STATISTICAL ANALYSIS AND DATA MINING

Original Article

A Bayesian criterion for cluster stability

Hoyt Koepke, Bertrand Clarke

First published: 06 February 2013
https://doi.org/10.1002/sam.11176

Abstract

We present a technique for evaluating and comparing how clusterings reveal structure inherent in the data set. Our technique is based on a criterion evaluating how much point - to - cluster distances may be perturbed without affecting the membership of the points. Although similar to some existing perturbation methods, our approach distinguishes itself in five ways. First, the strength of the perturbations is indexed by a prior distribution controlling how close to boundary regions a point may be before it is considered unstable. Second, our approach is exact in that we integrate over all the perturbations; in practice, this can be done efficiently for well - chosen prior distributions. Third, we provide a rigorous theoretical treatment of the approach, showing that it is consistent for estimating the correct number of clusters. Fourth, it yields a detailed picture of the behavior and structure of the clustering. Finally, it is computationally tractable and easy to use, requiring only a point - to - cluster distance matrix as input. In a simulation study, we show that it outperforms several existing methods in terms of recovering the correct number of clusters. We also illustrate the technique in three real data sets. © 2013 Wiley Periodicals, Inc. Statistical Analysis and Data Mining, 2013