Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 6, June 2015, pg.853 – 857

RESEARCH ARTICLE

AN ADVANCE APPROACH IN CLUSTERING HIGH DIMENSIONAL DATA

Kavita R Dhoot, Prof. Manoj. N. Chaudhari

[1]M.Tech.-CSE, Priyadarshini Bhagwati College of Engg, Nagpur, Maharashtra, India

[2]CSE Department, Priyadarshini Bhagwati College of Engg, Nagpur, Maharashtra, India

[1] kavya.dhoot@gmail.com; [2] manojchaudhari@gmail.com

Abstract: Clustering high dimensional data becomes challenging 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 reducing the intra-cluster distance and maximizing the inter-cluster distance which improves the cluster quality.

Keywords

High dimensional data, hubness phenomenon, Kernel mapping, and K-nearest neighbor

1. INTRODUCTION:

Clustering is an unsupervised process of grouping elements together. A cluster is a collection of data objects that are similar to one another within the same cluster and are dissimilar to the objects in other clusters. There are different clustering techniques which, such as hierarchical, partitional, and density-based and subspace [1]. Clustering methods can be used for detecting the underlying structure of the data distribution. Algorithms from the fourth group search for clusters in some lower dimensional projection of the original data, and have been generally preferred when dealing with data that are high dimensional [2], [3], [4], [5].

Partitional clustering methods start with an initial partition of the observation and optimize these partitions according to utility function or distance function. Hierarchical clustering methods works by grouping data objects into a tree of clusters. It can be further classified as either agglomerative or divisive, depending on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Density-based clustering methods regard clusters as dense regions of objects in the data space that are separated by regions of low density. Subspace clustering methods search for groups of clusters within different subspaces of the same data set. This paper mainly focused on partitional clustering. To overcome the problems in partitional clustering methods on high dimensional data, a new algorithm which is based on combination of kernel mappings [7] and hubness phenomenon[6].

2. RELATED WORK

Even though hubness has not been given much attention in data clustering, hubness information is drawn from k nearest-neighbor lists, which have been used in the past to perform clustering in various ways. These lists may be used for computing density estimates, by observing the volume of space determined by the k-nearest neighbors. Density based clustering methods often rely on this kind of density estimation [14], [15], [16]. The implicit assumption made by density-based algorithms is that clusters exist as high density regions separated from each other by low-density regions. In high-dimensional spaces this is often difficult to estimate, due to data being very sparse. There is also the issue of choosing the proper neighborhood size, since both small and large values of k can cause problems for density based approaches [17]. Enforcing k-nearest-neighbor consistency in algorithms such as K-means was also explored [18]. The most typical usage of k-nearest-neighbor lists, however, is to construct a k-NN graph [19] and reduce the problem to that of graph clustering.

Consequences and applications of hubness have been more thoroughly investigated in other related fields: classification [20], [21], [22], [23], [24], image feature representation [25], data reduction [23], [26], collaborative filtering [27], text retrieval [28], and music retrieval [29], [30], [31]. In many of these studies it was shown that hubs can offer valuable information that can be used to improve existing methods and devise new algorithms for the given task

Kernel k-means maps data points from the input space to the high dimensional feature space through a non-linear transformation [8]. The kernel based clustering minimizes the clustering error in feature space.

It is believed that the kernel k-means, which is used with the non-parametric histogram intersection kernel [7], is good for image clustering. In this paper we have proposed a new clustering algorithm which uses the concept of kernel and hubness phenomenon.

3. HUBNESS PHENOMENON

Hubness is an act of high dimensional data to contain points that frequently occur in k-nearest neighbor lists of other points. Let S ⊂ Rd be a set of high dimensional data points and let Nk(y) denote the number of k-occurrences of point y ∈ S, i.e., the number of times y occurs in k-nearest neighbor lists of other points from S. Hubness is an inherent property of high dimensional data which is related to distance concentration phenomenon [4]. The number of k-occurrences of point y ∈ S is referred as hubness score in rest of the text. The frequently occurring data points in k-neighbor sets are referred as hubs and very rarely occurring points are referred as anti-hubs.

3.1 Appearance of Hubs

The concentration of distances enables to view unimodal high dimensional data lying on a hypersphere centered at the data distribution mean. The variance of distances to the mean remains non-negligible for any countable number of dimensions, which indicates that some of the points still end up being closer to the data mean than other points [9]. The points closer to the mean tend to be closer to all other points in the dataset, for any dimensionality that observed. In high dimensional data, this act is made stronger. Such points will have a higher probability of being included in k-nearest neighbor sets of other points in the dataset, which increases their ability, and they emerge as neighbor-hubs.

Hubs can also exist in multimodal data, situated in the nearness of cluster centers [2]. The degree of hubness depends on the intrinsic data dimensionality, i.e., the number of variables needed to represent all pairwise distances in the data. Hubness phenomenon is related to high dimensional data regardless of the distance or similarity measure. The existence of hubs can be verified using Euclidean and Manhattan distances.

