UNIVERSITY OF NAMUR

U.PORTO

Workshop in Symbolic Data Analysis

Namur, Belgium

June 7-9th, 2011

Programme

www.info.fundp.ac.be

Organized within the framework of the 40th anniversary of the
Faculty of Computer Science of the University of Namur

WORKSHOP IN SYMBOLIC DATA ANALYSIS

Foreword

With the steady development of computer capabilities, recorded datasets that can be massively huge and present complex structures are the more and more common. Therefore, analysis techniques that cope with the size and complexities of the data are constantly required by companies' end up users, who expect analysts to provide interpretations of these data and be able to give quick answers to their questions.

Symbolic Data Analysis (SDA), providing a framework where the variability and complex structures observed may effectively be considered in the data representation, and methods be developed that take it into account, is hence a fast growing discipline, with an increasing field of application. In the recent years, it continued to develop at an intensive rate. The community of researchers in SDA includes nowadays teams in many different countries from all over the world, who are developing new methodologies and applying SDA in a large number of different fields.

About one and a half years after the last Workshop on Symbolic Data Analysis (SDA), held in Wienerwaldhof, time has come to gather again people from different teams working in this field, to review recent developments, from both a theoretical and methodological point of view, to know about practical applications and new software available, as well as to discuss lines for future research. The program, mixing oral presentations and open discussions, intends to foster interaction so as to open the way to future cooperation between participants. All researchers and teams who develop research or recently became interested in this domain are invited.

Namur, June 7th.

Paula Brito
Monique Noirhomme

Programme

Tuesday, June 7th

13h30 - 14h15 Welcome

14h15 - 14h30 **Dean's word**
Jean-Marie Jacquet, Dean of Faculty of Computer Science, University of Namur, Belgium
Presentation, discussion of objectives
Monique Noirhomme-Fraiture, University of Namur, Belgium and Paula Brito, Universidade do Porto, Portugal

Session 1 - Data and Applications

14h30 - 15h00 **Symbolic Data Analysis for Complex Data**
Edwin Diday, CEREMADE, University Paris-Dauphine, France

15h00 - 15h30 **An Example of Distribution-valued Data**
Masahiro Mizuta, Hokkaido University, Sapporo, Japan

15h30 - 16h00 **Symbolic Data Analysis of Cancer Care Trajectories in the Region of Burgundy: Application to Lung Cancers**
Gilles Nuem, CHU Dijon, France, Filipe Afonso, Carole Toque, Myriam Touati, Edwin Diday, Université de Paris Dauphine, Paris, France and Catherine Quantin, Université de Bourgogne, France

16h00 - 16h30 *Coffee Break*

Session 2 - Statistical Approaches

16h30 - 17h00 **Some Maximum Likelihood Estimations for Interval Data**
Lynne Billard, University of Georgia, USA

17h00 - 17h30 **Linear Regression with Histogram-valued Variables**
Sonia Dias, Escola Superior de Tecnologia e Gestão, Portugal and Paula Brito, Université de Porto, Portugal

17h30 - 18h00 **Linear Regression of Interval-valued Data based on Complete Information in Hypercubes**
Huiwen Wang, Rong Guan and Junjie Wu, School of Economics and Management, Beihang University, Beijing, China

Wednesday, June 8th

Session 3 - Clustering and Classification I

09h00 - 09h30 **Partitioning Methods On Dissimilarity and Similarity Matrices Set**
Francisco de A.T. de Carvalho, Universidade Federal de Pernambuco, Recife, Brazil, Marc Csernel and Yves Lechevallier, INRIA-Rocquencourt, Le Chesnay, France

09h30 - 10h00 **Clustering of Symbolic Objects Represented with Discrete Distribution**
Simona Korenjak-Cerne, Natasa Kejzar, Vladimir Batagelj, University of Ljubljana, Slovenia

Programme

10h00 - 10h30 **Clustering Symbolic Data Based on Quantile Representation**
Paula Brito, Université de Porto, Portugal and Manabu Ichino, Tokyo Denki University, Japan

10h30 - 11h00 *Coffee Break*

Session 4 - Clustering and Classification II

11h00 - 11h30 **Self-Organizing map for interval-valued data**
Chantal Hajjar, Université Libanaise, Beirut, Lebanon and Hani Hamdan, SUPELEC, France

11h30 - 12h00 **A Similarity-based Decision Tree Algorithm for Symbolic Data Classification**
Teh Amouh, University of Namur, Belgium, Benoît Macq, University of Louvain-La-Neuve, Belgium and Monique Noirhomme-Fraiture, University of Namur, Belgium

12h00 - 12h30 **Ordinal Modal Symbolic Data. An Application to Teacher Evaluation**
Carmen Bravo and Jose Miguel Garcia-Santesmases, Universidad Complutense de Madrid, Spain

12h30 - 14h00 *Lunch*

Session 5 - Factorial Analysis

14h00 - 14h30 **Principal Component Analysis and Metabins for Symbolic Data**
Edwin Diday, CEREMADE, University Paris-Dauphine, France

14h30 - 15h00 **Principal Component Analysis for Aggregated Symbolic Data**
Junji Nakano, Akiyoshi Fukui, Nobuo Shimizu, The Graduate University for Advanced Studies, Japan

15h00 - 15h30 **Correspondence Analysis for Symbolic Multi-Valued Variables**
Oldemar Rodriguez, University of Costa Rica

15h30 - 16h00 **INTERSTATIS: The STATIS Method for Interval Valued Data**
David Corrales, Hewlett-Packard Development Company, San José, Costa Rica and Oldemar Rodriguez, University of Costa Rica

16h00 - 16h30 *Coffee Break*

Session 6 - Time-series

16h30 - 17h00 **Some Advances in Symbolic Time Series Forecasting**
Javier Arroyo, Universidad Complutense de Madrid, Spain

17h00 - 17h30 **Modeling Interval Time Series with Space-time Processes**
Paulo Teles and Paula Brito, Universidade do Porto, Portugal

17h30 - 18h00 **A Temporal Symbolic Data Analysis based on Beanplots**
Carlo Drago, Carlo Lauro and Germana Scepi, University of Naples Frederico II, Italy

Programme

Thursday, June 9th

Session 7 - Software I

9h00 - 9h30 **Symbolic Data Analysis and *R***
Natasa Kejzar, Simona Korenjak-Cerne and Vladimir Batagelj, University of Ljubjana, Slovenia

9h30 - 10h00 **MAINT.DATA: Modeling and Analysing Interval Data in *R***
A. Pedro Duarte Silva, Universidade Catolica Portuguesa at Porto, Portugal and Paula Brito, Universidade do Porto, Portugal

10h00 - 10h30 **Bringing the Power of Complex Type Data Analysis into *R***
Teh Amouh and Monique Noirhomme-Fraiture, University of Namur, Belgium

10h30 - 11h00 *Coffee Break*

Session 8 - Software II

11h00 - 11h30 **RSDA: An *R* Package for Symbolic Data Analysis**
Juan de Dios Murillo, National University, Costa Rica, Oldemar Rodriguez, University of Costa Rica and Jhonny Villalobos, National University, Costa Rica

11h30 - 12h00 **The SYR Software for Symbolic Data Analysis of Complex Data**
Edwin Diday, CEREMADE, University Paris-Dauphine, France

12h00 - 14h00 *Lunch*

Session 9 - New Developments in SDA

14h00 - 14h30 **Matrix Visualization for Symbolic Data Analysis**
Chun-houh Chen, Academia Sinica, Taipei, Taiwan, Chiun-How Kao, National Taiwan University of Science and Technology, Taipei, Taiwan, Junji Nakano, The Institute of Statistical Mathematics, Tokyo, Japan, Sheau-Hue Shieh, National Taipei University, New Taipei City, Taiwan, Yin-Jing Tien, Academia Sinica, Taipei, Taiwan and Chuan-kai Yang, National Taiwan University of Science and Technology, Taipei, Taiwan

14h30 - 16h00 **Open Discussion**

16h00 - 16h30 *Coffee Break*

LIST OF PARTICIPANTS

AFONSO Filipe (Roissy Charles de Gaulle, France)		p. ..
AMOUH Teh (Namur, Belgium)	tam@info.fundp.ac.be	p. ..
ARROYO Javier (Madrid. Spain)	javier.arroyo@fdi.ucm.es	p. ..
BATAGELJ Vladimir (Ljubljana, Slovenia)	vladimir.batagelj@fmf.uni-lj.si	p. ..
BILLARD Lynne (Athena Georgia, USA)	lynne@stat.uga.edu	p. ..
BRAVO Maria Del Carmen (Madrid. Spain)	mcbravo@pas.ucm.es	p. ..
BRITO Maria-Paula (Porto, Portugal)	mpbrito@fep.up.pt	p. ..
CHEMSEDDINE Maher (Namur, Belgium)	mch@info.fundp.ac.be	p. ..
CHEN Chun-houh (Taipei, Taiwan)	cchen@stat.sinica.edu.tw	p. ..
DIAS Sonia (Viana do Castelo, Portugal)	sdias@estg.ipvc.pt	p. ..
DIDAY Edwin (Paris, France)	diday@ceremade.dauphine.fr	p. ..
DRAGO Carlo (Le Chesnay, France)	c.drago@mclink.it	p. ..
GUAN Rong (Le Chesnay, France)	florence.barbara@inria.fr	p. ..
HAMDAN Hani (Gif-sur-Yvette, France)	hani.hamdan@supelec.fr	p. ..
HARDY André (Namur, Belgium)	andre.hardy@fundp.ac.be	p. ..
KEJZAR Natasa (Ljubljana, Slovenia)	natasa.hejzar@fmf.uni-lj.si	p. ..
KORENJAK-CERNE Simona (Ljubljana, Slovenia)	simona.cerne@ef.uni-lj.si	p. ..
LECHEVALLIER Yves (Paris, France)	yves.lechevallier@inria.fr	p. ..
MIZUTA Masahiro (Hokkaido, Japan)	mizuta@iic.hokudai.ac.jp	p. ..
NAKANO Junji (Tokyo, Japan)	nakanoj@ism.ac.jp	p. ..
NOIRHOMME-FRAITURE Monique (Namur, Belgique)	mno@info.fundp.ac.be	p. ..
NUEMI Gilles (Dijon, France)	gilles.nuemi@chu-dijon.fr	p. ..
RODRIGUEZ Oldemar (San José, Costa Rica)	oldemar.rodriguez@ucr.ac.cr	p. ..
SAAD HAJJAR Chantal (Gif-sur-Yvette, France)	chantal.hajjar@usj.edu.lb	p. ..
SILVA Pedro (Porto, Portugal)	psilva@porto.ucp.pt	p. ..
TELES Paulo (Porto, Portugal)	pteles@fep.up.pt	p. ..

