The "Best K" for Entropy-based Categorical Data Clustering

Keke Chen, Wright State University - Main Campus
Ling Liu

Document Type
Presentation
Publication Date
2005

Abstract
With the growing demand on cluster analysis for categorical data, a handful of categorical clustering algorithms have been developed. Surprisingly, to our knowledge, none has satisfactorily addressed the important problem for categorical clustering – how can we determine the best K number of clusters for a categorical dataset? Since categorical data does not have the inherent distance function as the similarity measure, traditional cluster validation techniques based on the geometry shape and density distribution cannot be applied to answer this question. In this paper, we investigate the entropy property of the categorical data and propose a BkPlot method for determining a set of candidate "best Ks". This method is implemented with a hierarchical clustering algorithm ACE. The experimental results show that our approach can effectively identify the significant clustering structures.

Comments
This paper was presented at the Scientific and Statistical Database Management Conference (SSDBM05), Santa Barbara, CA, June 2005.

Repository Citation
It seems a good place to start is using hierarchical clustering (since I don't know k a-priori) maybe following the recommendations here: https://stackoverflow.com/questions/23943391/how-to-cluster-users-based-on-tags. I have a few questions about this. First, is this a good approach? And is Mahout the best tool for this? Any obvious ways to simplify? And finally, how can I combine the tag clustering approach with the other more structured data in the rows? There are tons of good answers already for clustering categorical data on CrossValidated – smci Nov 28 '16 at 14:54. why don't you point to one you think is exemplary – Alexey Grigorev Nov 28 '16 at 15:32.