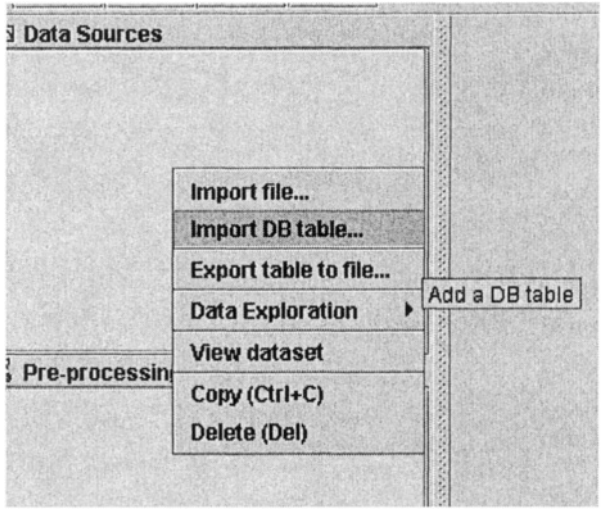


*Figure 2.* Context-sensitive Menus – Data Sources

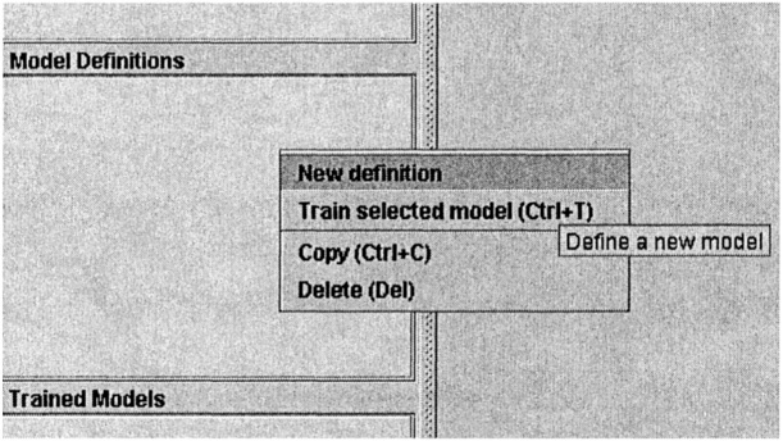


*Figure 3.* Context-sensitive Menus – Model Definitions

This simple design decision ensures that the interface consistently shows the right information at the right time, and applies the selected operation to

the intended object. A similar result is achieved in Weka's new Knowledge Flow GUI, where sets of operations are listed as icons under context-sensitive tabs.

## 3.2 Support for Multiple Streams

Typical DM projects are experimental in nature and thus require the ability to support multiple data or execution flows for analysis and comparison. SwissAnalyst allows complex projects to be defined that incorporate many data sources and operations, and various process flows between them.

Streams (or learning chains) thus defined are context-sensitive so that whenever any component of a stream is selected, the entire stream is highlighted on screen as shown in Figure 4.

Stream highlighting provides the user with a simple visual cue as to the overall process. Note also that although default unique names are automatically created by SwissAnalyst for objects (see, for example, the Data Exploration panel in Figure 4), these can easily be changed by double-clicking the objects.

Furthermore, comments may be attached to all steps, thus allowing the user to record specific insight or clarification. Comment fields are absent from most DM software, yet they often prove essential in documenting the DM process, facilitating maintenance (changes to models, etc.) and knowledge transfer.

A small note on streams involving more than one pre-processing step is worth making. Although one is indeed able to apply pre-processing to a pre-processed step (i.e., some object in the Pre-processing panel), one should avoid doing so as only one object can be highlighted in the top-level stream. Hence, only one (here, the last one) of the pre-processing steps will be highlighted. Instead, one should use a single pre-processing step and edit it at will since each pre-processing step is essentially a sub-stream. There is nothing to be gained in functionality in having more than one pre-processing steps, since they are linked sequentially anyway, which can be handled at the sub-stream level (see Section 3.7).

