conferenceseries.com SciTechnol

Boris Mirkin, J Comput Eng Inf Technol, 6:5
DOI: 10.4172/2324-9307-C1-014

4th International Conference on

BIG DATA ANALYSIS AND DATA MINING

September 07-08, 2017 | Paris, France

Boris Mirkin[1,2]

[1]Higher School of Economics, Russia

[2]Birkbeck University of London, UK

A complementary square-error clustering criterion and initialization of K-means

Clustering is a set of major data analysis techniques. The square-error clustering criterion underlies most popular clustering methods including K-means partitioning and Ward agglomeration. For the K-means, the square-error criterion to be minimized is the sum of squared Euclidean distances from all the objects to their respective cluster centers/means, W(S,c), where S is the sought partition of the set of objects and c is the set of within-cluster means. The method's popularity stems from the simplicity of computation and interpretation. Yet there is a catch: the user is to specify both the number of clusters and the initial locations of cluster centers, which can be of an issue sometimes. To tackle the problem, the current author proposes using the complementary criterion. It is not difficult to prove that there is a complementary criterion, B(S,c), to be maximized, such that W(S,c)+B(S,c)=T where T is the data scatter. The complementary criterion B(S,c) is the sum of individual cluster contributions, equal each to the product of the cluster's cardinality and the squared Euclidean distance from the cluster's center to 0. Therefore, the complementary criterion leads to a set of anomalous clusters, which can be found either one-by-one or in parallel. Our experiments show that methods emerging in this perspective are competitive, and frequently superior, to other initialization methods.

Biography

Boris Mirkin holds a PhD in Computer Science and DSc in Systems Engineering from Institute for Systems Studies of the Academy of Sciences and State Committee of Science and Technology of the USSR. He published a dozen monographs and a hundred refereed papers. He develops methods for clustering and interpretation of complex data within the data recovery perspective. Currently, these approaches are being extended to automation of text analysis including use of hierarchical ontologies.

mirkin@dcs.bbk.ac.uk

Notes: