8

Open-Source Tools for Data Mining in Social Science

Paško Konjevoda and Nikola Štambuk
Ruđer Bošković Institute
Croatia

1. Introduction

Data mining can be defined as the application of machine learning algorithms (Mitchell, 1997) for semiautomatic or automatic extraction of information from data stored in databases (Chakrabarti et al., 2009; Witten et al., 2011). The goal of data mining is to extract knowledge from the data set in human-understandable structures. In recent years data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and engineering. Textbooks on data analysis in social sciences largely deal with classical statistical methods, while data mining is usually mentioned only briefly (Moore, 2010). For example, in most of those books, the methods for solving the classification problem are separated from the analysis of data structure. Discriminant analysis and logistic regression are the most common classification methods found in social science textbooks, with PCA and factor analysis as representatives of the methods for analysis of data structures (Foster et al., 2006; Harlow, 2005; Tabachnick & Fidell, 2007). Some data mining techniques, like classification trees, allow simultaneous analysis of classification and data structure, and are far easier to interpret than discriminant, PCA or factor types of analyses (Rokach & Malmon, 2008; Zhang & Singer, 2010). We believe that it is not appropriate to separate statistics and data mining, because those methodologies are complementary to each other. Fortunately, several good textbooks that successfully combine statistics and data mining have been published in recent years (Larose, 2005, 2006; Myatt, 2007, 2009). Commercial applications for data mining are very expensive, and as such inaccessible to many institutions and students. One of the solutions is to use open-source programs that allow high-quality statistical and data mining software to be available to all students and researchers (Janert, 2011). The reason for using open-source software is not just commercial. Many open-source programs have such a quick development cycle that commercial software can not compete with. A classic example is R (R Development Core Team, 2011) which has more than 50,000 procedures for analysis and visualisation of data. Unable to follow the development of R, many commercial vendors of statistical software have added the option of calling R from their products. Some examples are SAS JMP, SPSS, STATISTICA, Genstat, etc. Therefore, we can say that R has become the gold standard for data analysis and visualization. Furthermore, the use of open-source software enables standardization and reproducibility of studies. This problem is particularly pronounced in numerical analysis, when the coefficients of the options within the program can significantly

affect the results and conclusions (Štambuk & Konjevoda, 2011). The purpose of this paper is to describe the open-source data mining programs that the authors have found useful in their work (Štambuk et al., 2007a, 2007b). Advantages and disadvantages of those programs are described, the web addresses where they can be found are listed, and the most relevant textbooks and manuals that describe how to work with these programs are cited.

2. R

R is an open-source programming language and software environment for statistical computing and data visualization (R Development Core Team, 2011). The R project was started in 1995 by a group of statisticians at the University of Auckland, and has continued to grow ever since. It is named after the first initials of the first two R authors (Robert Gentleman and Ross Ihaka). Academic researchers in various fields of applied statistics have adopted R for statistical software development and data analysis. It has become a *de facto* standard among statisticians for developing statistical software. There are a lot of niches in terms of R users, including: environmental statistics, econometrics, medical and public health applications, bioinformatics, and social sciences (Hilbe, 2010; Vinod, 2010) among others. Pre-compiled binary versions of R are provided for various operating systems on the address: http://cran.r-project.org/ (R, 2011). R is well documented, and free *R Journal* (2-3 issues per year) is also available on the same web address. Two main reasons for the world-wide success of R are its extensibility and superb data visualization (Fig. 1, Fig. 2). There are more than 2500 packages which enormously extend functionality of R. However, it is not easy for beginners (and even advanced users) to manage a huge number of procedures (over 50,000) that packages contain. A significant problem presents a command-line orientation of R. There are some graphical user interfaces under development. However, with the exception of Rattle (Fig. 3) none of them currently have the maturity and reachiness of menu-driven functionality associated with commercial statistical and data mining software.

2.1 Recommended R books

There are many books on R. Unfortunately, the vast majority of those books are too difficult for beginners, and it often happens that students completely withdraw from R after reading too advanced textbooks. Therefore, we cite the manuals which are simple and useful for the beginners.

2.1.1 Introductory books

1. *R in action: data analysis and graphics with R* (Kabacoff, 2011)**.** If you were to read only one book on R, it should be this one. It is a little masterpiece of pedagogy and clarity of writing, and deserves a detailed study.
2. *A beginner's guide to R* (Zuur et al., 2009) explains technical details of working with R. Therefore, it is oriented on users who already know statistics, but want to learn R.
3. *Using R for introductory statistics* (Verzani, 2005) teaches introductory statistics and R. It is well written and examples are nice. However, some students find it difficult to master both subjects at the same time.
4. *Statistics: an introduction using R* (Crowley, 2005) is similar to the previous book (3), but slightly more advanced.

5. *Business analytics for managers* (Jank, 2011) is a user-friendly introduction to regression analysis with R. It also explains some of advanced techniques, like multivariate visualization, regression trees, and nonparametric regression.