Tuesday, June 7th

SESSION 1 - DATA AND APPLICATIONS CHAIR: L. BILLARD

Symbolic data analysis for complex data

Edwin DIDAY

Paris Dauphine University, France

Contact author: diday@cerenade.dauphine.fr

Abstract:

In recent years, Symbolic Data Analysis (i.e. SDA), where the units are categories, classes or concepts described by intervals, distributions, sets of categories and the like, becomes a challenging task since many application fields generate massive amounts of «Complex data» which come from different sources, or live in high dimensional spaces. «Complex data» are based on several kinds of observations described by standard numerical or (and) categorical data contained in several related data tables. Sometimes, it refers to distributed data or structured data as hierarchical data, spatial-temporal data or heterogeneous data. In practice, «Complex Data» refers to complex objects like images, video, audio or text documents. «Symbolic Data Analysis « is a new paradigm in which theory, tools and practice have shown its ability to extract new knowledge from such complex data that are difficult to store and to analyze with traditional techniques.

In this talk, firstly we show that symbolic data are complex data as they cannot be reduced to standard data without losing much information; secondly we show how SDA can be useful on complex data of different kinds illustrating by some industrial applications: Text Mining (from telephone calls to a company), multidata tables (for Nuclear Power Plot), hierarchical Data (for the study of pig respiratory diseases).

References:

Afonso, F., Diday, E., Badez, N., Genest, Y. (2010). Symbolic Data Analysis of Complex Data: Application to nuclear power plant. In: *Proc. of COMPSTAT'2010*, Paris.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. 321 pages. Wiley series in computational statistics. Wiley, Chichester, ISBN 0-470-09016-2.

Diday, E. and Noirhomme-Fraiture, M. (Eds. and co-authors) (2008). *Symbolic Data Analysis and the SODAS software*. Wiley, Chichester, ISBN 978-0-470-01883-5.

Fablet, C. *et al* (2010). Classification of Hierarchical-Structured Data with Symbolic Analysis. Application to Veterinary Epidemiology. In: *Proc. of COMPSTAT'2010*, Paris.

An Example of Distribution-valued Data

Masahiro MIZUTA

Information Initiative Center, Hokkaido University, Japan

Contact author: mizuta@iic.hokudai.ac.jp

Keywords: Radiotherapy, IMRT, DVH

Abstract:

In this paper, we introduce dose volume histogram (DVH) data for an important example of distribution valued data.

The descriptions of symbolic data include interval values, histogram values and distribution values. The author has proposed methods for distribution valued data: cluster analysis, discriminant analysis, MDS, etc.

Radiotherapy, also called radiation therapy, plays a pivotal role in the treatment of solid tumors i.e. cancer. Especially, Intensity Modulated Radiation Therapy (IMRT) is commonly used to treat cancers. In cancer therapy, a major difficulty is to destroy tumor cells without harming the normal tissue or organs at risk (OAR). The Linear-Quadratic (LQ) model was employed for the surviving fraction of tumor cells and normal tissues. The evaluation of the planning of IMTR is based on DVH, which we can regard as distribution valued data.
We will talk about the basic concepts of IMRT, DVH, and the survival rate of tumor and OAR from the viewpoint of SDA.

References:

[1] Shirato, H., Mizuta, M., Miyasaka, K. (1995). A mathematical model of the volume effect which postulates cell migration from unirradiated tissues. *Radioth. Oncol.*,35:pp.227-231.

[2] Mizuta, M. (2008). Multidimensional scaling - when the dissimilarity data are distributions. *Statistical Computation and Visualization 2008: A satellite workshop for IASC 2008.*

[3] Mizuta, M. (2009). MDS for distribution valued dissimilarity data, *IFCS@GFKL Classification as a tool for research,* p234.

[4] Mizuta, M. (2010). Discriminant Analysis for Distribution Valued Data. *The 30th International Symposium of Forecasting.*p151.

Session 1 - Data and Applications

Clustering of Cancer Care Trajectories in The Region of Burgundy: Application to Lung Cancers

Gilles Nuemi[1], Myriam Touati[2], Filipe Afonso[3], Edwin Diday[2], Catherine Quantin[1,*]

1 Department of biostatistics and medical informatics laboratory,Teaching hospital of Dijon, France

2 CEREMADE CNRS UMR 7534, Université de Paris Dauphine, Paris, France

3 Syrokko, Paris, France

*__Contact author__: catherine.quantin@chu-dijon.fr

Abstract:

Introduction

With the new version of the Cancer Plan (2009-2013), the treatment of cancers in hospitals is now submitted for authorization. The assessment of the adequacy of care provision depends partly on the quantification of the existing situation and the regulatory threshold minimum annual activity.

Objectives

The main objective is to describe and analyse the trajectories of patients with lung cancer receiving hospital care in the Burgundy region using national administrative data (collected in hospitals).

Materials and methods

The definition of trajectories of hospitalization for cancer was based on the linking of hospitalizations for the same patient, to identify on the one hand the initial hospitalization and then to monitor patients identified as cancer cases for 1 year from the date of their 1st hospitalization. These data contain variations presented in the form of bar chart value variables in the Symbolic Data Analysis framework (Billard and Diday, 2006, Diday and Noirhomme-Fraiture, 2008).

A descriptive analysis of the management was done. Also a cluster analysis to reconstruct classes of trajectories was conducted using specific classification methods for symbolic data.

Results

The stays of 487 patients with lung cancer who underwent surgery for lung cancer were selected. Fifty-six trajectories of the institutions attended were reconstructed.

A symbolic database was build using these identified trajectories as concepts. Explanatory variables, like the hospital first attended, gender or patient's declared department of residence were used to conduct the clustering process.

This process based on symbolic data highlighted 4 distinct classes of trajectories that were described using predefined descriptive variables like age, gender, cancer care provided or the length of stay.

Discussions and conclusions

This is an important step in the description of collaboration between institutions in the region, which could then reveal those that should be developed as part of the next project for the organization of regional health systems.

References:

Billard, L. & Diday, E. 2006. *Symbolic Data Analysis: Conceptual Statistics and Data Mining*, Wiley.

Diday, E. & Noirhomme-Fraiture, M. 2008. *Symbolic Data Analysis and the SODAS Software*, Wiley.

Session 2 - Statistical Approaches

Some maximum Likelihood estimations for interval data

Lynne BILLARD[1], Jennifer LE-RADEMACHER[2], Wei XU[3], Jin-Huarng GUO[4]

1 Department of Statistics, University of Georgia, Athens GA 30602

2 Division of Biostatistics, Medical College of Wisconsin, Milwaukee WI 63226

3 MacroGenics Inc., Rockville MD 20850

4 Department of Applied Mathematics, National Pingtung University of Education, Taiwan

*__Contact author__: lynne@stat.uga.edu

Keywords: Interval data; means; variance; covariance

Abstract:

We obtain maximum likelihood estimators for the mean and variance of interval, histogram-, and triangular-valued variables (see Le-Rademacher and Billard, 2011). These match the empirically based results of Bertrand and Goupil (2000) for interval data and Billard and Diday (2003) for histogram data. These derivations can be extended to find maximum likelihood estimators of covariance functions, which in turn compare with the empirical covariance functions give in Billard (2008).

References:

Bertrand, P. and Goupil, F. (2000). Descriptive statistics for symbolic data. In: *Analysis of Symbolic Data: Explanatory Methods for Extracting Statistical Information from Complex Data* (eds. H.-H. Bock and E. Diday), Springer Berlin, 106-124.

Billard, L. (2008). Sample covariance functions for complex quantitative data. In: *Proceedings World Congress, International Association of Statistical Computing*, Japan.

Billard, L. and Diday, E. (2003). From the statistics of data to the statistics of knowledge: Symbolic data analysis. *Journal of the American Statistical Association* 98, 470-487.

Le-Rademacher, J. and Billard, L. (2011). Likelihood functions and some maximum likelihood estimators for symbolic data. *Journal of Statistical Planning and Inference* 141, 1593-1602.

Linear Regression with Histogram-valued Variables

Sonia Diaz[1], Paula Brito[2]

1 Instituto Politécnico Viana do Castelo, Portugal

2 Univerity of Porto, Portugal

Keywords: Histogram-valued variables, Imprecise data, Linear regression, Mallows distance, Symbolic Data Analysis

Abstract:

In recent years, concepts and methods of classical statistics have been adapted to different types of symbolic variables, such as histogram-valued variables (for a recent survey, see [Noirhomme-Fraiture and Brito (2011)]). In this case, for each observation, the frequency or probability distribution that the histogram-valued variable takes, may be represented by a histogram or, alternetively, by its quantile function. Considering histograms with only one interval with frequency/probability one, leads to interval-valued variables as a special case. For this reason, proposed methods for histogram-valued variables are frequently a generalization of their counterparts for interval-valued variables. In 2000, Billard and Diday proposed a first linear regression model for interval-valued variables and later expanded the work to histogram-valued variables [Billard and Diday (2006)]. Alternative models have been independently proposed for interval-valued variables [Lima Neto and de Carvalho (2009, 2010)] and histogram-valued variables [Irpino and Verde (2010)].