## 3.3 XML Save / Reload

Complete projects, including induced mining models, are saved in XML format for ease of re-use and communication with other applications. Hence, models can be defined, trained and saved for later application to new datasets (e.g., unlabeled data to produce predictions).

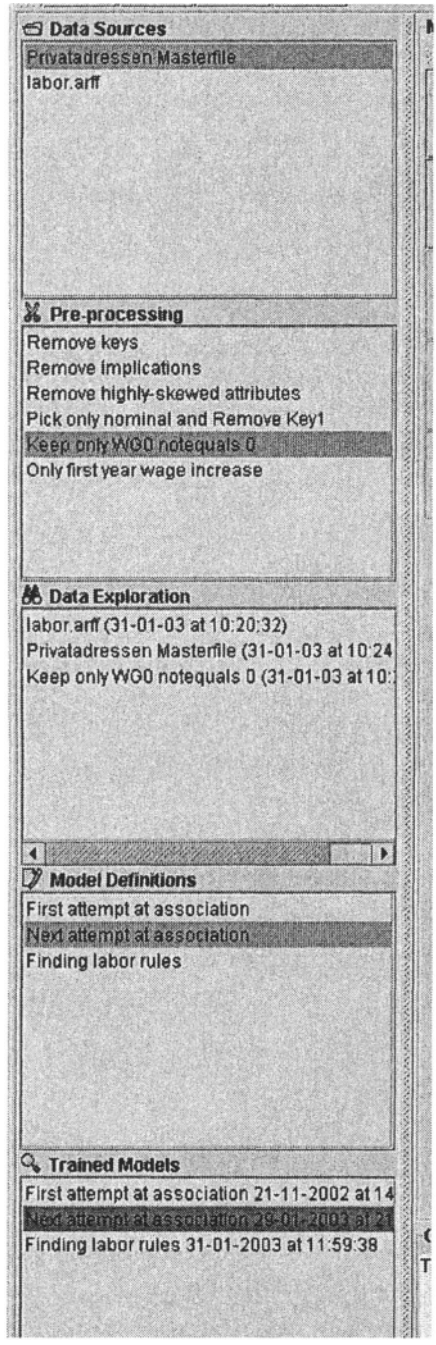


*Figure 4.* Data Mining Streams

The following is an excerpt from a sample XML project file.

```
<xml version="1.0" encoding="utf-8">
   <CommonInfo>
      <Value ProjectName="Project 1.xml"></Value>
...
   </CommonInfo>
      <Section name="Datasets">
         </Dataset>
         <Dataset name="contact-lenses.arff">
      <value FileName="contact-lenses.arff"></value>
      ...
         </Dataset>
      </Section>
      <Section name="PreProcesses">
         <PreProcess name="Keep only WG0 notequals 0">
      <value Parameter="SourceType">PreProcess</value>
...
         </PreProcess>
      </Section>
      <Section name="Statistics">
         <Statistic name="labor.arff (31-01-03 at 10:20:32)">
      <value Parameter="SourceType">DataSet</value>
...
         </Statistic>
      </Section>
      <Section name="ModelDefinitions">
         <MiningModel name="First attempt at association">
      <value Parameter="SourceType">PreProcess</value>
      ...
         </MiningModel>
      </Section>
      <Section name="Trainingresults">
         <Training name="First attempt at association 21-11-2002 at 14:22:53">
      <value Parameter="SourceType">MiningModel</value>
...
         </Training>
      </Section>
</xml>
```

Once a model is trained/saved, it can be executed or applied to any compatible dataset. Such an execution produces a dataset (in the Data Sources panel) equivalent to the original one with two additional columns, one for the predicted values and one for the corresponding confidences in those predictions. Using the Pre-processing capabilities of SwissAnalyst, it is then easy to extract useful actionable data, such as: all records (or adequate projections thereof) predicted as some value with a confidence level above some threshold (e.g., names and addresses of all prospects predicted to respond with a confidence higher than 85%).

