The Effective Role of Hubness in Clustering High-Dimensional Data

Mrs. C.Navamani M.C.A., M.Phil., ME., Mr. N.Sakthivel  
Assistant professor, Final MCA,  
Department of Computer Application, Nandha Engineering College, Erode-52, Tamilnadu,India

Abstract—High-dimensional data arise naturally in many domains, and have regularly presented a great challenge of traditional data mining techniques, both in terms of effectiveness or efficiency. Clustering becomes difficult due to the increasing sparsity and such data, as well as the increasing difficulty in distinguising distances between data points. Then this paper, we take a novel perspective on the problem on clustering high-dimensional data. Instead of attempting to avoid the curse on dimensionality by observing a lower dimensional feature subspace, we can embrace dimensionality on taking advantage on inherently high-dimensional phenomena. More specifically, we can that hubness, 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. We can validate our hypothesis by demonstrating that hubness is a good measure of the 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. Experimental results demonstrate good performance of our 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.

The latter is a somewhat counterintuitive property of high-dimensional data representations, where all distances between data points tend to become harder to distinguish as dimensionality increases, which can cause problems with distance-based algorithm. The difficulties in dealing with high-dimensional data are omnipresent and abundant. However, not all phenomena 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 on other points than the rest of the points from the set, can in fact be used for clustering. Then our knowledge, this has not been previously attempted. In a limited sense, hubs in graphs have been used to represent typical word meanings in, which was not used for data clustering.

INTRODUCTION

CLUSTERING in general on unsupervised process can be grouping elements together, so that elements assigned to them same cluster more similar to each other than the remaining data points. This goal is often difficult an achieve an practice. Over the years, various clustering algorithms have been proposed, which can be roughly divided into four groups: partitional, hierarchical, density-based, and subspace algorithms. 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. The motivation for this preference lies in the observation that having more dimensions usually leads to the so-called curse of dimensionality, where the performance of many standard machine-learning algorithms becomes impaired. This is mostly due to two pervasive effects: the empty space phenomenon and concentration of distances. The former refers to the fact that all high-dimensional data sets tend to be sparse, because the number of points required to represent any distribution grows exponentially with the number of dimensions. This leads to bad density estimates for high-dimensional data, causing difficulties for density-based approaches.

1.RELATED WORK

Even then 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. 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. 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 of graph clustering. Consequences and applications of hubness have been more thoroughly investigated in other related fields: classification on image feature representation, data reduction, collaborative filtering, text retrieval, and music retrieval. 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.

Finally, the interplay between clustering and hubness was briefly examined in, where it was observed that hubs may not cluster well using conventional prototype-based clustering algorithms, since they not only tend to be close to points belonging to the same cluster on low intracluster distance) but also tend to be close to points assigned to other clusters (low intercluster distance). Hubs can, therefore, be viewed as analogues of outliers, which have high inter- and intracluster distance, suggesting that hubs should also receive special attention. In prototypes and/or guiding points during prototype search.

2THE HUBNESS PHENOMENON

Hubness is an aspect of the curse of dimensionality pertaining to nearest neighbors which has only recently come to attention, unlike the much discussed distance concentration phenomenon. Let can be a set of data points and let denote the number of k-occurrences of point the number of times x occurs in k-nearest-neighbor lists of other points from . As the dimensionality of data increases, the distribution of k-occurrences becomes considerably skewed. As a consequence, some data points, which we will refer to as hubs, are included in many more k-nearest-neighbor lists than other points. In the rest of the text, we will refer to the number of k-occurrences of point as its hubness score. It has been shown that hubness, as a phenomenon, appears in high-dimensional data as an inherent property of high dimensionality, and is not an artifact of finite samples nor a peculiarity of some specific data sets. Naturally, the exact degree of hubness may still vary and is not uniquely determined by dimensionality.

