ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

AN EFFICIENT APPROACH FOR CLUSTERING HIGH DIMENSIONAL DATA USING HUB POINTS

S.Vinodhini1 , R. Sivaraj2 and R. Devi Priya3
1 PG-Student, Department of Computer science and Engineering Velalar College of Engineering and Technology vinodhini.subu@gmail.com
2 Assoc.Professor, Department of Computer science and Engineering Velalar College of Engineering and Technology rsivarajcse@gmail.com
3Assistant Professor, Department of Information Technology, Kongu Engineering College scrpriya@gmail.com

ABSTRACT

High-dimensional data arise naturally in many domains. It presented a great challenge for traditional data mining techniques, both in terms of effectiveness and efficiency. Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty in distinguishing distances between data points. This project takes a novel perspective on the problem of clustering high-dimensional data. Instead of attempting to avoid the curse of dimensionality by observing a lower dimensional feature subspace, this project embraces dimensionality by taking advantage of inherently high-dimensional phenomena. More specifically, it is showed that hubness, i.e., the tendency of high-dimensional data to contain points (hubs) that frequently occur in k-nearest-neighbor lists of other points, can be successfully exploited in clustering. This project demonstrates that hubness is a good measure of point centrality within a high-dimensional data cluster, and by proposing several hubness-based clustering algorithms, showing that major hubs can be used effectively as cluster prototypes or as guides during the search for centroid-based cluster configurations. The results demonstrate good performance of the algorithms in multiple settings, particularly in the presence of large quantities of noise. The proposed methods are tailored mostly for detecting approximately hyperspherical clusters and need to be extended to properly handle clusters of arbitrary shapes.

INTRODUCTION

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

managing data, it also includes analysis and prediction. Classification technique is capable of processing a wider variety of data than regression and is growing in popularity.Data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining," a shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw material.

Thus, such a misnomer that carries both "data" and "mining" became a popular choice. Many other terms carry a similar or slightly different meaning to data mining, such as knowledge mining from data, knowledge extraction, data/pattern analysis, data archaeology, and data dredging.Data mining is a synonym for another popularly used term, Knowledge Discovery from Data, or KDD process.

RELATED WORKS

Even though hubness has not been given much attention indata clustering, hubness information is drawn from knearest-neighbor lists, which have been used in the past toperform clustering in various ways. These lists may be used for computing density estimates, by observing the volumeof space determined by the k-nearest neighbors.The implicit assumption made by density-based algorithms is that clusters exist as highdensity 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 theissue of choosing the proper neighborhood size, since both small and large values of k can cause problems for densitybased approaches.Enforcing k-nearest-neighbor consistency in algorithms such as K-means was also explored. The most typical usage of k-nearest-neighbor lists, however, is to construct a k-NN graph and reduce the problem to that ofgraph clustering.In many of these studies it was shown that hubscan offer valuable information that can be used toimprove existing methods and devise new algorithms forthe given task.Finally, the interplay between clustering and hubness was briefly examined in, where it was observed that hubs may not cluster well using conventional prototype-basedclustering algorithms, since they not only tend to be close to points belonging to the same cluster

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

(i.e., have low intracluster distance) but also tend to be close to points assigned to other clusters (low intercluster distance). Hubs
can, therefore, be viewed as (opposing) analogues of outliers,which have high inter- and intracluster distance, suggesting that hubs should also receive special attention.

CLUSTERING HIGH DIMENSIONAL DATA

The difficulties in dealing with high-dimensional data are omnipresent and abundant. However, not all phenomena that arise are necessarily detrimental to clustering techniques. We will show in this paper that hubness, which is the tendency of some data points in high-dimensional data sets to occur much more frequently in k-nearest-neighbor lists of other points than the rest of the points from the set, can in fact be used for clustering.In this project focused on exploring the potential value of using hub points in clustering by designing hubness-aware clustering algorithms and testing them in a high-dimensional context. The hubness is a good measure of point centrality within a high-dimensional data cluster and that major hubs can be used effectively as cluster prototypes. Centroids and medoids in K-means iterations tend to nverge to locations close to high-hubness points, which implies that using hubs instead of either of these could actually speed up the convergence of the algorithms, leading straight to the promising regions in the data space.

A simple way to employ hubs for clustering is to use them as one would normally use centroids. Even though points with highest hubness scores are without doubt the prime candidates for cluster centers, there is no need to disregard the information about hubness scores of other points in the data.

