World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:12, No:09, 2018

Dissimilarity Measure for General Histogram Data and Its Application to Hierarchical Clustering

Authors : K. Umbleja, M. Ichino

Abstract : Symbolic data mining has been developed to analyze data in very large datasets. It is also useful in cases when entry specific details should remain hidden. Symbolic data mining is quickly gaining popularity as datasets in need of analyzing are becoming ever larger. One type of such symbolic data is a histogram, which enables to save huge amounts of information into a single variable with high-level of granularity. Other types of symbolic data can also be described in histograms, therefore making histogram a very important and general symbolic data type - a method developed for histograms - can also be applied to other types of symbolic data. Due to its complex structure, analyzing histograms is complicated. This paper proposes a method, which allows to compare two histogram-valued variables and therefore find a dissimilarity between two histograms. Proposed method uses the Ichino-Yaguchi dissimilarity measure for mixed feature-type data analysis as a base and develops a dissimilarity measure specifically for histogram data, which allows to compare histograms with different number of bins and bin widths (so called general histogram). Proposed dissimilarity measure is then used as a measure for clustering. Furthermore, linkage method based on weighted averages is proposed with the concept of cluster compactness to measure the quality of clustering. The method is then validated with application on real datasets. As a result, the proposed dissimilarity measure is found producing adequate and comparable results with general histograms without the loss of detail or need to transform the data.

Keywords : dissimilarity measure, hierarchical clustering, histograms, symbolic data analysis
Conference Title : ICDMA 2018 : International Conference on Data Mining and Applications
Conference Location : Tokyo, Japan
Conference Dates : September 10-11, 2018

Open Science Index, Computer and Information Engineering Vol:12, No:09, 2018 publications.waset.org/abstracts/92018.pdf