ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.9

Use of Hub Concept for Image Clustering by Extracting Features: A Survey

Sneha Sumare[1], Swati Khavare[2]
Assistant Professor
Department of Computer Science and Engineering
Bharati Vidyapeeth's College of Engineering, Kolhapur, Maharashtra, India

Abstract:
Many data domains such as personal data, image data have large number of attributes which is difficult to cluster using traditional data mining techniques due to high dimensionality of data. Image clustering is one of the essential problems in automatic image processing, since high dimensional data exhibit high hubness which is data points appears at high density area of data. We find that image data set under several feature can be represented and cluster using hubness. We propose a novel approach of hubness in clustering of high dimensional data like images. Each feature of image including its resolution will treat as dimension of the image and using these all dimensions we apply clustering using hub concept. Hub is the data point that frequently occurs in k-nearest neighbor of other data points. This hub point can be used effectively as centroid in cluster prototype which will considerably speed up the convergence of algorithm.

Keywords: Hub concept, image clustering, image feature extraction, high-dimensionality, hub-computation.

I. INTRODUCTION

Clustering is a process of grouping similar data elements together so that they posses similar feature to other members in the same group and dissimilar to data points in other clusters. Image clustering and categorization is a means for high-level description of image content. The goal is to find a mapping of the archive images into classes (clusters) such that the set of classes provide essentially the same prediction, or information, about the image archive as the entire image set collection The features on which clustering being performed is depends on data type. This feature acts like dimension of that data [1]. It is well known that many machine learning algorithms plagued by curse of dimensionality, since many real word data sets (e.g. Image Data Sets) consist of very high dimensional feature space. This comprises a set of properties which tends to become more pronounced as data dimensionality increases [2]. The main cause of all is unavoidable sparseness of data. In high dimensionality, there is not enough data to make reliable density estimation. The goal of clustering over high dimensional data becomes difficult due to empty space phenomenon and concentration of distances. This leads to bad density clusters for density based approaches. Furthermore, the property of high dimensional data representation presents distance between data points large enough to become harder to distinguish [4]. This hardness increases with dimensionality since the distance between two points in space become larger and larger. The data domains like images are most interested data for clustering purpose. Also, image clustering is essential Ease of Use for purposes like image retrieval, image classification, object detection etc. The local features are of intermediate complexity, which means that they are distinctive enough to determine likely matches in a large database of features [5]. For increase in features for clustering it become harder to compare with others. We focus on hub point in image clustering by designing hubness aware clustering algorithm to clustering of high dimensional data like image data [4]. An image data will be represented using its different features. These features usually contain information extracted from Color, texture, edges and any property which we feel important for image clustering [6]. The methods like k-mean and KNN are much powerful and widely used in classification methods. But these methods reduce their performance as increase in dimensionality of data. There are some more difficulties in dealing with high-dimensional data are omnipresent and abundant [4]. Same as other these two is also due to sparseness of data point. Hubs appear as a consequence of the geometry of high-dimensional space, and the behavior of data distributions within them [2]. Hubness 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 [4]. It is a high dimensional phenomenon which concern k-nearest-neighbor set. Denoted by Nk(x) the number of k occurrences of x i.e. the number of times x appears in k-nearest neighbor list of other points in data. The distribution of Nk (x) exhibits significant skew in high dimensional cases, skew which increases with intrinsic dimensionality of the data [2]. This leads to the emergence of hubs, influential points which affect the reasoning procedure of nearest-neighbor based methods for many data points. Fig. 1 shows illustrative example by which we get idea of "Hub" in data.

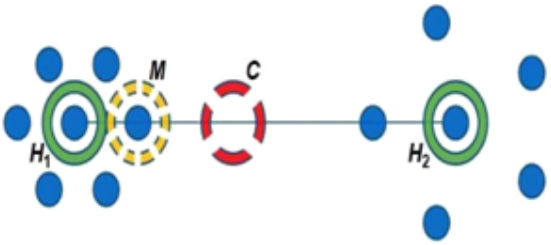

Figure.1. Illustration of Hub
The red dashed circle marks the centroid (C), yellow dotted circle the medoid (M), and green circles denote two elements of highest hubness (H1;H2), for neighborhood size 3. Through image clustering, we aim at developing techniques that support effective searching and browsing of large image digital libraries based on automatically derived image features [19].

Recent methods provide image retrieval based on neighborhood of the query image using similarity measure. But target images with high feature similarities to the query image may be quite different from the query image in terms of semantics due to the semantic gap [19]. We will tackle the problem by using hub-based high-dimensional image clustering technique.