In this work, we propose a new linear regression model for histogram-valued variables. This model aims at finding an error measure to quantify the difference between the observed and estimated values of the histogram-valued variables, defining a linear regression model for histogram-valued variables without forcing a direct linear relationship, and measuring the goodness-of-fit of the model.

To quantify the difference between the observed and estimated values of the histogram-valued variables, the Mallows distance is used. As this distance relies on the quantile functions to represent the values that the histogram-valued variables take for each observation, the proposed model uses this representation, and is defined by the quantile functions of each independent histogramvalued variable and those of their symmetrics. For the estimation of the model parameters, a quadratic optimization problem, subject to non-negativity constraints on the unknowns, is solved. A goodness-of-fit measure is then deduced. Using the proposed linear regression model, examples adapted from other works using histogram and interval-valued variables are presented.

References:

Billard, L. and Diday, E. (2006). Symbolic Data Analysis: Conceptual Statistics and Data Mining. Wiley, Chichester.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. Statistical Analysis and Data Mining, (in press).

Irpino, A. Verde, R. (2010). Ordinary Least Squares for Histogram Data Based on Wasserstein Distance. In COMPSTAT'2010, 19th Conference of IASC-ERS (Physica Verlag), pp. 581-589.

Lima Neto, E.A., de Carvalho, F.A.T. (2009). Linear regression models for symbolic intervalvalued. Computational Statistics and Data Analysis, 54, pp. 333-347.

Linear Regression of Interval-valued Data Based on Complete Information in Hypercubes

Huiwen WANG, Rong GUAN and Junjie WU

School of Economics and Management, Beihang University, China

*Contact author: rongguan77@gmail.com

Keywords: Interval-valued Data, Linear Regression, Hypercubes, Complete Information Method (CIM)

Abstract:

Symbolic Data Analysis (Diday (1987), Bock and Diday (2000), Diday and Noirhomme-Fraiture (2008)) has indicated a promising direction for solving the problem of databases explosion in reallife applications. This paper will mainly focus on linear regression for interval-valued data, one category of the most widely used symbolic data. The technique attempts to model the relationship between one or more explanatory variables and a response variable by fitting a linear equation to the interval-valued observations. Despite of the well-known methods such as CM (Billard and Diday (2000)), CRM (Lima Neto and De Carvalho (2008)) and CCRM (Lima Neto and De Carvalho (2010)), further study is still needed to build a regression model that can capture the complete information in interval-valued observations.

To this end, in this paper, we propose the novel Complete Information Method (CIM) for linear regression modeling. By dividing hypercubes into informative grid data, CIM de.nes the inner product of interval-valued variables, and transforms the regression modeling into the computation of some inner products. Experiments on both the synthetic and real-world data sets demonstrate the merits of CIM in modeling interval-valued data, and avoiding the mathematical incoherence introduced by CM and CRM.

References:

E. Diday (1987). The Symbolic Approach in Clustering and Related Methods of Data Analysis. *In Bock, H.H. 1987, Classi.cation and Related Methods of Data Analysis, Proc. IFCS-87 (Aachen, Germany)*, pp. 673–684.

H.H. Bock and E. Diday (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*. Springer-Verlag, Berlin.

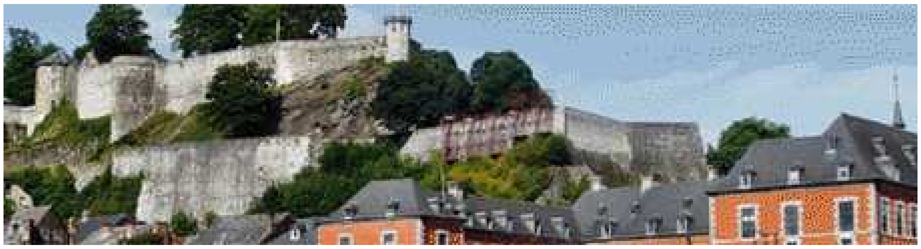
E. Diday and M. Noirhomme-Fraiture (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley-Interscience, Chichester.

L. Billard and E. Diday (2000). Regression analysis for interval-valued data. In *Kiers, H.A.L. Rasson, J.-P. Groenen, P.J.F. Schader, M., 2000, Data Analysis, Classi.cation and Related Methods, Proc. IFCS-00 (Cracow, Poland)*, pp. 369–374.

E.A. Lima Neto and F.A.T. De Carvalho (2008). Centre and range method to fitting a linear regression model on symbolic interval data. *Computational Statistics and Data Analysis 52*, 1500–1515.

E.A. Lima Neto and F.A.T. De Carvalho (2010). Constrained linear regression models for symbolic interval-valued variables. *Computational Statistics and Data Analysis 54*, 333–347.

Wednesday, June 8th



Session 3 - Clustering and Classification I Chair: A. Hardy

Partitioning Methods on Dissimilarity and “Similarity Matrices Set“

Francisco de A.T. de Carvalho[1], Marc Csernel[2], Yves Lechevallier[1,*]

1 Centro de Informatica, Universidade Federal de Pernambuco, Brazil

2 Institut National de Recherche en Informatique et en Automatique (INRIA), Domaine de Voluceau-Rocquencourt, Le Chesnay, France

Contact author: yves.lechevallier@inria.fr

Keywords: Partitioning Clustering Algorithms, Relational Data, Collaborative Clustering, Multiple Dissimilarity Matrices

Abstract:

In this presentation, we introduce partitioning clustering models and algorithms that are able to partitioning objects taking into account simultaneously their relational descriptions given by multiple dissimilarity matrices. The aim is to obtain a collaborative role of the different dissimilarity matrices in order to obtain a .nal consensus partition.

These matrices could have been generated using different sets of variables and a fixed dissimilarity function or using a fixed set of numerical or symbolic variables and different dissimilarity functions, or using different sets of variables and dissimilarity functions.

These methods, which are based on the dynamic clustering algorithm for relational data as well as on the dynamic clustering algorithm based on adaptive distances, are designed to furnish a partition and a prototype for each cluster as well as to learn a relevance weight for each dissimilarity matrix by optimizing an adequacy criterion that measures the .tting between clusters and their representatives.

These relevance weights change at each algorithm iteration and can either be the same for all clusters or different from one cluster to another.

The usefulness of these partitioning algorithms are shown on two time trajectory real world datasets.

References:

L. Kaufman, P.J. Rousseeuw(1990). *Finding Groups in Data* Publisher, Wiley, New York.

W. Pedrycz (2002). Collaborative fuzzy clustering. *Pattern Recognition Letters* 23, 675–686.

H. Frigui, C. Hwanga, F. C.-H. Rhee (2007). Clustering and aggregation of relational data with applications to image database categorization. *Pattern Recognition* 40, 3053–3068.

E. Diday, G. Govaert (1977). Classi.cation Automatique avec Distances Adaptatives. *R.A.I.R.O. Informatique Computer Science 11*, 329–349.

E. Diday, G. Govaert, Y. Lechevallier, J. Sidi (1980). Clustering in Pattern Recognition. *Proc. 5th Conf. Pattern Recognition Miami Beach*, FL, 284–290.

F.A.T De Carvalho, M. Csernel, Y. Lechevallier (2009). Clustering contrained symbolic data. *Pattern Recognition Letter 30*, 1037–1045.

Session 3 - Clustering and Classification I

Clustering of Symbolic Objects Represented with Discrete Distributions

Simona Korenjak-Cerne*, Nataša Kej•ar, Vladimir Batagelj

University of Ljubljana

Contact author: simona.cerne@ef.uni-lj.si

Keywords: k-means clustering, Ward's clustering, alternative error measures

Abstract:

Many real data, especially in social sciences, are composed in the most part of categorical (nominal) values. When condensed according to a given partition of units they produce a special kind of symbolic objects. Their components can be plotted as histograms (numeric data) or barcharts (nominal data) and are not to be confused with histogram symbolic data defined in Billard and Diday, 2006. This condensed representation preserves more information than the representation with only mean values and enables us to consider the variables of different types in the clustering process at the same time. Another advantage of such a symbolic data representation is that it enables as to combine related data sets into one (e.g. ego-centered networks: ego-SO = ego-properties + SO of ego's alters).

We will present the R package **clamix** that implements two clustering methods based on the data descriptions with discrete distributions: the adapted leaders method and the adapted agglomerative hierarchical clustering Ward's method. Both methods are compatible - they can be viewed as two approaches for solving the same clustering optimization problem. We applied both methods on several data sets and some of the results on TIMSS data set and on population pyramids will be presented.

For clustering discrete distributions, we proposed several alternative error functions. They characterize errors between clustered units and a cluster representative-leader (units and leaders are not necessary defined in the same space). For the proposed error functions the adapted leaders methods and compatible agglomerative hierarchical clustering methods were developed that are implemented in R-package **clustddist**. Some of the obtained results of clustering US patents, described with citation distributions will also be presented.

References:

Batagelj, V. (1988). Generalized Ward and Related Clustering Problems. In: *H.H. Bock (Ed.): Classification and Related Methods of Data Analysis*, North-Holland, Amsterdam.