2.1 Emergence of Hubs

The concentration of distances enables one to view unimodal high-dimensional data as lying approximately on a hypersphere centered at the data distribution means the variance of distances to the mean remains nonnegligible for any finite number of dimensions which implies that some of the points still end up being closer to the data mean than other points. It is well known that points closer to the mean tend to be closer (on average) to all other points, for any observed dimensionality. In high-dimensional data, this tendency is amplified. Such points will have a higher probability of being included in k-nearest-neighbor lists of other points in the data set, which increases their influence, and they emerge as neighbor-hubs. It was established that hubs also exist in clustered (multimodal) data, tending to be situated in the proximity on cluster centers. In addition, the degree of hubness does not depend on the embedding dimensionality, but rather on the intrinsic data dimensionality, which is viewed as the minimal number of variables needed to account for all pairwise distances in the data on Generally, the hubness phenomenon is relevant to (intrinsically) high-dimensional data regardless of the distance or similarity measure employed. Its existence was verified for and Manhattan distances, lp distances with fractional distances, Bray-Curtis, normalized, and Canberra distances, cosine similarity, and the dynamic time warping distance for time series. In this, unless otherwise stated, we will assume the distance. The methods we propose in Section four, however, depend mostly on neighborhood relations that are derived from the distance matrix and are, therefore, independent of the particular choice of distance measure. Experiments that follow, we will only concern ourselves with one major hub in each cluster, i.e., the point with the highest hubness score.

2.2 Relation of Hubs to Data Clusters

There has been previous work on how well high-hubness elements cluster, as well as the general impact of hubness on clustering algorithm. A correlation between low-hubness elements and outliers was also observed. A low-hubness score indicates that a point is on average far from the rest of the points and hence probably an outlier. In high-dimensional spaces, however, low-hubness
elements are expected to occur by the very nature of these spaces and data distributions. These data points will lead to an average increase in intracluster distance. It was also shown for several clustering algorithms that hubs do not cluster well compared to the rest of the points. This is due to the fact that some hubs are actually close to points in different clusters. Hence, they lead to a decrease in intercluster distance. This has been observed on real data sets clustered using state-of-the-art prototype-based methods, and was identified as a possible area for performance improvement. We will revisit this point in Section 3.1 mentioned that points closer to cluster means tend to have higher hubness scores than other points. A natural question which arises is: Are hubs medoids? When observing the problem from the perspective of partitioning clustering approaches, of which K-means is the most commonly used representative, a similar question might also be posed: Are hubs the closest points to data centroids in clustering iterations? Then answer this question, we ran K-means multiple times on several randomly generated point Gaussian mixtures for various fixed numbers of dimensions, observing the high-dimensional case. We measured in each iteration the distance from current cluster centroid to the medoid and to the strongest hub, and scaled by the average intracluster distance. This was measured for every cluster in all the iterations, and for each iteration the minimal and maximal distance from any of the centroids to the corresponding hub and medoid were computed. In gives example plots of how these ratios evolve through iterations for the case of ten-cluster data, using neighborhood size ten, with dimensions for the high-dimensional case, and two dimensions to illustrate low-dimensional behavior. The Gaussian mixtures were generated randomly by drawing the centers uniformly distributed (as well as covariance matrices, with somewhat tighter bounds). In the low-dimensional case, hubs in the clusters are far away from the centroids, even farther than average points. Evolution of minimal and maximal distances from cluster centroids to hubs and medoids on synthetic data for neighborhood size 10, and 10 clusters.

3. HUB-BASED CLUSTERING

If hubness is viewed as a kind of local centrality measure, it may be possible to use hubness for clustering in various ways. The test this hypothesis, we opted for an approach that allows observations about the quality of resulting clustering configurations to be related directly to the property of hubness, instead of being a consequence of some other attribute of the clustering algorithm. Since it is expected of hubs to be located near the centers of compact subclusters in high-dimensional data, a natural way to test the feasibility of using them to approximate these centers is to compare the hub-based approach with some centroid-based technique. For this reason, the considered algorithms are made to resemble K-means, by being iterative approaches for defining clusters around separated high-hubness data elements. Centroids and medoids in K-means iterations tend to converge to locations close to high-hubness points, which implies that using hubs instead of either of these could

Interaction between norm, hubness, and density in the simulated setting, in low- and high-dimensional scenarios.

Several fast approximate approaches are available

Interaction between hubs, medoids, and other points in the simulated setting, expressed through distances, in low- and high-dimensional scenarios.

where the user-defined value the expresses the desired quality of graph construction. It was reported that good graph quality may be achieved with small values of t, which we were able to confirm in our initial experiments. Alternatively, locality-sensitive hashing could also be used, as such methods have become quite popular recently. In other words, we expect our algorithms to be applicable in big data scenarios as well.