## 3.4 Enhanced Data Exploration

When a dataset is loaded into SwissAnalyst (and whenever it is later selected in the Data Sources panel), a “data dump” of about 50 instances/records is automatically displayed on screen (in tabular form) so

that users may get a quick and intuitive feel for the data. It is also possible to view the entire dataset.

Further data exploration is accessible via a sub-menu when selecting any dataset, either in the Data Sources panel or the Pre-processing panel, and results are stored in the Data Exploration panel.

In addition to providing detailed data statistics and a simple bar-chart-like viewer for value distributions, SwissAnalyst can automatically warn the user of pathological attributes, such as single-valued attributes, keys, attributes with a high proportion of missing values, etc. This information, summarized on a single report, shown in Figure 5, is useful in the data preparation phase of the DM process for the correct handling of such attributes (e.g., ignore, fill-in missing values, etc.).

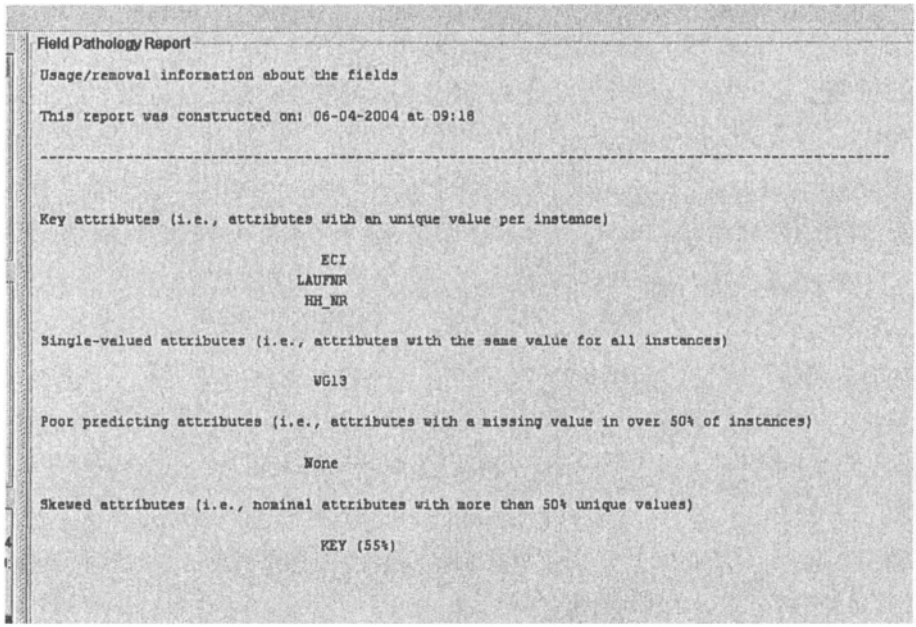


*Figure 5.* Field Pathology Report

## 3.5 Improved Results Section

The output/results section of Weka has been re-organised and enhanced. Results have been split into five categories:

- Basics : provides general information about the model, the data mining stream and any associated comments.

- Model: describes the induced model (e.g., text-based representation of decision tree, list of associations), including graphical representations when applicable (e.g., J4.8).
- Statistics: presents the model's statistics (i.e., confusion matrix, etc.).
- Graphs: displays response, lift and gain charts as well as ROC curves for all class values.
- Comparison: displays a 3-column table, which, for each instance of the test set, shows the actual class value, the predicted class value and the confidence in the prediction.

Some of these categories are disabled for certain types of mining algorithms (e.g., there are no graphs when performing clustering).

Figure 6 gives an illustration of the results section, with the graphs category selected. All graphs can be printed easily by right-clicking the mouse.