2.1.2 R Graphical User Interfaces (R GUI)

1. *R through Excel: A spreadsheet interface for statistics, data analysis, and graphics* (Heiberger & Neuwirth, 2009) describes RExcel (Microsoft Excel add-in). It allows access to the R from within Excel. Detailed information about RExcel available at the web page: http://rcom.univie.ac.at/ (RExcel, 2011).
2. *Getting started with Rstudio* (Verzani, 2011). This short book (75 pages) explains how to use RStudio, an integrated development environment (IDE) for R (Fig. 4). It includes a variety of features intended to make working with R more productive and straightforward. RStudio is available at the page: http://rstudio.org/ (RStudio, 2011).

2.1.3 Reference books about R

Reference books are must-have for any serious user of R. *The R book* (Crowley, 2007), *A handbook of statistical analysis using R* (Everitt, 2010) and *R cookbook* (Teetor, 2011) are recommended manuals of beginner-intermediate level.

2.1.4 Data mining books

1. *Data mining with R and Rattle* (Williams, 2011). This book describes Rattle (Fig. 3), a tab-based graphical user interface for data mining using R (Williams, 2011). The book is readable and easily written. Rattle runs under GNU/Linux, Macintosh OS X, and MS Windows operating systems. Rattle is freely available at: http://rattle.togaware.com/ (Rattle, 2011). Rattle is probably the most mature R GUI; simple, but powerful.
2. *Data mining with R: learning with case studies* (Torgo, 2011). The book is more advanced than *Data mining with R and Rattle*. It is based on examples from ecology, economy and bioinformatics.

2.1.5 Graphics with R

1. *R Graphics* (Murrell, 2006) is a detailed description of using R for production of publication quality graphs. It is clear and straightforward, but it is not recommended as the first book for beginners.
2. *Lattice: Multivariate data visualization with R* (Sarkar, 2008) describes the package *lattice* for multivariate data visualization (Fig. 1.).
3. *ggplot2: elegant graphics for data analysis* (Wickham, 2009) describes *ggplot2* package that simplifies many of the details of creating statistical graphics. Graphs produced with *ggplot2* are both beautiful and meaningful.

2.1.6 R programming

A first course in statistical programming with R (Brown & Murdoch, 2007) is entry-level introduction to programming. No previous knowledge of R is required, but the reader must know statistics and some calculus.

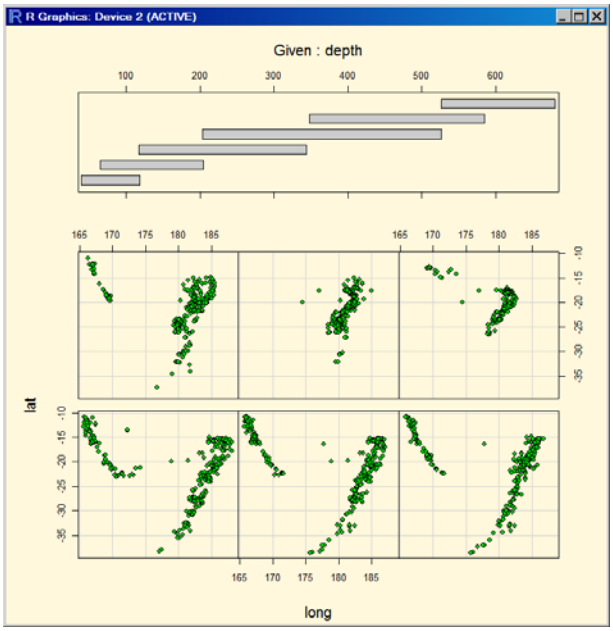

Fig. 1. An example of multivariate data visualization with R. The plot was made using *lattice* package (Sarkar, 2008).

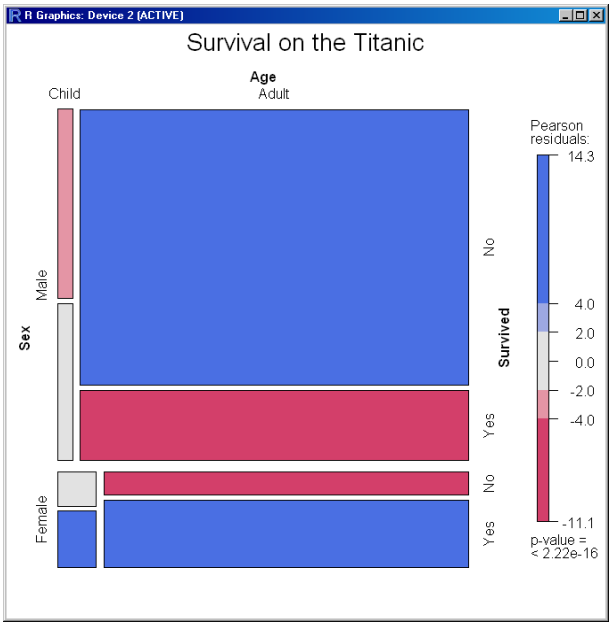

Fig. 2. An example of multivariate visualization of categorical data with R. The plot was made using *vcd* package.