Billard, L., Diday, E. (2006). *Symbolic data analysis. Conceptual statistics and data mining*. Wiley.

Hartigan, J.A. (1975): **Clustering algorithms**. Wiley-Interscience, New York.

Kej•ar, N., Korenjak-Cerne, S., and Batagelj, V. (2011): Clustering of discrete distributions: A case of patent citations, accepted in *Journal of Classification*.

Korenjak-Cerne, S., Batagelj, V., Japelj-Pavešic, B. (2011): Clustering Large Data Sets Described with Discrete Distributions and its Application on TIMSS Data Set. In: L. Billard (Ed.): *Statistical Analysis and Data Mining*, Special Issue on SDA.

Clustering Symbolic Data Based on Quantile Representation

Paula BRITO[1],*, Manabu ICHINO[2]

1 University of Porto, Portugal

2 Tokyo Denki University, Japan

*** Contact author**: mpbrito@fep.up.pt

Keywords: Conceptual clustering, Quantile representation, Symbolic data

Abstract:

Quantile representation [Ichino (2008)] provides a common framework to represent symbolic data described by variables of different types. The principle is to express the observed variable values by some prede.ned quantiles of the underlying distribution and is based on the fact that a monotone property of symbolic objects is characterized by the nesting structure of the Cartesian join regions. For interval-valued variables, a distribution is assumed within each observed interval, this may be Uniform as in [Bertrand and Goupil (2000)] or other; for a histogram-valued variable, quantiles of any histogram may be obtained by simply interpolation, assuming a Uniform distribution in each class (bid); for categorical multi-valued variables, quantiles are determined from the ranking de.ned on the categories based on their frequencies. In the simplest case, when quartiles are chosen, the representation for each variable is de.ned by the 5-uple (Min, Q1, Q2, Q3, Max). Notice, however, that the chosen quantiles need not be equally distributed.

The fact of having a common representation then allows for an uni.ed analysis of the data set, taking all variables simultaneously into account. An appropriate dissimilarity is used to compare data units. In a numerical clustering context, hierarchical and pyramidal models, with several aggregation indices, may be applied and clusters are formed on the basis of quantile proximity.

On a conceptual clustering approach, clusters are represented, for each variable, by a mixture of the quantile-distributions of the merged clusters and then compared on the basis of the current quantile representation. The proposed hierarchical/pyramidal clustering model follows a bottomup approach; at each step, the algorithm selects the two clusters with closest quantile representation to be merged. The newly formed cluster is then represented according to the same model, i.e., a quantile representation for the new cluster is determined from the uniform mixture cumulative distribution. Notice, however, that even if Uniform distibutions are assumed for the initial data, the clusters sucessively formed are generally not Uniform on each variable, putting in evidence different pro.les. Examples with real data sets illustrate the proposed method.

References:

Bertrand, P. and Goupil, F. (2000). Descriptive Statistics for Symbolic Data. In *Analysis of Symbolic Data*, Springer, Heidelberg, pp. 106–124.

Brito, P. and Ichino, M. (2010). Symbolic clustering based on quantile representation. In *Proc. COMPSTAT 2010*, Paris, France.

Ichino, M. (2008). Symbolic PCA for Histogram-Valued Data. In *Proc. IASC 2008*, Yokohama, Japan.

Ichino, M. (2011, in press). The quantile method for symbolic Principal Component Analysis. *Statistical Analysis and Data Mining 4*, 184-198, Wiley (in press).

Session 4 - Clustering and Classification II

A Similarity-based Decision Tree Algorithm for Symbolic Data Classification

Teh Amouh[1],* , Benot Macq[2], Monique Noirhomme-Fraiture[1]

1 Faculty of Computer Science, University of Namur, Belgium

2 Faculty of Engineering, University of Louvain-la-Neuve, Belgium

*** Contact author**: tam@info.fundp.ac.be

Keywords: Symbolic data classification, Decision tree, Similarity measure

Abstract:

In classification problems, we have a training set consisting of observations on a class variable *Y* for *n* entities described by means of *p* predictor variables $Y_1, \cdots, Y_p$. Variable *Y* is a dependent single-valued variable that takes its values from a set $C = \{1, 2, \cdots, K\}$ *of K* classes. The goal is to find a model for predicting the value *Y (e)* of *Y* for any new entity e, when given the values $Y_1(e), \cdots, Y_p(e)$ of variables $\mathrm{Y}_1, \cdots, \mathrm{Y}_p$ describing the new entity *e*.

Decision tree algorithms use a measure of node impurity based on the observed Y values in a node to split the node by (sometime exhaustively) searching over possible splits of the node in order to choose the one that minimizes the average impurity of the child nodes. The process is applied recursively on each child node until a stopping criterion is met.

There are many decision tree algorithms available in the literature. There can be some variations between these algorithms but basically they all allow two possible cases regarding the values of the predictor variables Y_j: ordered values or nominal categories.

Symbolic data, in their generality, are neither ordered values nor nominal categories. Nevertheless, some available solution approaches in the SDA literature try to apply to symbolic predictor variables decision trees that were originally designed for classical type predictor variables. Instead of trying to adapt tree learning algorithms to the characteristics of complex data, these approaches adjust symbolic variables by making assumptions about their complex values in order to be able to apply existing tree learning algorithms. For instance, a uniform distribution is assumed over each interval value. In some approaches [Chavent (2000)], restrictions such as "all predictor variables must have the same symbolic data type" apply. Some other approaches can deal with intervals and distributions simultaneously [Périnel and Lechevallier (2000)], but in this case any other complex types are either not accepted, or transformed into distributions with uniform probability functions. A Bayesian decision tree approach has also been proposed for two-class classification problems [Rasson et al. (2008)]. This approach does not generalize to m-class problems.

Just as nominal categories, symbolic data are unordered. But unlike nominal categories, measures of similarity/dissimilarity are possible with symbolic data. We propose a decision tree algorithm that uses similarity measures in its splitting strategy.

Session 4 - Clustering and Classification II

References

Chavent, M. (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Chapter 11, Section 2: Criterion-Based Divisive Clustering for Symbolic Data, pp. 299-311. Berlin: Springer-Verlag.

Périnel, E. and Y. Lechevallier (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data*, Chapter 10, Section 3: Symbolic Discrimination Rules, pp. 244-265. Berlin: Springer-Verlag.

Rasson, J.-P., P. Lallemand, and S. Adans (2008). *Symbolic Data Analysis and the SODAS Software*, Chapter 17: Bayesian decision trees, pp. 333-340. New York, NY, USA: Wiley-Interscience.

Ordinal Modal Symbolic Data. An Application to Teacher Evaluation

M. Carmen Bravo[1*], Jose Miguel Garcia-Santesmases[2]

[1] Servicio Informático de Apoyo a Docencia e Investigación, Edificio Real Jardín Botánico Alfonso XIII, Universidad Complutense de Madrid, 28040 Madrid, Spain,

[2] Departamento de Estadística e Investigación Operativa, Facultad de Ciencias Matemáticas, Universidad Complutense de Madrid, 28040 Madrid, Spain

*Contact author: mcbravo@pas.ucm.es

Keywords: Consensus Measure, Symbolic Data Analysis, Ordinal Modal Symbolic Data

Abstract:

This paper addresses the problem of analyzing the dispersion of set of objects described by ordinal modal symbolic data. Homogeneous groups are identified in case of a significant dispersion.

We start from the consensus measures given by Leik (1966) and Tastle et al. (2005) for a group of individuals that answer a single question on an ordinal scale that represents the ratings or rankings of each individual preference. A consensus measure for ordinal variables is the complementary of a dispersion measure. We make two extensions of these measures to symbolic data and several issues**.**

1) Under equally spaced assumption between scale scores we propose a new consensus measure for one issue that can be easily generalized to several issues and to a set of classes of individuals (García-Santesmases & Bravo (2010)). This measure satisfies the consensus rules given by Tastle et al. (2005). We define a consensus matrix for a partition of a group of individuals based on the consensus measure between pairs of classes. The consensus measure of a partition is defined based in this matrix and it is easily extended into the framework of symbolic data.

2) Due to the fact that one of the most relevant characteristics of the Leik index is its independence to the sample size and that it makes no assumptions on distances between issue item scale values, this measure is suitable for any ordinal scale. We extent this measure to any ordinal modal symbolic data.

   We apply these measures to a data set that represents 34 teachers rated by their students (1350) on 12 items on a 1 to 4 scale. Each teacher is described by ordinal modal symbolic data. A partitioning around medoids algorithm based solution is proposed (Kaufman & Rousseeuw (1990)). The dissimilarity index used is built out of the consensus measure between pairs of teachers.
   On going work: To develop a clustering optimization algorithm with criteria based on the consensus measure for a set of symbolic data objects.

References:

Bock, H.H., Diday, E. (Eds.) (2000). *Analysis of Symbolic Data. Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer Verlag, Heidelberg.

Session 4 - Clustering and Classification II

Garcia-Santesmases, J.M., Bravo M.C. (2010). Consensus Analysis Through Modal Symbolic Objects. In: *Compstat 2010 proceedings*, Springer, ISBN 978-3-7908-2603-6, 1055-1062.

Kaufman,L., Rousseeuw, P.J. (1990). *Finding Groups in Data: An Introd. to Cluster Analysis*, Wiley.

Leik, K.R. (1966): A measure of ordinal consensus. *The Pacific Sociological Review,* 9:85-90.

Tastle, W.J., Wierman, M.J., Dumdum, U.R. (2005). Ranking Ordinal Scales Using the Consensus Measure. *Issues in Information Systems* 6:96-1.

