Soft Data Analysis based on Cluster Scaling

Mika Sato-Ilic

University of Tsukuba

Abstract

There is a growing need to analyze today's vast and complex societal data; however, conventional data analysis that is dependent on statistical methods cannot deal with the frequently complex data types that make up this data. As early as 2000, the following six challenges were reported as important future challenges in the core area statistical research in the 21st century. The six challenges pointed out are 1) Scales of Data, 2) Data Reduction and Compression, 3) Machine Learning and Neural Networks, 4) Multivariate Analysis for Large p, Small n (High Dimension Low Sample Size Data), 5) Bayes and Biased Estimation, and 6) Middle Ground between Proof and Computational Experiment.

Soft data analysis which is soft computing based multivariate analysis is the core area in which to combine conventional statistical methods and machine learning or data mining methods, and has a strong capability to solve the statistical challenges in the 21st century. In soft data analysis, we have developed cluster-scaled models which use the obtained cluster as the latent scale for explaining data. While the original scale does not have the capacity to work as the scale for complex data, a scale that is extracted from the data itself will have the capability to deal with the vast and complex data.

This presentation outlines the problems and challenging issues of statistical data analysis caused by the new vast and complex data, how our cluster-scaled models are related with these issues, and how the models solve the problems with some applications.