Abstract

Clustering high dimensional data becomes difficult due to the increasing sparsity of such data. One of the inherent properties of high dimensional data is hubness phenomenon, which is used for clustering such data. Hubness is the tendency of high-dimensional data to contain points (hubs) that occurs frequently in k-nearest neighbor lists of other data points. The k-nearest-neighbor lists are used to measure the hubness score of each data point. The simple hub based clustering algorithms detect only hyperspherical clusters in the high dimensional
dataset. But the real time high dimensional dataset contains more number of arbitrary shaped clusters. To improve the performance of clustering, a new algorithm is proposed which is based on the combination of kernel mapping and hubness phenomenon. The proposed algorithm detects arbitrary shaped clusters in the dataset and also improves the performance of clustering by minimizing the intra-cluster distance and maximizing the inter-cluster distance which improves the cluster quality.

References

- I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts", in Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2004, pp. 551–556.
- E. Plaka and L. E. Kavraki (2007), "Distributed computation of the Knn graph for large high dimensional point sets", Journal of Parallel and Distributed Computing, 67(3): 346-

Index Terms
Keywords
High Dimensional Data  Hubness Phenomenon  Kernel Mapping  And K-nearest Neighbor.