Innovations in Soft Data Analysis

Mika Sato-Ilic[1] and Lakhmi C. Jain[2]

[1] Faculty of Systems and Information Engineering, University of Tsukuba,
Tsukuba, Ibaraki 305-8573, Japan
mika@risk.tsukuba.ac.jp
[2] Knowledge-Based Intelligent Engineering Systems Centre,
University of South Australia,
Adelaide, SA 5095, Australia
Lakhmi.Jain@unisa.edu.au

Abstract. The amount of data is growing at an exponential rate. We are faced with a challenge to analyze, process and extract useful information from the vast amount of data. Traditional data analysis techniques have contributed immensely in the area of data analysis but we believe that the soft data analysis techniques, based on soft computing techniques, can be complementary and can process complicated data sets. This paper provides an introduction to the soft data analysis paradigms. It summarizes the successful and possible applications of the soft computing analysis paradigms. The merits and demerits of these paradigms are included. A number of resources available are listed and the future vision is discussed. This paper also provides a brief summary of the papers included in the session on "Innovation in Soft Data Analysis".

1 Introduction

Soft data analysis (SDA) is based on soft computing which is a consortium of methodologies including fuzzy logic, neural networks, probabilistic reasoning, genetic algorithms, and chaotic systems which complement each other.

Soft computing reflects the pervasiveness of imprecision and uncertainty which exists in the real world. On the other hand, hard computing does not reflect this imprecision and uncertainty. The guiding principal of soft computing is to exploit the tolerance of imprecision and uncertainty in order to achieve tractability, robustness, and low solution cost.

Recently, in the area of data analysis, many new methods have been proposed. One reason for this is that traditional data analysis does not adequately reflect the imprecision and uncertainty of real world data. For instance, analyses for uncertainty data including interval-valued data, modal data, functional data, categorical data, and fuzzy data have been proposed. In the area of neuroscience, data which have a larger number of variables than the number of objects are also important concerns in multivariable data analysis. Conventional multivariate analyses can not treat such data. Huge amounts of data are also a problem and data mining is placing an increased emphasis on revealing the latent structure existing in such data. For example, in the analyses of point of sale (POS) data in

B. Gabrys, R.J. Howlett, and L.C. Jain (Eds.): KES 2006, Part II, LNAI 4252, pp. 110–114, 2006.

the marketing area the key challenge is interfacing the emulator with the POS data traffic. In another area Deoxyribonucleic Acid (DNA) data in information biology also has a similar problem resulting from the amount of data.

The second reason is the increase of interest and recognition for the specific features of data which must be treated as non-ignorable features of the data. Several types of data have asymmetric features. Input and output data in an input-output table in the economic area or in human relationship data such as in psychometrics are typical examples. Traditional techniques treat the asymmetry feature as noise affecting the symmetric structure of the data. However, based on the premise that the systematic asymmetry is of value and the information contained should be captured, new methods have been proposed.

The third reason is that with the advance of computer technology and the resulting expansion of computer ability, improved visualization techniques of data, results of data analysis, and the features of data analysis have flourished. The features of complex data analysis involving imprecision and uncertainty have witnessed a crystallization of the exploratory visualization techniques for data.

The fourth reason is the limited precision in the data. In order to obtain precision from real data, we need to make many assumptions about the latent data structures. Under the many assumptions necessary for representing real data structures, even if we obtain a precise result for the data, since no one can know the real data structure, we can not prove that the assumptions actually represent the real data structure.

From this background, SDA is constituted as a consortium of data analyses aimed at reflecting imprecision and uncertainty existing in real data. Fuzzy multivariate data analysis is a popular current methodology of SDA. Examples include: fuzzy clustering, fuzzy regression analysis, fuzzy principal component analysis, fuzzy quantification analyses (types I, II, III, and IV). Hybrid methods over neural networks, genetic algorithms, support vector machines, and fuzzy data analysis are also types of SDA. The role of SDA techniques is growing and their potential is fully accepted in real world data analysis. In SDA, functional data analysis, non-linear generalized models, and symbolic data analysis, are also new and powerful methods that exhibit further progress with substantial reliance on the traditional statistical data analysis.

