International Congress on Environmental Modelling and Software

4th International Congress on Environmental Modelling and Software - Barcelona, Catalonia, Spain - July 2008

Jul 1st, 12:00 AM

Pre and Postprocessing in KLASS

Karina Gibert

R. Nonell

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Pre and Postprocessing in KLASS

Karina Gibert[1,2], R. Nonell[1]

[1] Department of Statistics and Operation Research, Technical University of Catalonia, Barcelona
[2] Knowledge Engineering and Machine Learning Group, Technical University of Catalonia, Barcelona, Catalonia

Keywords: Knowldege Discovery from Databases, Artificial Intelligence, Statistics, Prior Knowledge, Clustering based on Rules, PostProcessing, Interpretation Support Tools, Reporting,Data Mining.

KLASS (Gibert et alt 2005b) is a software originally conceived for Knowledge Discovery (KDD) in real domains with complex structure (Gibert et alt 1999). It provides some mixtures of statistical and artificial intelligence tools to support KDD, including basic statistics and providing an integrated system to support the whole process of KDD including pre and post processing, provided that the main data mining technique to be used is related with clustering or rule induction (Gibert et alt 2005c).

Regarding preprocessing, KLASS offers functionalities for basic statistics (simple or by groups), histograms, boxplots (side-by-side), (letter)plots, cross-tables. The performance of the system is quite high, since the user has control over many parameters of de graphics (like the number of classes of a histogram, or the limits of the axis of a plot), being a very flexible tool. It also offers a complete module of data managing, including missing data treatment, creation of transformed variables either using mathematical expressions or via recodification or discretization (here the Boxplot based discretization is provided, which discretizes the numerical variable in such a way that the resulting qualitative variable maximizes association degree with a previously discovered class variable, Pérez-Bonilla et alt 2007). Construction of a prior expert knowledge base (which can be non-complete) is also available and it can be used to biass a posterior clustering process, by means of the Clustering Based on Rules option (Gibert et alt 1999), in such a way that the final classes hold the semantic constraints expressed by the rules.

Regarding the postprocessing, Klass offers some interesting tools to support the interpretation of a clustering results, apart from the classical representation of the dendrogram; It also provides the Class Panel Graph (Gibert et alt 2005), which is a very interestig possibility in clustering contexts to understand better the meaning of the classes. It also implements the CCCS methodology (Pérez-Bonilla 2007) for assigning concepts to every class, improving even more the support to the understanding of the results. There is also a function for visualizing knowledge bases, either probabilized or not, and selecting the rules with degrees of certainty over a certain threshold.

One of the particularities of the system is that it is designed in such a way that the outputs, either graphical or numerical or textual,

are produced in LaTeX font files, which are directly processed by the kernel of KLASS and automatically sent to the LaTeX viewer and displayed on the screen. From the final user point of view, this makes no difference with other systems, since graphical representations are directly displayed on the screen as well as other results. However, as reporting the results of the KDD process is always involved with the ellaboration of technical papers, KLASS also includes a reporting facility in such a way that the user can specify a set of steps to be performed sequentially and a single big LaTeX document including all the results is produced. The user only needs to edit this document and add personal comments on it to get a complete report of the analysis. KLASS provides either standard or personalized reports. It is a flexible possibility since it is possible to automatically transform every result of the single steps into PostScript or PDF documents, which can be managed as usual, for example, pasting it into a Word document.

If the document to be produced is long and with complex structure and contains hard mathematical notation, LaTeX offers nice advantages and the LaTeX results provided by KLASS are really useful. In this cases, LaTeX is a widely used text processor, owing to the excellent support it provides to the generation of high quality mathematical formulae and scientific notation. However, including graphical representations from commercial statistical packages in a LaTeX document requires the use of special LaTeX packages to deal with graphical formats and makes a little bit more complicate the ellaboration of the document, which requires transformation to PostScript or PDF to be completely visualized. Since the results of KLASS are produced in native LaTex code, inclusion of those graphics in the final report becomes trivial.

On the other hand, making the native LaTeX code accessible to the user permits the user to adjust labels or size of the titles of graphical representations. In this way, the quality of the image is maintained to its final use.

Acknowledgments: The project TIN2004-01368 has partially financed the development of GESCONDA.

References

Gibert, K., Annicchiarico, R., Cortés, U. and Caltagirone, C. 2005a. *Knowledge Discovery on Functional Disabilities: Clustering Based on Rules Versus Other Approaches*. IOS Press.

Gibert, K., Nonell, R., Velarde, JM, Colillas, MM 2005b. Knowledge discovery with clustering: Impact of metrics and reporting phase by using klass. *Neural Network World*, 319-326.

Gibert, K., Nonell, R, 2005c. Descriptive statistics with KLASS. Supporting LaTeX documents ellaboration, In Procs. 3rd World Conferencs on Computational Statistics and Data Analysis (IASC 2005), pp 90. Limassol, Cyprus.

Gibert, K. and Sonicki, Z. 1999. Clustering based on rules and medical research. *Journal on Applied Stochastic Models in Business and Industry*, formerly JASMDA 15(4): 319-324.

Pérez-Bonilla A, Gibert, K (2007) Automatic generation of conceptual interpretation of clustering. In Progress in Pattern Recognition, Image analysis and Applications.Lecture Notes in Computer Science v 4756, pp 653-663. Springer.