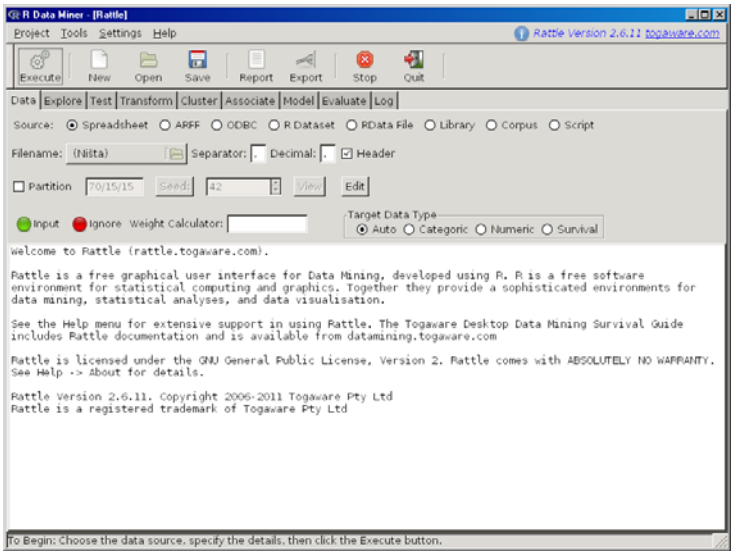

Fig. 3. Rattle is a tab-based graphical user interface for data mining using R (Williams, 2011). It runs under GNU/Linux, Macintosh OS X, and MS Windows operating systems.

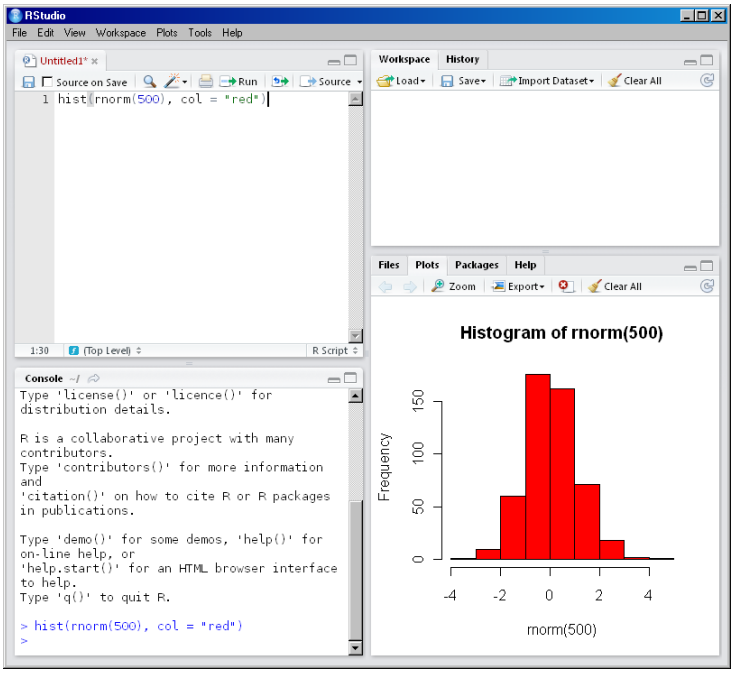

Fig. 4. RStudio is an integrated development environment (IDE) for R (Verzani, 2011). It includes a variety of features intended to make working with R more productive and straightforward.

3. WEKA

WEKA is a collection of machine learning algorithms for data mining tasks, freely available at: http://www.cs.waikato.ac.nz/ml/weka/ (WEKA, 2011). WEKA is Java based software, and works well under Windows, GNU/Linux and Mac Os X operating systems. It contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization (Fig. 5, Fig. 6, Fig. 7). The algorithms can either be applied directly to a dataset, or called from a user's Java code. It is also well-suited for developing new machine learning schemes.

A number of books use WEKA to illustrate principles of data mining (Larose, 2005, 2006; Myatt, 2007, 2009). Two books are especially useful for potential users:

1. *Data mining: practical machine learning tools and techniques* (Witten et al., 2010). This book, written by creators of WEKA, is now in its third edition, and is a standard reference on WEKA.
2. *Data mining techniques and applications: an introduction* (Du, 2010) is a short (about 300 pages) and readable introduction to data mining. All examples are explained using WEKA software.

WEKA is probably the most successful open source data mining software. It has inspired development of other programs with more elaborated graphical user interface and better visualization methods. Two of them are KNIME (Konstanz Information Miner) available from: http://www.knime.org/ (KNIME, 2011), and RapidMiner, available from: http://rapid-i.com/ (RapidMiner, 2011).