Self-Organizing map for interval-valued data

Chantal HAJJAR[1,2] , Hani HAMDAN[1,*]

1 SUPELEC, Department of Signal Processing and Electronic Systems, France

2 Université Libanaise, Beirut, Lebanon

*** Contact author**: Hani.Hamdan@supelec.fr

Keywords: Self-organizing maps, Clustering, Interval-valued data, Meteorological stations

Abstract:

The Self-Organizing Maps have been widely used as multidimensional unsupervised classifiers. The aim of our contribution is to develop a self-organizing map for interval data. As a result of the increasing use of such data in data mining, many clustering methods for interval data have been proposed this last decade. In this contribution, we propose an algorithm to train the self-organizing map for interval data. This algorithm is based on the batch training algorithm for the self-organizing maps. Different variants of our algorithm are proposed depending on the distance used to compare two vectors of intervals: L2 distance, Hausdorff distance or city-block distance. In order to show the usefulness of our approach, we apply the self-organizing map on real interval data issued from meteorological stations in China, Lebanon and France. In fact, by using minimal and maximal values to record the temperature, we get a more realistic view on the variation of the weather conditions instead of using average temperature values. In our experiments, we show that the meteorological stations installed in regions close to each other geographically or having similar climate, are assigned to the same cluster or to a neighbor cluster on the self-organizing map.

In the first part of the presentation, we explain briefly the theory of the self-organizing maps and their training algorithms. In the second part, we present our algorithm to train the map using interval data. In the third part, we show the results of our experiments. Finally, we conclude by giving an idea on our future work.

SESSION 5 - FACTORIAL ANALYSIS   CHAIR: M. MIZUTA

Principal component Analysis and Metabins for Symbolic Data

Edwin DIDAY

CEREMADE, University Paris-Dauphine, France

Contact author: diday@ceremade.dauphine.fr

Keywords: Bar chart variables, Clustering, Metabins, Principal Component Analysis, Social network

Abstract:

We propose a strategy for extending standard Principal Component Analysis (PCA) to several kinds of complex data. This leads to variables whose values are «bar charts» (i.e., sets of categories called bins with their relative frequencies). Metabins are ordered sets of such bins (one for each variable), which mix together bins of the different bar charts and enhance interpretability. Some theoretical results lead to the representation of the bar chart variables inside a hypercube covering the correlation sphere.

As shown in Figure 1 the metabins can be associated to each concept, a clustering and a social network on the PCA may also be obtained from the NETSYR software.

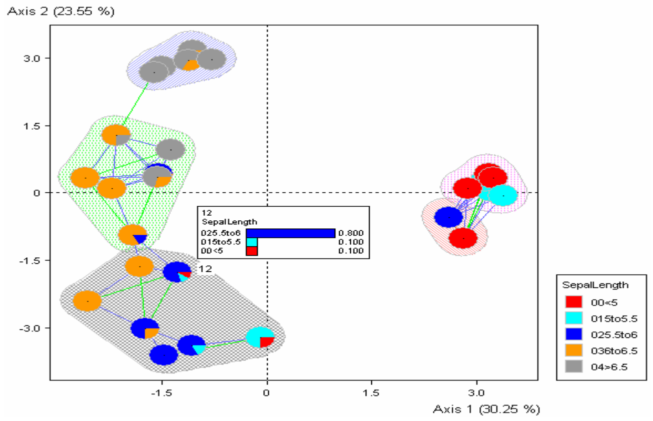


Figure 1: Symbolic PCA of Thirty clusters of Fisher Data

References:

Diday E. (2011). Principal Component Analysis for Categorical Histogram Data: Some Open Directions of Research. In**:** B. Fichet et al. (eds.), *Classification and Multivariate Analysis for Complex Data Structures*, Studies in Classification, Data Analysis, and Knowledge Organization.

Ichino, M. (2008). Symbolic PCA for histogram-valued data. In: *Proceedings IASC*, Yokohama, Japan, 5–8 Dec 2008.

Session 5 - Factorial Analysis

Nagabhsushan, P., Kumar, P. (2007). Principal component analysis of histogram data. In: Liu, D. et al.(eds.) ISNN 2007, Part II, LNCS 4492, pp. 1012–1021. Springer, Berlin, Heidelberg.

Session 5 - Factorial Analysis

Principal Component Analysis for Aggregated Symbolic Data

Junji Nakano[1,2],*, Akiyoshi Fukui[2], Nobuo Shimizu[1,2]

1 The Institute of Statistical mathematics, Japan

2 The Graduate University for Advanced Studies, Japan

*** Contact author**: nakanoj@ism.ac.jp

Keywords: Correlation coefficient, Information reduction

Abstract:

Symbolic Data Analysis (SDA) handles symbolic data (SD), in which values of a variable can be more complex than the traditional data such as real numbers and categorical values. Typical SD take intervals, histograms and bar charts as variable values (Billard and Diday, 2006). SDA provides techniques for handling such SD, including several extensions of principal component analysis (PCA).

In this paper we propose to use correlation coefficients among variables in each SD for PCA of SD. We notice that SD for PCA often arise by aggregation of individuals in groups. In this situation, correlation coefficients are easily calculated together with usual SD information, and are naturally used in PCA of SD.

References:

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining,* Wiley, Chichester.

Correspondence Analysis for “Symbolic Multi–Valued Variables“

Oldemar Rodriguez

CIMPA, School of Mathematics, University of Costa Rica

*** Contact author**: oldemar.rodriguez@ucr.ac.cr

Keywords: Symbolic data analysis, contingency tables, interval contingency table, disjunctive complete table, symbolic multi–valued variables.

Abstract:

This paper sets a proposal of a new method and two new algorithms for Correspondence Analysis. In this method there are two multi–valued symbolic variables X and Y , that is to say, the modality that takes the variables for a given individual is a finite set formed by the possible modalities taken for the variables in a given individual. Then, starting from all the possible classic contingency tables an interval contingency table can be built, which will be the point of departure of the proposed method.

The method is illustrated with an example showing how the proposed method can process multiple–choice statistical questionnaires achieving better interpretations than classic correspondence analysis. This example was built using the program RSymCA in RSDA, an R package developed by the author for Symbolic Data Analysis.

References:

Benzécri, J.P. (1973). *L’Analyse des Données: Tomo 2: L’Analyse des Correspondances*. Dunod, Paris.

Billard, L. and Diday E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*. Wiley, London.

Bock H-H. and Diday E. (eds.) (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer Verlag, Germany.

Castillo, W. y Rodriguez O. (1997). Algoritmo e implementacion del factorial de correspondencias. *Revista de Matematicas: Teoria y Aplicaciones, Editorial Universidad de Costa Rica* 6, 13–17.

Cazes P., Chouakria A., Diday E. et Schektman Y. (1997) Extension de l’analyse en composantes principales à des données de type intervalle. *Rev. Statistique Appliquées XLV (3)*, 5–24.

INTERSTATIS: The STATIS Method for Interval Valued Data

David Corrales[1] and Oldemar Rodriguez[2],*

1 Hewlett–Packard Development Company, San Jose, Costa Rica

2 CIMPA, School of Mathematics, University of Costa Rica

Contact authors: david corrales@acm.org, oldemar.rodriguez@ucr.ac.cr

Keywords: INTERSTATIS, STATIS, interval arithmetic, interval PCA, interval data, compromise

Abstract:

The STATIS method, proposed by L'Hermier des Plantes and Escou.er, is used to analyze multiple data tables, each with information from the same set of individuals. The differences and similitudes between said tables are analyzed by means of a structure called the compromise. In this paper we present a new algorithm for applying the STATIS method when the input consists of interval data. This proposal is based on Moore's interval arithmetic (Moore (1979)) and the Centers Method for Principal Component Analysis with interval data, Cazes et al. (1997). In addition to presenting the INTERSTATIS method in an algorithmic way, important implementation considerations are also discussed. Finally, an example, using the program RSDA, an R package developed by the author for Symbolic Data Analysis, is shown, alongside the interpretation of its results.

References:

Billard, L. and Diday E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*. Wiley, London.

Bock H-H. and Diday E. (eds.) (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Germany.

Cazes P., Chouakria A., Diday E. & Schektman Y. (1997) Extension de l'analyse en composantes principales à des données de type intervalle. *Rev. Statistique Appliquées XLV (3)*, 5–24.

Escoufier, Y. (1980). L'analyse conjointe de plusieurs matrices de données. *Biométrie et Temps. Paris: Societé Francaise de Biométrie 1*, 59–76.

Lavit, Ch. (1988). *Analyse Conjointe de Tableaux Quantitatifs*. Ed. Masson, Paris.

L'Hermier Des Plantes, H. (1976). *Structuration des tableaux a trois indices de la statistique*. Thèse de troisième cycle. Université de Montpellier, France.

Moore, R.E. (1979). *Methods and Applications of Interval Analysis*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, USA.

Some Advances in Symbolic Time Series Forecasting

Javier ARROYO*

Facultad de Informática. Universidad Complutense de Madrid, Spain

***Contact author**: javier.arroyo@fdi.ucm.es

Keywords: Interval time series, Histogram time series, Candlestick time series, Symbolic autocorrelation

Abstract:

This presentation will summarize the works in the field of symbolic time series forecasting made by its author during the last two years. In his doctoral dissertation (Arroyo, 2008), the author extended the catalogue of symbolic data analysis methods to include forecasting methods such as smoothing and k-nearest neighbor methods. However, some important forecasting issues, such as autocorrelation functions and autoregression models, were not tackled in this work. A couple of years later some steps have been taken to address these issues (González-Rivera, G. and Arroyo, J., 2010; Arroyo, J., and González-Rivera 2010). In addition, Arroyo et al. (2011) proposes a novel application of the exponential smoothing methods in Arroyo (2008) to the context of Value-at-Risk.