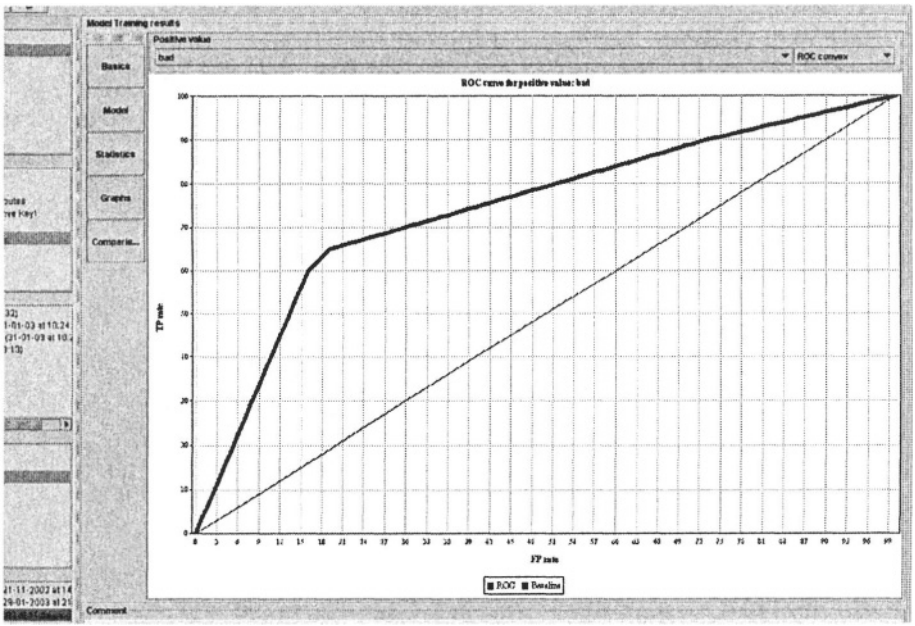


*Figure 6.* Results Section

Although Weka does offer support for some visualization/graphs (e.g., ROC curve), results are only displayed for one class value at a time, requiring the user to go back to the main window to select another class value, and opening a new window for the display. SwissAnalyst allows users to view the ROC curve of each class value simply by selecting the class value from a drop-down list in the graphs section.

Furthermore, from the same interface and with a similar drop-down list mechanism, SwissAnalyst displays response charts, as well as lift and gain charts.

## 3.6 Model Pre-selection

To assist novice users and avoid the unnecessary generation of error messages, SwissAnalyst implements a simple pre-selection scheme for classification algorithms.

Based on the features' types for a given data source, SwissAnalyst displays only those classification algorithms that are applicable (e.g., if the class attribute is continuous, then ID3 is not shown as an option). Figure 7 shows an example, where input features are of mixed types and the class is continuous.

