Title	A Study on data integration based on its structure and characters(Digest_要約)
Author(s)	岡﨑, 威生
Citation	
Issue Date	2014-09-10
URL	http://hdl.handle.net/20.500.12000/29679
Rights	

（様式第３号）

論　文　要　旨

論　文　題　目

A Study on data integration based on its structure and characters

（構造と特徴によるデータ再構成に関する研究）

Data science is a set of complex academic domain that is the integrated technical knowledge concept of mathematics, probability models, statistics, pattern recognition, machine learning, visualization and database, so on. In the 2000's, it has been generalized as advances in computing with data. Because of the development of smart devices, web, the mobile and social media, the underlying data has come to be produced in large quantities. It is a Big data. However, Data science is not requesting the Big data at all times. Classically it was starting from the inference with the small sample, it is a result that has been scaled up by the development of database and data warehousing techniques, or the importance of the data science being recognized. Data mining is the flourishing trend of Data science. It can be said to think from the point of view of data analysis, it is a change to the hypothesis finding type analysis from the conventional hypothesis confirming type. In other words, the small sample problem has been transitioned to Big data, into a composite data problem. However, a large amount of data does not necessarily coincide with the height of quality. When the start as a big data, an attempt was made to hypothesis discovered by data mining, how objectivity and validity of the hypothesis that has been extracted are guaranteed ? Originally the discovered hypothesis should be verified again. At the time of verification, rather than big data itself that was used at the time of the discovery, variable selection suitable for verification or performing data conversion is important. That is, the data reconstruction.

The purpose of this study is to propose procedures of data reconstruction for the improvement of the statistical analysis accuracy at the hypothesis confirmation. Especially 3 topics were focused such as (1) Data pooling (2) Outlier finding (3) Data normalization. On data pooling, I proposed the TE type estimator that is an estimate for the population mean and variance in the normal population, and evaluated the statistical features. I derived the probability density functions, expectations and MSE of the estimator, proved the unbiasedness of the estimator. MAE that was the evaluation index by L_1-measure was proposed. For the non-parametric statistics median, I derived the similar statistical features. To decide on the optimal significance level for the preliminary test, relative risk criterion and the mini max criterion were considered. As the pooling data issue in the latent structural model, factor analysis was taken up and I proposed a parameter estimation procedure for the purpose of the observation variable supplement. I considered the supplement possibility from the view of non-discriminating conditions, carried out numerical evaluations by benchmark data. For outlier finding, I proposed a two-dimensional arrangement technique that could correspond to the asymmetric relationship. Decomposing the symmetric and skewness part of the target data, and I visualized in two dimensions, respectively. The improved point for skewness part representation enable us to detect outliers effectively. Also in the symmetrical part representation, I improved to increase the robustness of the Powell method and the BFGS method. To the DNA microarray data that is the foundation genomic data, I modeled comprehensively bias caused by experimental methods, environment and the observation method. I formulated an integrated representation method of application and the order in each bias correction method on it. Then, I proposed a method to select the optimal normalization procedure by BIC comparison. Experiments on mouse, yeast, and Homo sapience showed effectiveness.

It is expected that the technology for data acquisition such as sensors, networks and smart devices evolves dramatically. It is an order of magnitude faster speed, compared to the processing strategy of the data obtained.In that situation, these four research results will contribute to the higher reliability of modern data science or data analysis.

氏　名　　岡　﨑　威　生