The aforementioned works deal with intervals and histograms in a financial context. However, other symbolic data arise in the context of finance, such as candlesticks and arrows. Candlesticks summarize the variability of a trading session by means of two pairs of values: the opening-closing and the lowest-highest prices of each session. Arrows are similar to candlesticks, but they also provide information about the time where the highest and the lowest prices occur. The works by Arroyo (2010) and Arroyo and Bomze (2010) propose the use of a k-nearest neighbors method to forecast these new kinds of symbolic time series.

References:

Arroyo, J. (2008). Métodos de predicción para series temporales de intervalo y de histograma. Universidad Pontificia Comilas (Doctoral Thesis).

Arroyo, J. (2010). Forecasting candlesticks time series with locally weighted learning methods. In *Classification as a Tool for Research*. pp. 603-611.

Arroyo, J. and Bomze, I. (2010). Shooting arrows in the stock market. In *COMPSTAT 2010* (Paris, France).

Arroyo, J., and González-Rivera (2010). Interval autoregression: an application to volatility. In *CFE 2010* (London, UK).

Arroyo, J., González-Rivera, G., Maté, C. and Muñoz San Roque, A. (2011). Smoothing methods for histogram-valued time series: an application to value-at-risk, *Statistical Analysis and Data Mining 4*, pp. 216-228.

González-Rivera, G. and Arroyo, J. (2011) Time series modeling of histogram-valued data. The daily histogram time series of SP500 intradaily returns. *International Journal of Forecasting (in press)*.

Modeling Interval Time Series with Space-time Processes

Paulo Teles*, Paula Brito

Faculdade de Economia & LIAAD-INESC Porto LA, Univ. Porto, Portugal

***Contact author**: pteles@fep.up.pt

Keywords: Interval data, Interval time series, Prediction, Space-time AR model

Abstract:

Symbolic Data Analysis provides a suitable framework for data where variability and/or uncertainty might be inherent to each observation. This is the case of interval data, where the observed values of the variables are intervals of IR. Interval-valued data may be represented by the lower and upper bounds of each observed interval or, alternatively, by its center and radius. When the interval-valued symbolic data are collected as an ordered sequence through time or any other dimension, they form an interval time series (ITS). Time series models such as ARIMA processes have been designed and are appropriate for single-valued observations. This leads to the proposal of a new approach for ITS, namely using Space-Time Autoregressive Models (STAR), which allows taking into account the existence of contemporaneous correlation or dependence between the intervals' lower and upper bounds (or center and radius). We start by setting up the bivariate STAR model for the ITS bounds and derive the corresponding bivariate model for its center and radius which is a Structural Vector Auto-regressive model (SVAR) with the same order. The parameters of the latter are functions of those in the former. Important particular cases and their consequences are analyzed. Prediction of the ITS bounds from the respective STAR model and of the center and radius from the respective SVAR model is then discussed. An application of this approach is given where the STAR model for the interval bounds is estimated and checked for adequacy. The corresponding model for the center and radius is derived and its parameter estimates are computed from those of the STAR model. Next, the ITS values are forecast for several periods (out-of-sample forecasts) showing good predictive performance. Finally, the equivalence between the forecasts obtained from the ITS bounds and the ITS center and radius is shown.

References:

Arroyo, J., Gonzalez-Rivera, G., Maté, C. (2011). Forecasting with interval and histogram data. Some financial applications. In: *Ullah, A. and Giles, D., Balakrishnan, N., Schucany, W., Schilling, E., Eds. Handbook of Empirical Economics and Finance*. New York: Chapman and Hall/CR, pp. 247-280.

Billard, L., Diday, E. (2003). From the Statistics of Data to the Statistics of Knowledge: Symbolic Data Analysis. *J. Amer. Statist. Assoc. 98*: 470-487.

Cressie, N.A.C. (1993). *Statistics for Spatial Data*. New York: John Wiley and Sons.

Diday, E., Noirhomme, M. (Eds.) (2008). *Symbolic Data and the SODAS Software*. Chichester: Wiley.

Finkenstädt, B., Held, L., Isham, V. (Eds.) (2007). *Statistical Methods for Spatio-Temporal Systems*. London: Chapman and Hall/CRC.

Teles, P., Brito, P. (2005). Modelling interval time series data. *Proceedings of the 3rd IASC World Conference on Computational Statistics and Data Analysis*. Limassol, Cyprus.

Session 6 - Time Series

A Temporal Symbolic Data Analysis based on Beanplots

Carlo Drago, Carlo Lauro* and Germana Scepi

University of Naples "Federico II", Naples, Italy

***Contact author**: clauro@unina.it

Keywords: Forecasting, Clustering, Beanplots

Abstract:

In this paper, we present a new approach for modeling, forecasting and clustering high frequency temporal data. Data are first grouped according to a suitable temporal window (say, daily, monthly, quarterly ecc) and modeled as to a peculiar density plot, called Beanplot (Kampstra 2002 Drago Scepi 2009), that have a more rich and powerful interpretation with respect to the classical histogram symbolic data. (Billard L., and Diday E.,2003 Arroyo Matè 2008 on histogram time series). The features of a Beanplot allows to analyze intra-period variability by a non parametric approach (model of mixtures) useful to detect structural change points by comparing location, size and shape of this representation. At the same time the parameter of the different Beanplots, offer a multiple vectorial time series that can be modeled through a Factorial Time Series Analysis (Gilbert Meijer 2005) in order to forecast data over the considered period and also for classification aims. In performing the clustering of models we use a suitable distance measure based on a convex function of parameters and the corresponding model goodness of fit index(Romano Giordano Lauro 2006), for example the Bayesian Information Criteria. We will present both the methods and the associated software developed in R on real financial time series.

References:

Arroyo J. Matè C. (2008) Forecasting Histogram Time Series with the K-neirest neighbour methods. International Journal of Forecasting 25 n.1 (2009) 192-207

Billard L., and Diday E. (2003) From the statistics of data to the statistics of knowledge: Symbolic data analysis. Journal of the American Statistical Association, 98 (462), 470-487.

Drago C. and Scepi G. (2009) Univariate and Multivariate Tools for Visualizing Financial Time Series. In Book of Short Paper, Seventh Meeting of the Classification and Data Analysis Group of the Italian Statistical Society, Catania Settembre 2009, Salvatore Ingrassia e Roberto Rocci (eds), Cleup editore, pp 481-485

Kampstra, P. (2008) Beanplot: A Boxplot Alternative for Visual Comparison of Distributions Journal of Statistical Software Vol. 28, Code Snippet 1, Nov. 2008 Beanplot Data Analysis in a Temporal Framework 5

Gilbert P.D and Meijer E. (2005) Time Series Factor Analysis with an Application to Measuring Money, Research Report 05F10, University of Groningen, Research Institute SOM (Systems, Organisations and Management).

Romano E., Giordano G. and Lauro C. (2006) An Inter-Models Distance for Clustering Utility Functions, Statistica Applicata-Italian Journal of Applied Statistics, Vol.17, n.2.

Thursday, June 9th

SESSION 7 - SOFTWARE I CHAIR: O. RODRIGUEZ

Symbolic data analysis and *R*

Nataša KEJ•AR*, Simona KORENJAK-CERNE, Vladimir BATAGELJ
University of Ljubljana

***Contact author**: natasa.kejzar@mf.uni-lj.si

Keywords: SDA software, XML

Abstract:

Throughout the years many methods and algorithms have been developed for analysis of sets of symbolic objects (Billard and Diday, 2006, Diday and Noirhomme-Fraiture, 2008). SODAS computer software is a well recognized tool that supports the analysis of such data sets. As a part of SODAS software the representation of the sets of symbolic data objects on files was developed. It has also a XML binding.

R is a widely used open-source statistical programming framework. For many researchers and other users of statistical methods it represents the main programming language in which simulations and new methods are probed. Currently few *R* packages for symbolic data analysis methods exist, i.e. Dudek (***SymbolicDA*** - package already implements a function parseSO for transformation of XML data set from SODAS into *R*), Kej•ar and Batagelj (***clamix***). However none of them is available on the CRAN yet.

At the meeting in Wienerwaldhof in 2009 we proposed that SODAS and *R* could be connected through the 'XML link' since R provides support for XML files. In the paper we present a discussion and a prototype implementation of this approach.

References:

Billard, L., Diday, E. (2006). *Symbolic data analysis. Conceptual statistics and data mining*. Wiley.

Diday, E. (Ed.), Noirhomme-Fraiture, M. (Co-Ed.) (2008). *Symbolic Data Analysis and the SODAS*. Wiley.

Dudek, A (2010). Analysis of symbolic data; Package 'symbolicDA'. Version 0.01-17, December 13, 2010. http://keii.ue.wroc.pl/symbolicDA

Kej•ar, N., Korenjak-Cerne, S., and Batagelj, V. (2011): Clustering of discrete distributions: A case of patent citations, accepted in *Journal of Classification*.

Session 7 - Software I

MAINT.DATA: Modeling and Analysing Interval “Data in R“

A. Pedro Duarte Silva[1],* , Paula Brito[2]

1 Faculdade de Economia e Gestao & CEGE, Universidade Catolica Portuguesa at Porto, Porto, Portugal

2 Faculdade de Economia & LIAAD-INESC Porto LA, Univ. Porto, Portugal

***Contact author**: psilva@porto.ucp.pt

Keywords: Symbolic data, Interval data, Parametric modelling of interval data, Statistical tests for interval data, Skew-Normal distribution

Abstract:

Symbolic Data Analysis [2], [5], [3], [6]) provided a framework where new variable types allow to take directly into account variability and/or uncertainty associated to each single “individual”, by allowing multiple, possibly weighted, values for each variable. We focus on the analysis of interval data, i.e., where elements are described by variables whose values are intervals of $\mathbb{R}$. Parametric inference methodologies based on probabilistic models for interval variables are developed in [4] where each interval is represented by its midpoint and log-range, for which Normal and Skew-Normal [1] distributions are assumed. The intrinsic nature of the interval variables leads to special structures of the variance-covariance matrix, which are represented by five different possible configurations.