NON-PARAMETRIC DETECTION

Distance concentration is the phenomenon that in certain conditions, the contrast between the nearest and the farthest neighbouring points vanishes as the data dimensionality increases. It affects high dimensional data processing, analysis, retrieval, and indexing, which all rely on some notion of distance or dissimilarity. Previous work has characterized this phenomenon in the limit of infinite dimensions. However, real data is finite dimensional, and hence the infinite-dimensional characterization is

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

insufficient. Here we quantify the phenomenon more precisely, for the possibly high but finite dimensional case in a distribution-free manner, by bounding the tails of the probability that distances become meaningless. As an application, we show how this can be used to assess the concentration of a given distance function in some unknown data distribution solely on the basis of an available data sample from it. This can be used to test and detect problematic cases more rigorously than it is currently possible, and we demonstrate the working of this approach on both synthetic data and ten real-world data sets from different domains.

Previous work has characterized the phenomenon of distance concentration asymptotically, in the limit of infinite dimensions. This enables analyses of a given distance function in a given data model family, and allows us to identify conditions on the data distribution that matter. E.g. the existence correlations among the features were shown to be a favorable trait whereas weakly dependent or independent features lead to meaningless distances in high dimensions. Another stream of research seeks to alleviate the problem by devising dissimilarity functions that suffer less in a worst-case scenario, e.g. on i.i.d. uniformly distributed features.

ALGORITHMS

- Hub-based clustering
- Hubness-proportional Clustering(HPC)
- K-means as k-hubs
- Shared-neighbor clustering

HUB-BASED CLUSTERING

- Centroids and medoids in K-means iterations tend to converge to locations close to high-hubness points.
- Computational complexity of hubness-based algorithms is mostly determined by the cost of computing hubness scores.

HUBNESS- PROPORTIONAL CLUSTERING (HPC)

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)
International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)
Vol. 2, Special Issue 10, March 2016

- Though points with highest hubness scores are without doubt the prime candidates for cluster centers.
- There is no need to disregard the information about hubness scores of other points in the data.

HUBNESS-PROPORTIONAL K-MEANS

- It is nearly identical to HPC, the only difference being in the deterministic phase of the iteration.
- Instead of reverting to k-hubs, the deterministic phase executes k-means updates.

SHARED-NEIGHBOR CLUSTERING

- It finds clusters of different sizes, shapes and densities from very large and high dimensional data sets.
- This algorithm first finds the list of nearest neighbors for each point and then redefines the similarity between points .
- The shared neighbor similarity takes the sum of the similarity of the points nearest neighbors as a measure of density.

CONCLUSION

Instead of attempting to avoid the curse of dimensionality by observing a lower dimensional feature subspace, embrace dimensionality by taking advantage of inherently high-dimensional phenomena. In this project shown that using hubs to approximate local data centers is not only a feasible option, but also frequently leads to improvement over the centroid-based approach.The proposed Hubness based probalistic clustering (HPC) method had proven to be more robust than the K-Means baseline on both synthetic and real-world data, as well as in the presence of high levels of artificially introduced noise. This initial evaluation suggests that using hubs both as cluster prototypes and points guiding the centroid-based search is a promising new idea in clustering high-dimensional and noisy data. Also, global hubness estimates are generally to be preferred with respect to the local ones. Hub-based algorithms are designed specifically for high-dimensional data.

ISSN 2395-695X (Print)
ISSN 2395-695X (Online)

International Journal of Advanced Research in Biology Engineering Science and Technology (IJARBEST)

Vol. 2, Special Issue 10, March 2016

REFERENCES

[1] Muller.E et al.,(2009), "Evaluating Clustering in Subspace Projections of High Dimensional Data," Proc. VLDB Endowment, vol. 2, pp. 1270-1281.
[2] Ning.K et al.,(2010), "Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology," BMC Bioinformatics, vol. 11, pp. 1-14.
[3] Chang.C.T et al., (2010), "Fast Agglomerative Clustering Using Information of k-Nearest Neighbors," Pattern Recognition, vol. 43, no. 12, pp. 3958-3968.
[4]Toma sev.N et al.,(2011) "Hubness-Based Fuzzy Measures for High-Dimensional k-Nearest Neighbor Classification," Proc. Seventh Int'l Conf.Machine Learning and Data Mining (MLDM), pp. 16-30.
[5] Radovanovi.M et al., (2010), "Time-Series Classification in Many Intrinsic Dimensions," Proc. 10th SIAMInt'l Conf. Data Mining (SDM), pp. 677-688.
[6]Radovanovic.M et al.,(2010), "Hubs inSpace: Popular Nearest Neighbors in HighDimensional Data,"J. Machine Learnin Research,vol. 11, pp. 2487-2531.