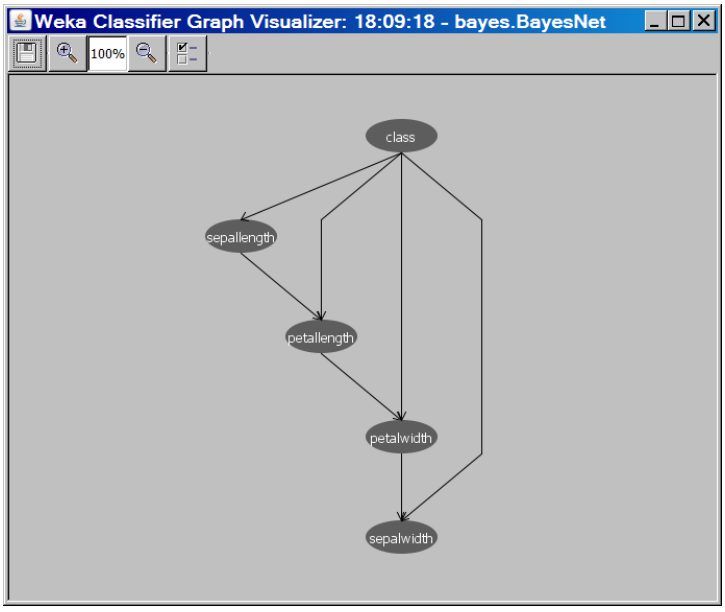

Fig. 5. WEKA implements various Bayesian network classifier learning algorithms.

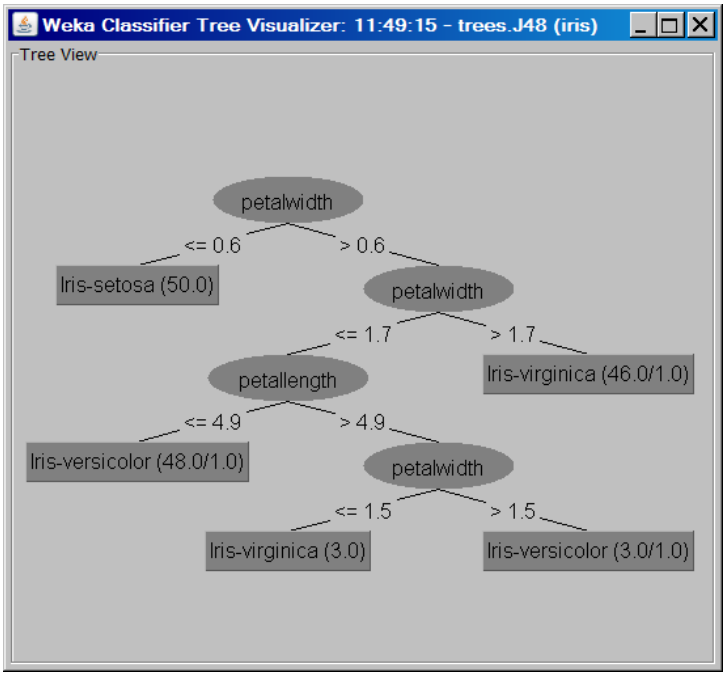

Fig. 6. An example of classification tree constructed with J48 classifier inside WEKA software.

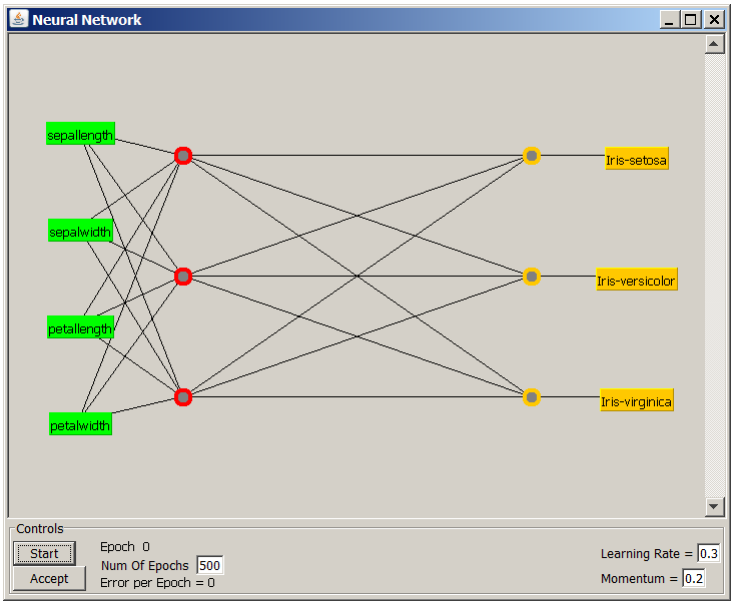

Fig. 7. A graphical representation of an artificial neural network (multilayer perceptron) implemented in WEKA.