Several successful SDA applications have been presented. The use of grey-tone pictures in image classification is a typical example. X-ray pictures, images by satellite and particles registered by optical devices are examples of objective data. Although the data are represented by pixels as digital values, the real data is continuous, so the analyses considering the imprecision and uncertainty must be adaptable. Fuzzy clustering for this data achieves suitable representation of the result including the uncertainty of boundaries over clusters. The methods show the robustness and low solution cost.

Applications to temperature records, growth curves, time series records, and hand writing samples have also proved to be successful. Such as analog sources of data performed continuously in time or space. These data and the results are represented by functions or trajectory in hyper spaces, fuzzy data analysis or

functional data analysis, and fuzzy weighted regression analysis are suitable for these data.

The application of discrimination to neural networks and support vector machines are well known. Many applications have been presented in the cognitive neuroscience to develop the mechanisms underlying object vision.

The innovative SDA techniques are based on the idea that the universe may be chosen as a space of functions which can include extremely large dimensions. Techniques which use fuzzy data, functional data, symbolic data, kernel method, and spherics are typical examples.

2 Soft Data Analysis Paradigms

Statistical data analysis assumes "systematic" uncertainty for observational data. The amount of systemization is represented by statistical distribution.

The concept of exploratory data analysis [1] where the emphasize is on the idea that real data structure is on the multiple aspects of the data and that a model (or a structure) is not assumed have been prospered. For exploratory data analysis, the concept of statistical science was an ideal solution.

However, the essence of observed data is not always based on "systematic" uncertainty. The necessity for analysis allowing for the most comprehensive uncertainty have been increased. As an ideal solution for analysis capturing unique features of data with the intention of discovering uncertainty and a set of robust and modern methods, SDA has been proposed.

3 Resources on SDA

We introduce several of the latest references and software support for SDA.

3.1 Literatures for SDA

- H. Bandemer and W. Nather, *Fuzzy Data Analysis*, Kluwer Academic Publishers, 1992.
- J.C. Bezdek, J. Keller, R. Krisnapuram, and N.R. Pal, *Fuzzy Models and Algorithms for Pattern Recognition and Image Processing*, Kluwer Academic Publishers, 1999.
- H.H. Bock and E. Diday eds., *Analysis of Symbolic Data*, Springer, 2000.
- C. Brunsdon, S. Fotheringham, and M. Charlton, Geographically Weighted Regression-Modelling Spatial Non-Stationarity, *Journal of the Royal Statistical Society*, Vol. 47, Part 3, pp. 431-443, 1998.
- N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*, Cambridge University Press, 2000.
- A.J. Dobson, *An Introduction to Generalized Linear Models*, Chapman and Hall, 1990.
- A.S. Fotheringham, C. Brunsdon, and M. Charlton, *Geographically Weighted Regression*, Wiley, 2002.

