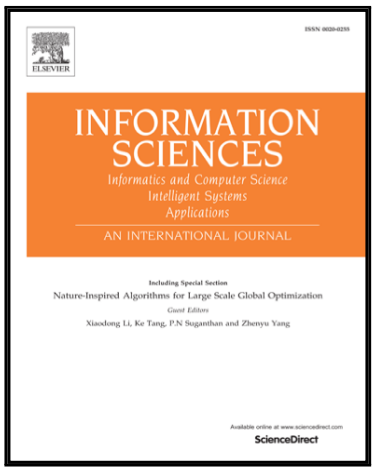


# I-nice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres

Md Abdul Masud, Joshua Zhexue Huang, Chenghao Wei, Jikui Wang, Imran Khan, Ming Zhong

PII: S0020-0255(17)30113-5
DOI: 10.1016/j.ins.2018.07.034
Reference: INS 13801

To appear in: *Information Sciences*

Received date: 7 January 2017
Revised date: 4 July 2018
Accepted date: 11 July 2018

Please cite this article as: Md Abdul Masud, Joshua Zhexue Huang, Chenghao Wei, Jikui Wang, Imran Khan, Ming Zhong, I-nice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.07.034

# I-nice: A New Approach for Identifying the Number of Clusters and Initial Cluster Centres

**Md Abdul Masud[a], Joshua Zhexue Huang*[a], Chenghao Wei[a], Jikui Wang[a], Imran Khan[b], Ming Zhong[a]**

*[a]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China.*
*[b]Shenzhen Key Laboratory of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China.*

**Abstract**

This paper proposes I-nice, which is a new method for automatically identifying the number of clusters and selecting the initial cluster centres in data. The method mimics a human being in observing peaks of mountains in field observation. The clusters in a dataset are considered as the hills in a field terrain. The distribution of distances between the observation point and the objects is computed. The distance distribution is modelled by a set of Gamma mixture models (GMMs), which are solved with the expectation-maximization (EM) algorithm. The best-fitted model is selected with an Akaike information criterion variant (AICc). In the I-niceSO algorithm, the number of components in the model is taken as the number of clusters, and the objects in each component are analysed with the $k$-nearest-neighbour method to find the initial cluster centres. For complex data with many clusters, we propose the I-niceMO algorithm, which combines the results of multiple observation points. Experimental results show that the two algorithms significantly outperformed two state-of-the-art methods (Elbow and Silhouette) in identifying the correct number of clusters in data. The results also show that I-niceMO improved the clustering accuracy and efficiency of the $k$-means clustering process.

*Keywords:* Clustering Algorithm, Initial Cluster Centres, Number of Clusters.

## 1. Introduction

Clustering is one of the key techniques in data analysis. It is the process of dividing the data of objects into a set of clusters in which the objects in the same clusters are close to each other according to a similarity measure, whereas the objects in different clusters are far from each other. One problem in cluster analysis is that the number of clusters in the data to be analysed must be known in advance because many clustering algorithms require the number of clusters as an input parameter to run the algorithms. However, the number of clusters that exist in real data is usually unknown. Therefore, a number is often guessed in practical cluster analysis, which often results in unsatisfactory results. Although several methods for estimating the number of clusters in data have been developed [13, 45, 42, 21, 47], they either produce incorrect results or are difficult to use in real applications. Therefore, finding the correct number of clusters from real data remains a classical problem in cluster analysis. It is also an active research topic.

In this paper, we propose an innovative approach to identifying the number of clusters in high-dimensional data. We consider a dataset as a terrain in which clusters are hills. We assign an observer to the terrain to observe and count the peaks of hills, which correspond to the dense regions of clusters and reflect the number of clusters in the data. Fig. 1(a) shows an example of the observation process in which three hills are situated at different distances from two

*Corresponding author
*Email addresses:* masud@szu.edu.cn (Md Abdul Masud), zx.huang@szu.edu.cn (Joshua Zhexue Huang*), chenghao.wei@yahoo.com (Chenghao Wei), wjkweb@163.com (Jikui Wang), imran.khan@sustc.edu.cn (Imran Khan), mingz@szu.edu.cn (Ming Zhong)