4. TANAGRA

TANAGRA (Rakotomalala, 2005) is open-source data analysis software for academic and research purposes which combines data mining techniques with statistical learning (Hastie et al., 2009; Mitchell, 1993; Witten et al., 2010). The program and detailed tutorials are available at: http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html (TANAGRA, 2011).

TANAGRA is a successor of SIPINA project (SIPINA, 2011) which implements various supervised learning algorithms, especially an interactive and visual construction of decision trees. TANAGRA is more powerful, it contains many supervised learning techniques, but also other paradigms such as clustering, factorial analysis, parametric and nonparametric statistics, association rule, feature selection, etc (Fig. 8). TANAGRA works under Windows operating systems and GNU/Linux if Wine is used.

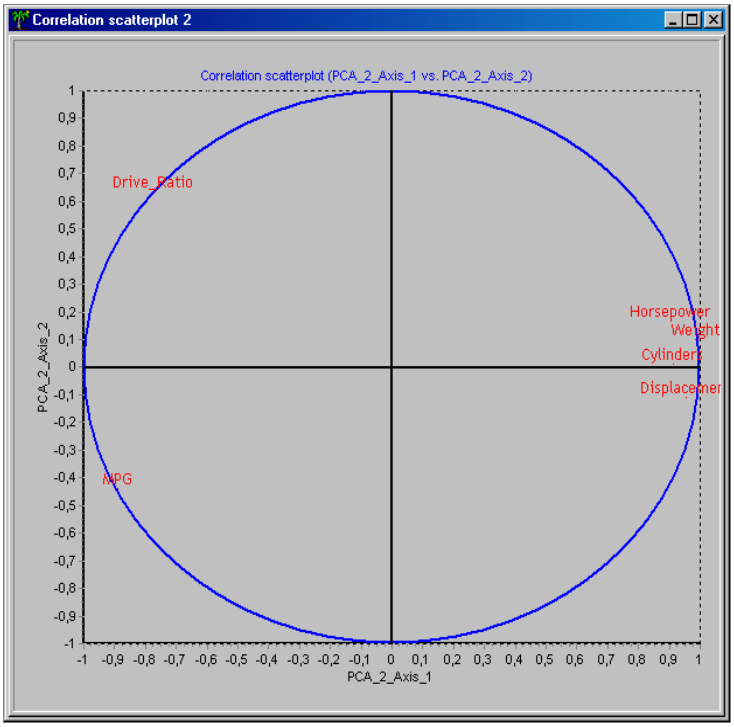

Fig. 8. Knowledge extraction with TANAGRA using combination of PCA and HAC

5. ORANGE

ORANGE is open-source data mining and visualization software. It is freely available from: http://orange.biolab.si/ (ORANGE, 2011). It enables design of data analysis process through user friendly visual programming (Fig. 9). ORANGE contains different visualizations, from scatter plots, bar charts, trees, to dendrograms, networks, and heatmaps (Fig. 10). Most major algorithms for data mining are represented. It works under Window, Mac Os X and GNU/Linux operating systems. Orange is integrated within Python, so this programming language can be used as a scripting language for repetitive tasks.

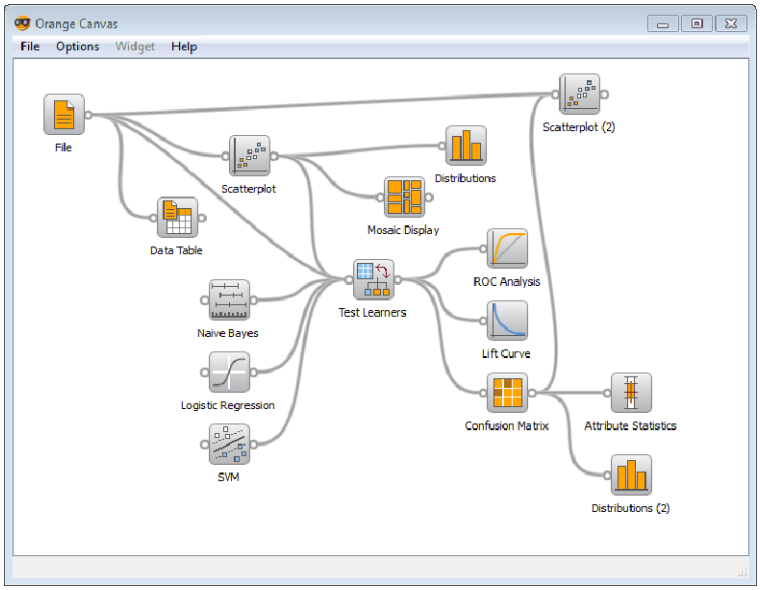

Fig. 9. ORANGE enables visual programming of data analysis process. It remembers user's choices, suggests most frequently used combinations, and intelligently chooses which communication channels to use.