- A. Ghosh and L.C. Jain eds., *Evolutionary Computation in Data Mining*, Springer, 2005.
- T.J. Hastie and R.J. Tibshirani, *Generalized Additive Models*, Chapman & Hall, 1990.
- F. Höppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*, Wiley, 1999.
- H. Ishibuchi, T. Nakashima, and M. Nii, *Classification and Modeling with Linguistic Information Granules*, Springer, 2005.
- L.C. Jain ed., *Soft Computing Techniques in Knowledge-Based Intelligent Engineering Systems*, Springer, 1997.
- H.A.L. Kiers, J.-P. Rasson, P.J.F. Groenen, and M. Schader eds., *Data Analysis, Classification, and Related Methods*, Springer, 2000.
- S. Miyamoto, *Fuzzy Sets in Information Retrieval and Cluster Analysis*, Cluver Academic Publishers, 1990.
- N. Pal and L. Jain, *Advanced Techniques in Knowledge Discovery and Data Mining*, Springer, 2005.
- W. Pedrycz, *Knowledge-Based Clustering*, Wiley, 2005.
- J.O. Ramsay and B.W. Silverman, *Applied Functional Data Analysis*, Springer, 2002.
- J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer, 2nd ed., 2005.
- Y. Sato, An Analysis of Sociometric Data by MDS in Minkowski Space, *Statistical Theory and Data Analysis II*, North Holland, pp. 385-396, 1988.
- M. Sato, Y. Sato, and L.C. Jain, *Fuzzy Clustering Models and Applications*, Springer, 1997.
- M. Sato-Ilic and L.C. Jain, *Innovations in Fuzzy Clustering*, Springer, 2006.
- B. Scholkopf and A.J. Smola, *Learning with Kernels*, The MIT Press, 2002.
- H. Tanaka and J. Watada, Possibilistic Linear Systems and their Application to the Linear Regression Model, *Fuzzy Sets and Systems*, Vol. 27, pp. 275-289, 1988.
- H. Tanaka and P. Guo, Possibilistic Data Analysis for Operations Research, Springer, 1999.
- Y. Yabuuchi and J. Watada, Fuzzy Principal Component Analysis and Its Application, *Journal of Biomedical Fuzzy Systems Association*, Vol. 3, No. 1, pp. 83-92, 1997.

3.2 Software Supports

- The R Project for Statistical Computing, http://www.r-project.org/
- MATLAB Fuzzy Logic Toolbox, http://www.mathworks.com/products/fuzzylogic/
- GWR for Geographically Weighted Regression, A.S. Fotheringham, C. Brunsdon, and M. Charlton: *Geographically Weighted Regression*, Wiley, 2002.
- SODAS (Symbolic Official Data Analysis System) for Symbolic Data Analysis, H.H. Bock and E. Diday eds.: *Analysis of Symbolic Data*, Springer, 2000.

4 Papers Included in This Session

Five papers are to be presented during this session. The first by Mr. Inokuchi and Prof. Miyamoto on a nonparametric fisher kernel using fuzzy clustering. Using a distribution of the degree of belongingness of objects to the clusters obtained by fuzzy clustering, a nonparametric fisher kernel is applied to the classification. As a new application of fuzzy clustering for fisher kernel, they present quite novel research.

The second by Prof. Ishibuchi, Dr. Nojima, and Mr. Kuwajima on a new methodology to find fuzzy classification systems. They demonstrate a high capability to find fuzzy classification systems with large interpretability using multiobjective rule selection. In reasoning for real world data, this technique is very progressive.

The third by Prof. Klawonn and Prof. Höppner on a new classification technique to create clusters of approximately the same size. Not only the similarity of objects in each cluster, but also the similarity of the sizes of clusters is considered. The idea is quite novel. For the use of real data application, this technique has enormous potential to solve a wide range of problems.

The fourth by Prof. Sato on a new method for clustering of mixed data using spherical representation. Several different transformations are introduced through the use of a probabilistic distance in probabilistic space, a distance in Riemannian space, and a map from a plane to the surface of the sphere. As a consortium of the several logical techniques involving over the probabilistic theory, differential geometry, spherics, and visual inspection for the representation for directional data, this research is quite novel.

The last paper by Prof. Sato-Ilic and Mr. Kuwata on a self-organized (dis)-similarity considering two spaces. One is the (dis)-similarity of classification structures of a pair of objects and the other is (dis)-similarity of a pair of objects. This is an attempt to discover "how to combine two spaces" which are a space of classification structures obtained as a result of fuzzy clustering and a space of observed data. Our final goal is to investigate the relation between the two spaces through a metric space defined (dis)-similarity.

5 Conclusion

This session provides an overview of SDA. Since SDA is based on traditional data analysis, the overview covers a wide range of topics. This overview of SDA described here is only the tip of an iceberg. However, most data analysts are aware of the present limitations for analyzing collected data. SDA provides a solution for these limitations.

Reference

1. Tukey, J.W.: Exploratory Data Analysis. Addison-Wesley Publishing. (1977)