In this work, we introduce the R package MAINT.DATA, which implements the proposed methodologies in the R statistical environment [7]. It introduces a data class for representing interval data. MAINT.DATA includes functions and methods for parametric modeling and analysing of interval data. In particular, it performs maximum likelihood estimation and statistical tests for the different configurations. (M)ANOVA and Linear and Quadratic Discriminant Analysis are also implemented for all considered configurations.

References:

Azzalini, A. and Dalla Valle, A. (1996). The multivariate Skew-Normal distribution. *Biometrika 83 (4)*, 715–726.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. Wiley, Chichester.

Billard, L. (2011). Brief overview of symbolic data and analytic issues. *Statistical Analysis and Data Mining* (in press).

Brito, P. and Duarte Silva, A.P. (2011). Modelling interval data with Normal and Skew-Normal distributions. *Journal of Applied Statistics* (in press).

Diday, E. and Noirhomme-Fraiture, M. (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley, Chichester.

Noirhomme-Fraiture, M. and Brito, P. (2011). Far Beyond the Classical Data Models: Symbolic Data Analysis. *Statistical Analysis and Data Mining* (in press).

R Development Core Team. *R: A language and environment for statistical computing. R Fondation for statistical computing*. Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.com, 2011.

Bringing The Power of Complex Type Data Analysis into *R*

Teh Amouh*, Monique Noirhomme-Fraiture

Faculty of Computer Science, University of Namur, Belgium

***Contact author**: tam@info.fundp.ac.be

Keywords: Data frame, Symbolic data types, Symbolic data table model, R data structures

Abstract:

The data.frame object provided by *R* does not apply as a tabular model for symbolic data collection since the data.frame model allows only scalar types data on columns. In order to bring the power of SDA methods [Diday and Noirhomme-Fraiture (2008)] into *R*, there is a need on a symbolic data table model in *R* where each cell could contain a –possibly weighted– listing of values taken from a predefined set, and not just a single quantitative or categorical value. In addition, the *R* implementation of SDA methods does not go without the ability to represent complex type data such as intervals, distributions and multivalued data in the *R* software framework. These complex data types are not natively provided by *R*. Therefore, each of the existing approaches to the *R* implementation of SDA methods uses its own conventions for the representation of symbolic data types. Such a situation is not ideal, as code sharing and reuse is desired in a research community. A common data types system would be desirable.

We design and implement symbolic data structures using both S3 and S4 object approaches available in *R* [Chambers (2008)]. Our data structures include table objects that extend the data.frame object and allow complex type data value in each cell. This talk is about the development of these data structures and underlying design principles. An *R* package containing these data structures will be available as a basic building bloc for the *R* implementation of symbolic data analysis methods. The main advantage of our approach lies in the fact that it proposes a well designed class system for symbolic data types representation in *R*. Such a class system can provide a commonly shared foundation for SDA methods implementation in *R*.

References:

Chambers, J. (2008). *Software for Data Analysis: Programming with R*. Springer.

Diday, E. and M. Noirhomme-Fraiture (Eds.) (2008). *Symbolic Data Analysis and the SODAS Software*. Wiley.

Session 8 - Software II

SESSION 8 - SOFTWARE II CHAIR: V. BATAGELJ

RSDA: An *R* package for Symbolic Data Analysis

Juan DE DIOS MURILLO[1]*, Oldemar RODRIGUEZ[2]** and Jhonny VILLALOBOS[1]***

1 School of Informatics, National University, Costa Rica

2 CIMPA, School of Mathematics, University of Costa Rica

Contact authors: *jmurillo@una.ac.cr, **oldemar.rodriguez@ucr.ac.cr, ***jvillalo@una.ac.cr

Keywords: Symbolic data analysis, interval principal components analysis, histogram principal components analysis, multi–valued correspondence analysis, INTERSTATIS

Abstract:

The *R* package *RSDA* is being developed for Symbolic Data Analysis. The main features of this package is the possibility to take into account different types of variables (continuous, interval, histogram or multi–valued). Centers interval principal components analysis, histogram principal components analysis, multi–valued correspondence analysis and INTERSTATIS methods have been implemented.

We are using PostgreSQL, a powerful open source object-relational database system, to store the symbolic objects. PostgreSQL is released under the PostgreSQL License, a liberal Open Source license, similar to the BSD or MIT licenses, so we have permission to use, copy, modify, and distribute this software and its documentation for any purpose, without fee.

Numerous graphics and a graphical user interface is being implemented within the *RS-DAcmdr* environment in order to propose an user friendly package. Also, the capabilities and features of the package are illustrated using two data examples.

References:

Billard, L. and Diday E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*. Wiley, London.

Bock H-H. and Diday E. (eds.) (2000). *Analysis of Symbolic Data. Exploratory methods for extracting statistical information from complex data*. Springer, Germany.

Chambers, J.M. (2008). *Software for Data Analysis: Programming with R*. Springer, New York.

Everitt B.S. and Hothorn T. (2010). *A Handbook of Statistical Analysis Using R*. Chapman & Hall book, Florida.

R Development Core Team (2007). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org.

The SYR Software for Symbolic Data Analysis of Complex Data

Filipe Afonso[1], Edwin Diday[2]

1 SYROKKO Company, Roissy Charles de Gaulle, France

2 Paris Dauphine University, France

Contact author: diday@cerenade.dauphine.fr

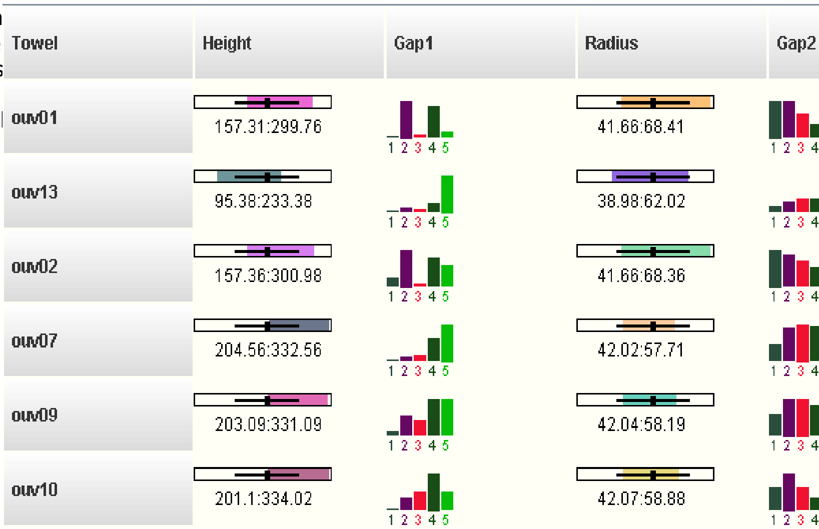
Keywords: Clustering, Principal component analysis, Symbolic data table sorting, Symbolic data analysis software

Abstract:

We present the main modules of the SYR software for Symbolic Data Analysis of Complex Data.

The SYR software is a SYROKKO company product. Its aim is to extract, from a data file (.txt, .csv) of several millions of units or an Access data base of hundreds of thousands units, a reduced number of units called "concepts" which summarize the initial data. These units are described by standard categorical or numerical variables, as well as by interval variables and by bar-chart and histogram-valued variables. These new kinds of variables allow keeping the internal variation of each concept. The software allow:

- Creating a symbolic data file
- Visualizing a symbolic data matrix tha
- Handling a symbolic data matrix (sele
- Using sorting methods allowing to s discriminant to the least discriminant
- Clustering, statistics, PCA etc. of any



Towel	Height	Gap1	Radius	Gap2
ouv01	157.31:299.76	1 2 3 4 5	41.66:68.41	1 2 3 4
ouv13	95.38:233.38	1 2 3 4 5	38.98:62.02	1 2 3 4
ouv02	157.36:300.98	1 2 3 4 5	41.66:68.36	1 2 3 4
ouv07	204.56:332.56	1 2 3 4 5	42.02:57.71	1 2 3 4
ouv09	203.09:331.09	1 2 3 4 5	42.04:58.19	1 2 3 4
ouv10	201.1:334.02	1 2 3 4 5	42.07:58.88	1 2 3 4

Figure 1: Output of TABSYR Module

Session 8 - Software II

References:

Afonso, F., Diday, E., Badez, N., Genest, Y. (2010). Symbolic Data Analysis of Complex Data: Application to nuclear power plant. In: *Proc. of COMPSTAT'2010*, Paris.

Diday, E. and Noirhomme-Fraiture, M. (eds and co-authors) (2008). *Symbolic Data Analysis and the SODAS software*. Wiley, Chichester, ISBN 978-0-470-01883-5.

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining*. 321 pages. Wiley series in computational statistics. Wiley, Chichester, ISBN 0-470-09016-2

SESSION 9 - NEW DEVELOPMENT IN SDA CHAIR: M. NOIRHOMME

Matrix Visualization for Symbolic Data Analysis

Chun-houh CHEN[1,*], Chiun-How KAO[2], Junji Nakano[3], Sheau-Hue SHIEH[4], Yin-Jing TIEN[1], Chuan-kai YANG[2]

1 Academia Sinica, Taipei, 11529 Taiwan

2 National Taiwan University of Science and Technology, Taipei, Taiwan

3 The Institute of Statistical Mathematics, Tokyo, Japan

4 National Taipei University, New Taipei City, Taiwan

Contact author: cchen@stat.sinica.edu.tw

Keywords: Dependent Data, EDA, Huge Data, GAP, Proximity

Abstract:

Matrix visualization techniques such as GAP (generalized association plots: Chen 2002, Tien et al. 2008, Wu et al. 2010) are useful exploratory data analysis tools for visualizing and clustering high-dimensional data structure. Various extensions of GAP for analyzing data with different scales or generated from different models have been developed to make GAP a more versatile environment.