6. PAST

The final application in our selection of data mining software is PAST (Hammer et al., 2001). PAST is freeware (not open-source), but we think it's a good replacement for R, especially for users who dislike a command line. Originally, it was aimed at paleontology, but now is also popular in many other fields. It includes common statistical, plotting and modeling functions.

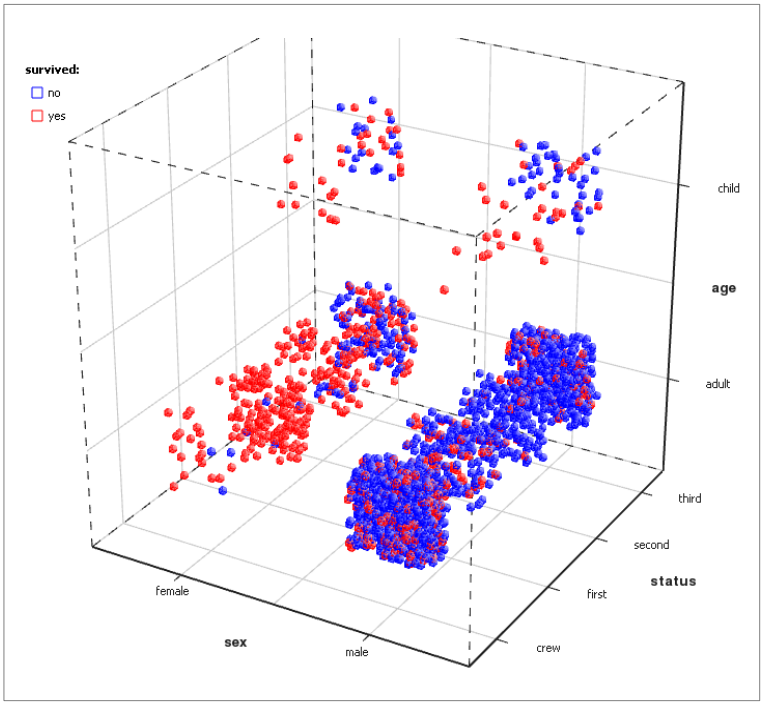

Fig. 10. ORANGE contains different visualization methods: from scatter plots, bar charts, trees, to dendrograms, networks, and heatmaps.

PAST is available on the following web address: http://folk.uio.no/ohammer/past/ (PAST, 2011). The manual accompanying PAST is detailed, and examples are well explained. The theoretical foundations of algorithms implemented in the PAST are described in the textbook *Paleontological data analysis* (Hammer & Harper, 2006).

PAST is a very user-friendly and feature-rich software (Fig. 11), especially remarkable for its large collection of cluster analysis algorithms (Everitt et al., 2011). It also implements a two-group permutation test, an excellent replacement for Hotelling's T2 test in situations where data do not follow a normal distribution (Good, 2011).

It works under Windows operating systems and GNU/Linux if Wine is used. PAST includes a simple, Pascal-like, scripting (programming) language that allows user to make his own algorithms within the PAST framework. The language has a full matrix support and a library of mathematical, statistical and user-interface functions.

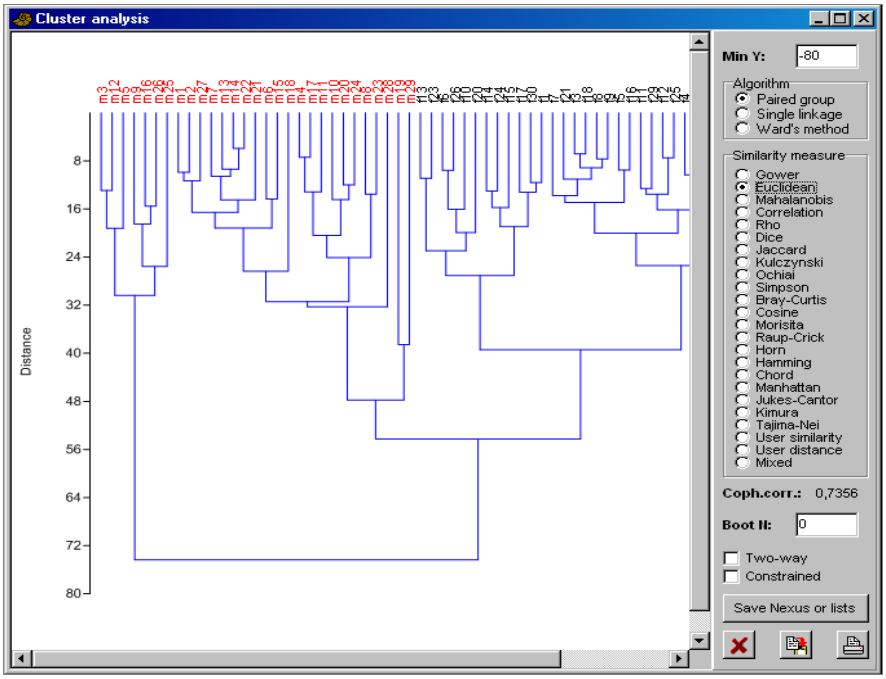

Fig. 11. PAST implements different unsupervised learning methods, including cluster analysis, neighbour joining, and *k*-means clustering.