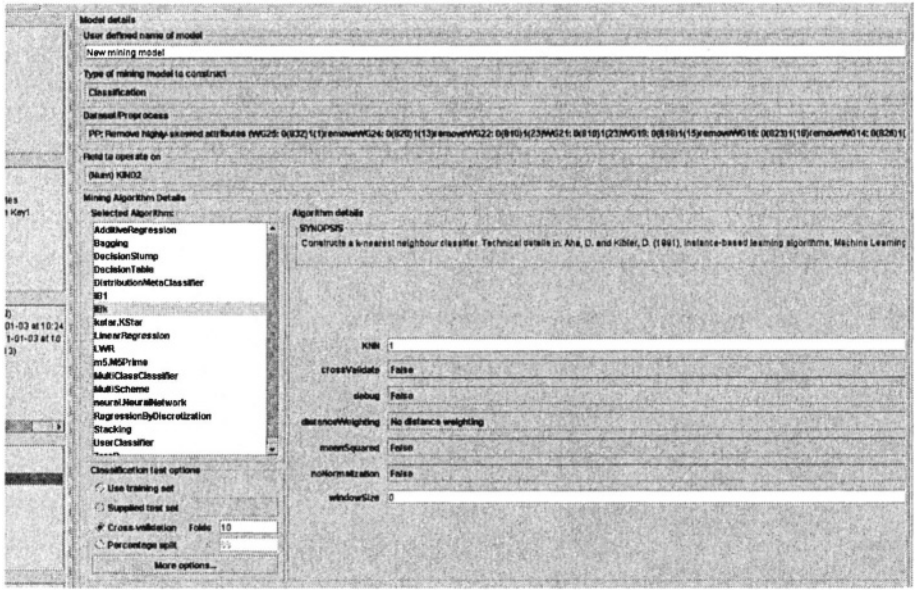


*Figure 7.* Model Pre-selection

Note that the pre-selection dynamically adapts as the user selects the type of mining model, the data source to operate on and the field to predict when applicable.

More sophisticated pre-selection mechanisms, such as the ontology-based approach of IDEA (Bernstein et al. 2002) or METAL's meta-learning approach (Farrand, 2002), both of which implemented in Weka, can easily

be added. Such mechanisms further assist users in selecting and combining DM operators most suitable to a particular task and objective.

## 3.7 Enhanced Pre-processing

As in Weka, the pre-processing phase of the DM process in SwissAnalyst allows users to apply more than one pre-processing algorithms to the selected data source, thus creating a kind of pre-processing sub-stream. Elements of such substreams can not only be added and removed, but modified and re-ordered at will.

Figure 8 shows a simple sub-stream with two pre-processing tasks, an instance selection step, which filters instances based on some condition (generally based on the value of some attribute), and an attribute selection step, which removes specific attributes from the analysis.

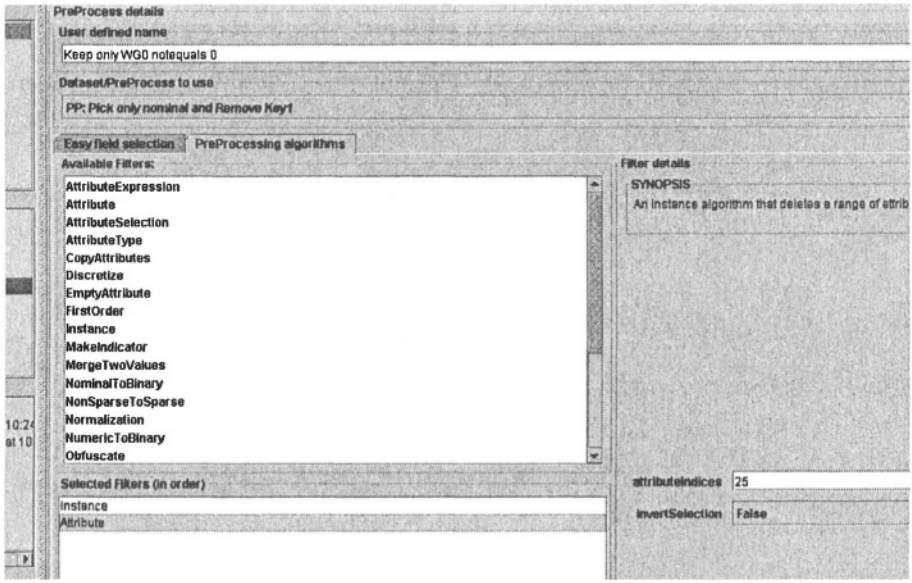


*Figure 8.* Pre-processing Sub-stream

Note that the "Easy Field Selection" tab allows users to quickly perform manual attribute selection (select/de-select) by clicking on individual attributes.

Pre-processing algorithms, known as filters in Weka, often apply to one or more attributes. These are specified by their index value. Historically, filters have been added to Weka from varying sources so that there is no standard indexing schemes, some filters assuming that index values start at 0

and others at 1. SwissAnalyst rectifies this by allowing uniform indexing, starting numbering at 1. Furthermore, to avoid confusion, all index values are automatically re-calculated when attributes are selected/de-selected by the user.