However all these GAP modules treat samples as independent ones and as base unit. We are currently developing GAP modules for data with various types of dependent structure. Hierarchical (multi-level) data, familial genetic data, longitudinal repeated measurements are three such GAP modules. We are also working on GAP modules for handling huge data sets. Both directions, dependent data and huge data, are closely related to that of symbolic data analysis (SDA).

In this talk we will summarize our current studies on matrix visualization for symbolic data analysis with some real data examples. Some difficulties and technical issues on implementing GAP modules for SDA will also be discussed.

References:

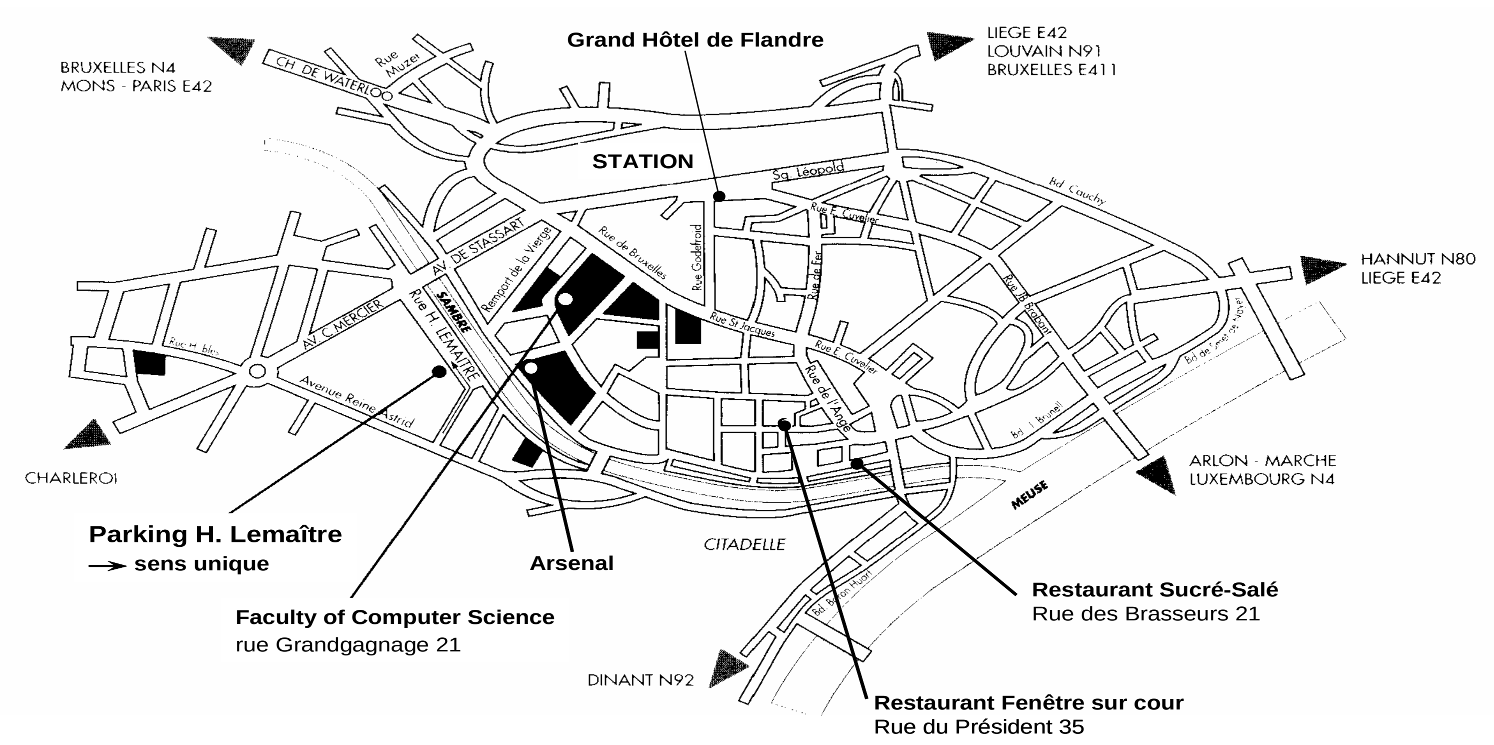
Chen, C. H. (2002). Generalized Association Plots for Information Visualization: The applications of the convergence of iteratively formed correlation matrices, *Statistica Sinica*, 12, 1-23.

Tien, Y. J., Lee, Y. S., Wu, H. M., and Chen, C. H. (2008). Methods for Simultaneously Identifying Coherent Local Clusters with Smooth Global Patterns in Gene Expression Profiles, *BMC Bioinformatics, 9*:155.

Wu, H. M., Tien, Y. J., and Chen, C. H. (2010). GAP: A graphical environment for matrix visualization and cluster analysis, *Computational Statistics and Data Analysis*, 54 (3), 767-778.

Notes

Notes



Grand Hôtel de Flandre

Parking H. Lemaître
→ **sens unique**

Arsenal

Faculty of Computer Science
rue Grandgagnage 21

Restaurant Sucré-Salé
Rue des Brasseurs 21

Restaurant Fenêtre sur cour
Rue du Président 35

http://www.fundp.ac.be/en/inf

Welcome to the Faculty of Computer Science!

Founded in 1968, the Faculty of Computer Science is one of the early pioneers of teaching and research in the areas of information and computer sciences.

From its very beginning, it has aimed at developing a unique curriculum of studies devoted to the design of information systems according to an interdisciplinary and holistic approach. Our teaching programs thus comprise courses related to classical computer science (such as courses on programming, databases, networks, operating systems, computer architecture) but also courses on mathematics and human sciences. Being close to the European headquarters in Brussels, our faculty also supports language learning and student mobility, in particular through a 6 months stage during the last year of the master.

The faculty has a team of 85 members, among which 25 are professors. It offers the degrees of bachelor, master and PhD. Our teaching is supported by research conducted, at an international level, in seven research units, in which problems are handled not only from the technical point of view but also with methodological, organizational and social concerns.
This encompasses both fundamental and applied research with research funds coming from regional, national as well as european projects.

Jean-Marie Jacquet,
Dean of the Faculty of Computer Science

Teaching
Detailed information about these programmes are only available in french. The principal language of teaching is French

Bachelor Degree (three years)
Computer Science

Master Degree (two years)
Computer Science (choice among 3 streams of specialisation : information systems, information & software engineering, fundamentals of computing)

Advanced Master in

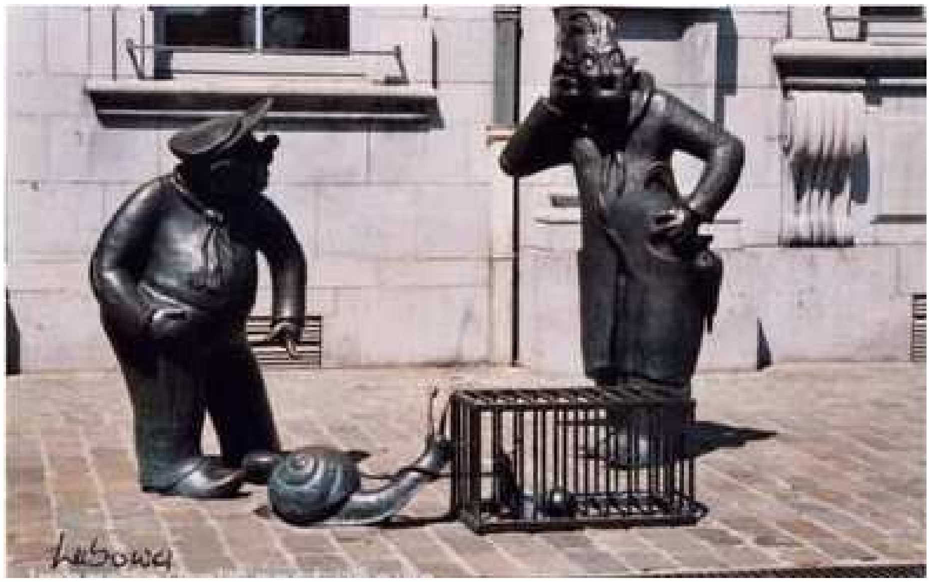
* Computer Science and Innovation
* Law of Information and Communication Technologies

Evening programme

* Bachelor in Computer Science
* Master in Computer Science (one year)

Doctoral Studies
Computer Science

Strengths
The Faculty of Informatics trains students for software development in its real context, that is, the private and public sectors. It emphasises the development of the main personal competences and qualities necessary for the specialist of XXIst century: mastery of the technology, independence, aptitude for team work and strong foreign language skills. During the second year, students do a work placement in one of the foreign organizations with which the Faculty maintains ongoing contacts.



" Djoseph and Françwès in the snail hunting "

According to the characters created by Jean Legrand.
Suzanne Godart's sculpture realized by the Studios of the ceramics,
in December, 2000
This work is Place d'Armes, Namur, Belgium

The snail is the symbol of Namur because
Namurois people, liking taking advantage of the life
and preferring to live peacefully and quietly,
are considered for "being slow".
(New proof of legendary Belgian auto-mockery)