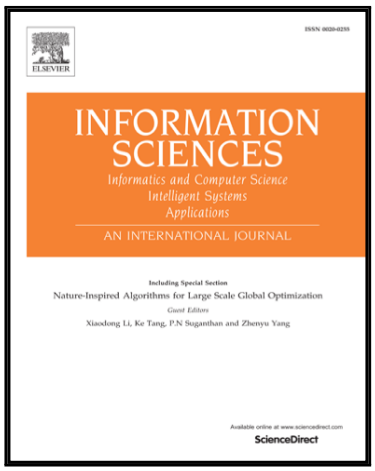

Curvature-based method for determining the number of clusters

Zhang Yaqian, Jacek Mańdziuk, Quek Hiok Chai, Goh Wooi Boon

PII: S0020-0255(16)31301-9
DOI: 10.1016/j.ins.2017.05.024
Reference: INS 12894

To appear in: *Information Sciences*

Received date: 15 October 2016
Revised date: 11 April 2017
Accepted date: 16 May 2017

Please cite this article as: Zhang Yaqian, Jacek Mańdziuk, Quek Hiok Chai, Goh Wooi Boon, Curvature-based method for determining the number of clusters, *Information Sciences* (2017), doi: 10.1016/j.ins.2017.05.024

Curvature-based method for determining the number of clusters

Zhang Yaqian[a], Jacek Mańdziuk[a,b,*], Quek Hiok Chai[a], Goh Wooi Boon[a]

[a]*School of Computer Science and Engineering, Nanyang Technological University, Singapore*
[b]*Faculty of Mathematics and Information Science, Warsaw University of Technology, Warsaw, Poland*

Abstract

Determining the number of clusters is one of the research questions attracting considerable interests in recent years. Majority of the existing methods require parametric assumptions and substantiated computations. In this paper we propose a simple yet powerful method for determining the number of clusters based on curvature. Our technique is computationally efficient and straightforward to implement. We compare our method with 6 other approaches on a wide range of simulated and real-world datasets. Theoretical motivation underlying the proposed method is also presented.

Keywords:
k-Means clustering, Number of clusters, Cluster analysis, Gap statistic, Hartigan's rule

1. Introduction

Many clustering algorithms suffer from the limitation that the number of clusters has to be specified by a human user [21][28][34]. However, as Salvador and Chan pointed out [27], in most cases, users do not have sufficient domain knowledge or prior information to select the correct number of clusters to return. Consequently, there have been a number of approaches published in the literature for choosing the right k after multiple runs of k-Means [4][13][16], being the most popular machine learning (ML) clustering algorithm. The notion of a *cluster* is not uniquely-defined as it heavily

*Corresponding author: Jacek Mańdziuk, J.Mandziuk@mini.pw.edu.pl