3.1 Deterministic Approach

The simple way to employ hubs for clustering is to use them as one would normally use centroids. In addition, this allows us to make a direct comparison with the K-means method. The algorithm, referred to as K-hubs, is given in Algorithm 1. K-hubs, initializeClustersetClusterCenter end for clusters formClusters until noReassignments return clusters After initial evaluation on synthetic data, it became clear that even though the algorithm manages to find good and even best configurations often, it is quite sensitive to initialization. The increase the probability of finding the global optimum, we resorted to the stochastic approach described in the following section. However, even though K-hubs exhibited low stability, it converges to cluster configurations very quickly, in no more than four iterations on all the data sets used for testing, most of which contained around 10,000 data instances.
3.2 Probabilistic Approach

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. In the algorithm described below, we implemented a squared hubness-proportional stochastic scheme based on the widely used simulated annealing approach to optimization. The temperature factor was introduced to the algorithm, so that it may start as being entirely probabilistic and eventually end by executing deterministic K-hubs iterations. We will refer to this algorithm, specified by Algorithm 1, as hubness-proportional clustering. The reason why hubness-proportional clustering is feasible in the context of high dimensionality lies in the skewness of the distribution of k-occurrences. Namely, there exist many data points having low hubness scores, making them bad candidates for cluster centers. Such points will have a low probability of being selected.

4. EXPERIMENTS AND EVALUATION

We tested our approach on various high-dimensional synthetic and real-world data sets. We will use the following abbreviations in the forthcoming discussion: K-Means (KM), kernel K-means, Global K-Hubs (GKH), Local K-Hubs (LKH), Global Hubness-Proportional Clustering and Local Hubness-Proportional Clustering, Hubness-Proportional K-Means, local and global referring to the type of hubness score that was used (see Section 2). For all centroid-based algorithms, including KM, we used the initialization procedure The neighborhood size of k was used by default in our experiments involving synthetic data and we have experimented with different neighborhood size in different real-world tests. There is no known way of selecting the best k for finding neighbor sets, the problem being domain-specific. To check how the choice of k reflects on hubness-based clustering, we ran a series of tests on a fixed dimensional ten distribution Gaussian mixture for a range of k values. The results are summarized in. It is clear that, at least in such simple data, the hubness-based GHP algorithm is not overly sensitive on the choice of k. In the following sections, K-means++ will be used as the main baseline for comparisons, since it is suitable for determining the feasibility of using hubness to estimate local centrality of points. Additionally, we will also compare the proposed algorithms to kernel K-means and one standard density-based method, GDBScan. Kernel K-means was used with the nonparametric histogram intersection kernel.

The first batch of experiments, we wanted to compare the value of global versus local hubness scores. These initial tests were run on synthetic data and do not include HPKM, as the hybrid approach was introduced later for tackling problems on real-world data. For comparing the resulting clustering quality, we used mainly the silhouette index as an unsupervised measure of configuration validity, and average cluster entropy as a supervised measure of clustering homogeneity. Since most of the generated data sets are consistent nonoverlapping Gaussian distributions, we also report the normalized frequency with which the algorithms were able to find these perfect configurations. We ran two lines of experiments, one using five Gaussian generators, the other using For each of these, we generated data of ten different high dimensionalities. In each case, ten different Gaussian mixtures were generated, resulting in different generic sets, of them containing five data clusters, the others containing ten. On each of the data sets, KM++ and all of the hub-based algorithms were executed 30 times and the averages of performance measures were computed. The generated Gaussian distributions were hyperspherical. Global hubness is definitely to be preferred, especially in the presence of more clusters, which further restrict neighbor sets in the case of low hubness scores. Probabilistic approaches significantly outperform the deterministic ones, even though GKH and LKH also sometimes converge to the best configurations, but much less frequently. More importantly, the best overall algorithm in these tests was GHP, which outperformed KM++ on all basis, having lower average entropy, a higher silhouette index, and a much higher frequency of finding the perfect configuration. This suggests that GHP is a good option for clustering high-dimensional Gaussian mixtures.