7. Conclusion

Open-source data mining software is completely comparable to commercial programs according to the criteria of functionality and reliability. Moreover, some of them, such as R and Weka, have become the gold standard for industrial applications. According to Rexer's Annual Data Miner Survey in 2010, R has become the data mining tool used by more data miners (43%) than any other (Rexer Analytics, 2010). Therefore, it can be concluded that the use of open source software for educational purposes is completely justified, because students will be able to continue using the same programs after they graduate.

8. Acknowledgments

The support of the Croatian Ministry of Science, Education and Sports is gratefully acknowledged (grant No. 098-0982929-2524).

9. References

Brown, W.J. & Murdoch, D.J. (2007). *A first course in statistical programming with R*, Cambridge University Press, ISBN 978-0-521-69424-7, Cambridge

Chakrabarti, S.; Cox, E.; Frank, E.; Güting, R.H.; Han, J.; Jiang, X.; Kamber, M.; Lightstone, S.S.; Nadeau, T.P.; Neapolitan, R.E.; Pyle, D.; Refaat, M.; Schneider, M.; Teorey, T.J.

& Witten, I.H. (2009) *Data mining: know it all*, Elsevier, ISBN 978-0-12-374629-0, Amsterdam

Crowley, M.J. (2005) *Statistics: an introduction using R*, Wiley, ISBN 0-470-02297-3, Chichester

Crowley, M.J. (2007) *The R book*, Wiley, ISBN-13 978-0-470-51024-7, Chichester

Du, H. (2010). *Data mining techniques and applications: an introduction*, Cengage Learning EMEA, ISBN 978-1-84480-891-5, Andover

Everitt, B.S.; Landau, S.; Leese, M. & Stahl, D. (2011). *Cluster analysis*, Wiley, ISBN 978-0-470-74991-3, Hoboken

Everitt, B.S. & Horthorn, T. (2010). *A handbook of statistical analysis using R*, CRC Press, ISBN 978-1-4200-7933-3, Boca Raton

Foster, J; Barkus, E. & Yavorsky, C. (2006). Understanding and using advanced statistics, Sage, ISBN 1-4129-0013-1, London

Good, P.I. (2011). *Analyzing the large number of variables in biomedical and satellite imagery*, Wiley, ISBN 978-0-470-92714-4, Hoboken

Hammer, Ø., Harper, D.A.T., & Ryan, P.D. (2001). PAST: Paleontological statistics software package for education and data analysis. *Palaeontologia Electronica*, 4, 1, June 2001, pp. 1-9, ISSN 1094-8074.

Hammer, Ø., Harper, D.A.T. (2006). *Paleontological dana analysis*, Blackwell, ISBN 978—14051-1544-5, Malden

Harlow, L.L. (2005). The *essence of multivariate thinking: basic themes and methods*, Lawrence Erlbaum, ISBN 0-8058-3730-2, Mahwah

Hastie, T.; Tibshirani, R. & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*, Springer, ISBN 0387848576, Dordrecht

Heiberger, R.M. & Neuwirth, E. (2009). *R through Excel: A spreadsheet interface for statistics, data analysis, and graphics*, Springer, ISBN 978-1-4419-0051-7, Dordrecht

Hilbe, J. (2010). Statistical software for social sciences, In: *Statistics in the social sciences: current methodological developments*, Kolenikov, S., Steinley, D. & Thombs, L. (ed.), pp. 175-190, Wiley, ISBN 978-0-470-14874-7, Hoboken

Janert, P.K. (2011). *Data analysis with open source tools*, O'Reilly, ISBN 978-0-596-80235-6, Beijing

Jank, W. (2011) *Business analytics for managers*, Springer, ISBN 978-1-4614-0405-7, New York

Kabacoff, R.I. (2011) *R in action: data analysis and graphics with R*, Manning, ISBN: 9781935182399, Shelter Island

KNIME (Date of access 19 October, 2011), Available from: <http://www.knime.org/>

Larose, D.T. (2005). *Discovering knowledge in data: an introduction to data mining*, Wiley, ISBN 0-471-66657-2, Hoboken

Larose, D.T. (2006). *Data mining methods and models*, Wiley, ISBN 978-0-471-66656-1, Hoboken

Mitchell, T.M. (1997). *Machine learning*, McGraw-Hill, ISBN 0070428077, New York

Moore, D.S. (2010). *The basic practice of statistics*, W. H. Freeman and Company, ISBN 978-1-4292-0121-6, New York