3.2 Relation of hub to centroid and medoid

In low dimensional data hubs in the clusters are far-off from the centroids, even out of average points. There is no relationship between cluster means and frequent neighbors in the low dimensional environment [9]. This fact may changes with the increase in dimensionality. The minimal distance from centroid to hub converges to minimal distance from centroid to medoid. This concept implies that some medoids are actually cluster hubs. As medoids the centroids are also closer to data hubs. This relationship brings us to get an idea that the points with high hubness scores are closer to centers of clustered sub regions of high dimensional space than other data points in the dataset. Hence these points can act as cluster representatives [6].

4. KERNEL BASED HUBNESS CLUSTERING

Hubness is viewed as a local centrality measure and is possible to use it for clustering high dimensional data in various ways. There are two types of hubness, namely global hubness and local hubness [9]. Local hubness can be defined as a restriction of global hubness on any given cluster of the current algorithm iteration. Local hubness score represents the number of k-occurrences of a point in k-nearest neighbor lists of elements within the same cluster. Global hubness represents the number of k-occurrences of a point in k-nearest neighbor lists of all elements of the dataset. This global hubness is used for determining the number of clusters automatically.

The high dimensional data contains more number of attributes, in which some attributes are more important for representing the data points. In order to identify the important attributes in the dataset, the Kernel Principal Component Analysis is used. The kernel principal components are used for defining the kernel function. By using the kernel function [7], i.e.,an appropriate non-linear mapping from the original input space to a higher dimensional feature space, clusters that are non-linearly separable in input space can be extracted.

5. CONCLUSION

The hubness phenomenon and kernel mapping will be used for clustering high dimensional data. Hubs are used to approximate local cluster prototypes is not only a feasible option, but also frequently leads to improvement over the centroid-based approach. Kernel hubness clustering algorithm be will designed specifically for high dimensional data. This algorithm is expected to offer improvement by providing higher inter-cluster distance and lower intra-cluster distance. Since kernel mapping is applied, the algorithm will detects arbitrary shaped clusters in the dataset. The hubs automatically determine the number of clusters to be formed. Hence users need not to specify the number of clusters as manually.

REFERENCES

[1] J. Han and M. Kamber (2006), “Data Mining: Concepts and Techniques,” 2nd ed.Morgan Kaufmann Publishers.

[2] C.C. Aggarwal and P.S. Yu, “Finding Generalized Projected Clusters in High Dimensional Spaces,” Proc. 26th ACM SIGMOD Int’l Conf. Management of Data, pp. 70-81, 2000.

[3] K. Kailing, H.-P. Kriegel, P. Kro¨ ger, and S. Wanka, “Ranking Interesting Subspaces for Clustering High Dimensional Data,” Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 241-252, 2003.

[4] K. Kailing, H.-P. Kriegel, and P. Kro¨ger, “Density-Connected Subspace Clustering for High-Dimensional Data,” Proc. Fourth SIAM Int’l Conf. Data Mining (SDM), pp. 246-257, 2004.

[5] E. Muller, S. Gunnemann, I. Assent, and T. Seidl, “Evaluating Clustering in Subspace Projections of High Dimensional Data,”

Proc. VLDB Endowment, vol. 2, pp. 1270-1281, w.

[6] N. Tomasev, M. Radovanovic, D. Mladenic, M. Ivanovic (2013), “The Role of Hubness in Clustering High-Dimensional data,” IEEE Transactions on Knowledge and Data Engineering, vol:pp, issue:99, ISSN:1041-4347.

[7] Grigorios F. Tzortzis and Aristidis C. Likas*,(2009), “*The Global Kernel K-Means Algorithm for Clustering in Feature Space” IEEE Transactions on Neural Networks, Vol. 20. No. 7,PP:1181-1194.

[8] 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.

[9] Miloˇs Radovanovi´c, Alexandros Nanopoulos, and Mirjana Ivanovi´c (2010), “Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data,” Journal of Machine Learning Research, pp. 2487-2531.

[14] T.N. Tran, R. Wehrens, and L.M.C. Buydens, “Knn Density-Based Clustering for High Dimensional Multispectral Images,” Proc. Second GRSS/ISPRS Joint Workshop Remote Sensing and Data Fusionover Urban Areas, pp. 147-151, 2003.

[15] E. Bic¸ici and D. Yuret, “Locally Scaled Density Based Clustering,”Proc. Eighth Int’l Conf. Adaptive and Natural Computing Algorithms (ICANNGA), Part I, pp. 739-748, 2007.

[16] C. Zhang, X. Zhang, M.Q. Zhang, and Y. Li, “Neighbor Number,Valley Seeking and Clustering,” Pattern Recognition Letters, vol. 28,no. 2, pp. 173-180, 2007.

[17] S. Hader and F.A. Hamprecht, “Efficient Density Clustering Using Basin Spanning Trees,” Proc. 26th Ann. Conf. Gesellschaft fu¨r Klassifikation, pp. 39-48, 2003.

[18] C. Ding and X. He, “K-Nearest-Neighbor Consistency in Data Clustering: Incorporating Local Information into Global Optimization,”Proc. ACM Symp. Applied Computing (SAC), pp. 584-589,2004.

[19] C.-T. Chang, J.Z.C. Lai, and M.D. Jeng, “Fast Agglomerative Clustering Using Information of k-Nearest Neighbors,” Pattern Recognition, vol. 43, no. 12, pp. 3958-3968, 2010.