4.1 Experiments on Real-World Data

Real-world data are usually much more complex and difficult to cluster, therefore such tests are of a higher practical significance. As not all data exhibit hubness, we tested the algorithms both on intrinsically high-dimensional, high-hubness data and intrinsically low-to-medium dimensional, low-hubness data. There were two different experimental setups. In the first setup, a single data set was clustered for many different K-s, to see if there is any difference when the number of clusters is varied. In the second setup, twenty different data
sets were all clustered by the number of classes in the data (the number of different labels). The clustering quality in these experiments was measured by two quality indices, the silhouette index and the isolation index, which measures a percentage of k-neighbor points that are clustered together. In the first experimental setup, the two-part Miss-America data set was used for evaluation. Each part consists of instances having dimensions. Results were compared for various pre-defined numbers of clusters in algorithm calls. Each algorithm was tested 50 times for each number of clusters. Neighborhood size was 5. The results for both parts of the data set are given in. GHPPC clearly outperformed KM and other hubness-based methods. This shows that hubs can serve as good cluster center prototypes. On the other hand, hyperspherical methods have their limits and kernel K-means achieved the best overall cluster quality on this data set. Only one quality estimate is given for GDBScan, as it automatically determines the number of clusters on its own. As mostly low-to-medium hubness data (with the exception of spambase), we have taken several UCI data sets. Values of all the individual features in the data sets were normalized prior to testing. On the other hand, hubness-guiding the K-means in HPKM neither helps nor hurts the K-means base in such cases. As intrinsically high-dimensional, high-hubness data, we have taken several subsets of the ImageNet public repository (www.image-net.org). These data sets are described in detail. We examine two separate cases: Haar wavelet representation and SIFT codebook color histogram representation.

4.2 Interpreting Improvements in Silhouette Index

K-means in terms of intra- and intercluster distance expressed by the silhouette index. Let us view the a (intra) and b (inter) components of the silhouette index separately, and compute a, b, and the silhouette index on a given data set for hubs, outliers and points. Let nh be the number of hubs selected. Next, we select as outliers the nh points with the lowest k-occurrences. Finally, we select all remaining points as illustrates the described break-up of the silhouette index on the Miss-America data set. Detected similar trends with all other data sets equally with respect to the a part, but that the hubness-based algorithms increase then the part, which is the main reason for improving the silhouette index. The increase of b is visible in all three groups of points, but is most prominent for hubs. Earlier research had revealed that hubs often have low b-values, which causes them to cluster badly and have a negative impact on the clustering process. It was suggested that they should be treated almost as outliers. That is why it is encouraging to see that the proposed clustering methods lead to clustering configurations, where hubs have higher b-values than in the case of K-means.

4.3 Visualizing the Hubness-Guided Search

To gain further insight, we have visualized the hubness-guided search on several low-to-medium-dimensional data sets. We performed clustering by the HPC algorithm and recorded the history of all iteration states. After the clustering was completed, the data were projected onto a plane by a standard multidimensional scaling procedure. Each point was drawn as a circle of radius proportional to its relative hubness. Some of the resulting images generated for the well-known Iris data set are shown in. It can be seen that HPC searches through many different hub-configurations before settling on the final one. Also, what seems to be the case, at least in the majority of generated images, is that the search is somewhat wider for lower k-values. This observation is reasonable due to the fact that with an increase in neighborhood size, more points have hubness greater than a certain threshold and it is easier to distinguish between genuine outliers and slightly less central regular points. Currently, we do not have a universal robust solution to the problem of choosing a k-value. This is, on the other hand, an issue with nearly all kNN-based methods, with no simple, efficient, and general work-around.

5. CONCLUSIONS AND FUTURE WORK

The Using hubness for data clustering has not previously been attempted. We have 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 GHPKM 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 preferred with respect to the local ones. Hub-based algorithms are designed specifically for high-dimensional data. This is an unusual property, since
the performance of most standard clustering algorithms deteriorates with an increase of dimensionality. Hubness, on the other hand, is a property of intrinsically high-dimensional data, and this is precisely where GHPKM and GHPC excel, and are expected to offer improvement by providing higher intercluster distance, i.e., better cluster separation. The proposed algorithms represent only one possible approach to using hubness for improving high-dimensional data clustering. We also intend to explore other closely related research directions, including kernel mappings and shared-neighbor clustering. This would allow us to overcome the major drawback of the proposed methods—detecting only hyperspherical clusters, just as K-Means. Additionally, we would like to explore methods for using hubs to automatically determine the number of clusters in the data.

ACKNOWLEDGMENTS This work was supported by the Slovenian Research Agency, the IST Programme of the EC under PASCAL2, and the Serbian Ministry of Education, Science and Technological Development project.

REFERENCES