The AttributeSelectedClassifier and the FilteredClassifier of Weka have been purposely removed as they can be reconstructed more naturally and in line with the DM process by applying first a pre-processing task (i.e., attribute selection or an arbitrary filter) and then a mining algorithm.

## 4. OBTAINING SWISSANALYST

Instructions to download SwissAnalyst may be obtained from datamining@elca.ch. Installation only requires the decompression of the archive to any directory and execution of the accompanying script file. SwissAnalyst has been successfully tested on both PC Windows and Unix platforms.

SwissAnalyst, powered by Weka, is also distributed as GNU GPL software. We will continue to improve it and encourage others to do likewise. No claim is made that the current version is either bug-free or DM panacea. It simply provides a clear process-oriented wrapper for Weka's functionality and offers a reasonable platform for further development.

## 5. CONCLUSION

This short paper describes SwissAnalyst, a complete process-driven DM environment powered by Weka.

We contend that SwissAnalyst includes enough functionality and process support, with an intuitive graphical user interface, to make it attractive for (license-free) business proofs of concept as well as for advanced research purposes. Its open source character facilitates future enhancements, based on experience. Our own work will focus on:

- Extending data exploration with additional information, flexibility and advice.
- Improving user support by combining IDEA (Bernstein et al., 2002) and an incremental form of the METAL advice strategy (METAL, 2002).
- Adding further model visualization tools.

## ACKNOWLEDGEMENTS

Special thanks to our early beta-testers who provided valuable feedback on SwissAnalyst.

## REFERENCES

Bernstein, A., Hill, S. and Provost, F. (2002): Intelligent Assistance for the Data Mining Process: An Ontology-based Approach. New York University – Leonard Stern School of Business, Center for Digital Economy Research, *CeDER Working Paper # IS-02-02.*

Farrand, J. (2002). *WekaMetal.* Software available at www.cs.bris.ac.uk/~farrand/wekametal.

Fischer, S., Klinkenberg, R., Mierswa, I. and Ritthoff, O. (2002). *YALE: Yet Another Learning Environment – Tutorial.* CI-136/02, Collaborative Research Center 531, University of Dortmund, ISSN 1433-3325. Software available at/yale.cs.uni-dortmund.de.

Giraud-Carrier, C. and Povel, O. (2003). Characterizing Data Mining Software. *Journal of Intelligent Data Analysis,* **7**(3): 181-192.

KDnuggets Poll (2002a). *Data mining tools you regularly use.* Full results available at www.kdnuggets.com/poll/data_mining_tools_2002_june2.htm.

KDnuggets Poll (2002b). *What main methodology are you using for data mining?* Full results available at www.kdnuggets.com/polls/methodology.html.

METAL (2002). *METAL: A Meta-Learning Assistant for Providing User Support in Machine Learning and Data Mining.* ESPRIT Project Nr 26.357. Official site at www.metal-kdd.org/.

Ritthoff, O., Klinkenberg, R., Fischer, S., Mierswa, I. and Felske, S. (2001). *YALE: Yet Another Learning Environment.* LLWA 01 – Tagungsband der GI-Workshop-Woche Lernen – Lehren – Wissen – Adaptivität, Forschungsberichte des Fachbereichs Informatik, Universität Dortmund, Nr. 763, ISSN 0933-6192.

SAS (1998). *From Data to Business Advantage: Data Mining, The SEMMA Methodology and SAS Software.* Available as a SAS Institute White Paper at www.sas.com.

SPSS (2000). *CRISP-DM 1.0: Step-by-step data mining guide.* Available as an SPSS White Paper at www.spss.com.

Thess, M. and Bolotnikov, M. (2003). XELOPES Library Documentation, Version 1.1.7, prudsys AG. Software available at www.prudsys.com.

Witten, I.H. and Frank, E. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations,* Morgan Kaufmann. Software available at www.cs.waikato.ac.nz/ml/weka/index.html.

Zupan, B. and Demsar, J. (2003). *Orange.* Software available at magix.fri.uni-lj.si/orange.