Murrell, P. (2006) *R Graphics*, CRC Press, ISBN 978-1-58488-486-6, Boca Raton

Myatt, G.J. (2007). *Making sense of data: a practical guide to exploratory data analysis and data mining*, Wiley, ISBN 978-0-470-07471-8, Hoboken

Myatt, G.J. (2009). *Making sense of data II: A practical guide to data visualization, advanced data mining methods, and applications*, Wiley, ISBN 978-0-470-22280-5, Hoboken

ORANGE (Date of access 19 October, 2011), Available from: <http://orange.biolab.si/>
PAST (Date of access 19 October, 2011), Available at: <http://folk.uio.no/ohammer/past/>
R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.
R (Date of access 19 October, 2011), Available at: <: http://cran.r-project.org/>
Rakotomalala, R. (2005). TANAGRA: a free software for research and academic purposes, *Proceedings of European Grid Conference 2005*, RNTI-E-3, vol. 2, pp. 697-702, Amsterdam, February 2005
RapidMiner (Date of access October 19, 2011), Available from: <http://rapid-i.com/>
Rattle (Date of access October 19, 2011), Available from: <http://rattle.togaware.com/>
RExcel (Date of access October 19, 2011), Available from: <http://rcom.univie.ac.at/>
Rexer Analytics (Date of access 19 October, 2011), Available at: http://www.rexeranalytics.com/Data-Miner-Survey-Results-2010.html
Rokach, J. & Malmon, O. (2008). *Data mining with decision trees: theory and applications*, World scientific, ISBN 978-981-277-171-1, Singapore
RStudio (Date of access October 19, 2011), Available from: <http://rstudio.org/>
Sarkar, D. (2008) *Lattice: Multivariate data visualization with R*, Springer, ISBN 978-0-387-75968-5, New York
SIPINA (Date of access October 19, 2011), Available from: http://eric.univ-lyon2.fr/~ricco/sipina.html
Štambuk, N. & Konjevoda P. (2011). The role of independent test set in modeling of protein folding kinetics, In: *Software tools and algorithms for biological systems (Book Series: Advances in Experimental Medicine and Biology, Volume: 696),* Arabnia, H.R. & Tran, Q.N. (ed.), pp. 279-284, Springer, ISBN 978-1-4419-7045-9, New York
Štambuk, A; Štambuk, N. & Konjevoda, P. (2007) Application of Kohonen Self-Organizing Maps (SOM) based clustering for the assessment of religious motivation, *Proceedings of the ITI 2007, 29th Int. Conf. on Information Technology Interfaces*, Cavtat, Croatia, June 25-28, 2007, pp. 87-91
Štambuk, A.; Štambuk, N. & Konjevoda, P. (2007) Analysis of Hoge Religious Motivation Scale by Means of Combined HAC and PCA Methods, *Proceedings of the ITI 2007, 29th Int. Conf. on Information Technology Interfaces*, Cavtat, Croatia, June 25-28, 2007, pp. 197-201
Tabachnick, B.G. & Fidell, F.S. (2007). *Using multivariate statistics*, Pearson, ISBN 0-205-45938-2, Boston
TANAGRA (Date of access October 19, 2011), Available from: <http://eric.univ-lyon2.fr/~ricco/tanagra/en/tanagra.html>
Teetor, P. (2011). *R cookbook*, O'Reilly, ISBN 978-0-596-80915-7, Beijing
Torgo, L. (2011). *Data mining with R: learning with case studies*, CRC Press, ISBN 978-1-4398-1018-7, Boca Raton.
Verzani, J. (2005). *Using R for introductory statistics*, CRC press, ISBN 1-58488-4509, Boca Raton
Verzani, J. (2011). *Getting started with RStudio*, O'Reilly, ISBN 978-1-449-30903-9, Beijing
Vinod, H.D. (Ed.) (2010). *Advances in social science research using R*, Springer, ISBN 978-1-4419-1763-8, New York

WEKA (Date of access October 19, 2011), Available from: <http://www.cs.waikato.ac.nz/ml/weka/ (Weka, 2011)>

Wickham, H. (2009). *ggplot2: elegant graphics for data analysis*, Springer, ISBN 978-0-387-98140-6, Dordrecht

Williams, G. (2011). *Data mining with Rattle and R: the art of excavating data for knowledge discovery*, Springer, ISBN 978-1-4419-9889-7, New York

Witten, I.H.; Frank, E. & Hall, M.A. (2010). *Data mining: practical machine learning tools and techniques*, Morgan Kaufmann, ISBN 978-0-12-374856-0, Amsterdam

Zhang, H & Singer, B.H. (2010). *Recursive partitioning and applications*, Springer, ISBN 978-1-4419-6823-4, Dordrecht

Zuur, A.F.; Ieno, E.N. & Meesters E.H.W.G. (2009). *A beginner's guide to R*, Springer, ISBN 978-0-387-93836-3, Dordrecht