II. LITERATURE REVIEW

Concept of hubness has not being given much attention in data clustering techniques. In 2006, Thanh N. Tran, RonWehrens, Lutgarde M.C. Buydens propose a new method of density based clustering algorithm called KNNCLUST is presented in [1]. In that, they estimate KNN Kernel density for multivariate data, and have two advantages over other methods. First, adaptive kernel width and second is smooth estimator. Nenad Tomasev, Milos Radovanovic, Dunja Mladenic, and Mirjana Ivanovic propose three approaches for hub based clustering: Deterministic approach, Probabilistic approach and Hybrid approach. They treat data points which will be referred to as hub and choose data points with high hubness score to approximate centroid of cluster[4]. Nenad Tomasev, Raluca Brehar, Dunja Mladenic and Sergiu Nedevschi studied Influence of hubness in object recognition by using image features like Haar, SIFT features and provide classification over data set [2]. Karin Kailing, Hans-Peter Kriegel, Peer Kroger, and Stefanie Wanka propose method for clustering high dimensional data using ranking of intersecting areas. They define the concept of "intersectingness" so that they able to detect all subspace containing clusters of arbitrary size and shape. Since, hubness information is drawn from k-nearest neighbor list 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 [7]. 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. Karin Kailing, Hans-Peter Kriegel, Peer Kroger presents new approach of density connectivity to conquer subspace clustering problem having advantages like well shaped and positioned clusters [3]. Enforcing k-nearest-neighbor consistency in algorithms such as K-means was also explored [8]. The most typical usage of k-nearest-neighbor lists, however, is to construct a k-NN graph [9] and reduce the problem to that of graph clustering. Consequences and applications of hubness have been more thoroughly investigated in other related fields: classification [2], [10], [11], [12], [8], image feature representation [7], data reduction [12], [13], collaborative filtering [14], text retrieval [15], and music retrieval [16], [17], [18]. 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 [12], 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 (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 [12]. In this paper, we have adopted the approach of using hubs as cluster prototypes and/or guiding points during prototype search.

III.CONCLUSION

We conclude that this is the first attempt of Use of hubness for clustering of image data type. We found that hubs act as local data centers for multidimensional sparse data. Hub is a feasible option and also improves the algorithm performance over the centroid-based approach. For high dimensional data like images, hub based algorithms are designed. Since, the efficiency and performance of well known and mostly used clustering algorithms decreases with dimensionality increases. But hub concept minimize the convergence time of algorithm.

IV. REFERENCES

[1].S .Gordon, H. Greenspan and J. Goldberger, "Applying the Information Bottleneck Principle to Unsupervised Clustering of Discrete and Continuous Image Representations" Proc. IEEE 9th Int'l Conf. on Computer Vision (ICCV 2003)

[2].N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "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, 2011.

[3].K. Kailing, H. P. Kriegel, and P. Kroger, "Density-Connected Subspace Clustering for High-Dimensional Data," Proc. Fourth SIAM Int'l Conf. Data Mining (SDM), pp. 246-257, 2004.

[4].N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic," The Role of Hubness in Clustering High-Dimensional Data," IEEE Transactions on Knowledge and Data Engineering, Vol. 26, NO. 3, pp. 739-751, March 2014.

[5].David G. Lowe, "Local Feature View Clustering for 3D Object Recognition," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, Kauai, Hawaii (December 2001)

[6].N. Tomasev, R. Brehar, D. Mladenic, and S. Nedevschi, "The Influence of Hubness on Nearest-Neighbor Methods in Object Recognition," Proc. IEEE Seventh Int'l Conf. Intelligent Computer Comm. and Processing (ICCP), pp. 367-374, 2011.

[7].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 Fusion. Over Urban Areas, pp. 147-151, 2003.

[8].N. Tomasev and D. Mladenic, "Nearest Neighbor Voting in High Dimensional Data: Learning from Past Occurrences," Computer Science and Information Systems, vol. 9, no. 2, pp. 691-712, 2012.

[9].K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, "INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification," Proc. 15th Pacific-Asia Conf.

Knowledge Discovery and Data Mining (PAKDD), Part II, pp. 149-160, 2011.

[10].N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic, "A Probabilistic Approach to Nearest-Neighbor Classification: Naïve Hubness Bayesian kNN," Proc. 20th ACM Int'l Conf. Information and Knowledge Management (CIKM), pp. 2173-2176, 2011.

[11].M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Time-Series Classification in Many Intrinsic Dimensions," Proc. 10th SIAM Int'l Conf. Data Mining (SDM), pp. 677-688, 2010.

[12].M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Hubs in Space: Popular Nearest Neighbors in High-Dimensional Data," J. Machine Learning Research, vol. 11, pp. 2487-2531, 2010.

[13].K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, "INSIGHT: Efficient and Effective Instance Selection for Time-Series Classification," Proc. 15th Pacific-Asia Conf. Knowledge Discovery and Data Mining (PAKDD), Part II, pp. 149-160, 2011.

[14].A. Nanopoulos, M. Radovanovic, and M. Ivanovic, "How Does High Dimensionality Affect Collaborative Filtering?" Proc. Third ACM Conf. Recommender Systems (RecSys), pp. 293-296, 2009.

[15].M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "On the Existence of Obstinate Results in Vector Space Models," Proc. 33rd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval, pp. 186-193, 2010.

[16].J.J. Aucouturier and F. Pachet, "Improving Timbre Similarity: How High Is the Sky?" J. Negative Results in Speech and Audio Sciences, vol. 1, 2004.

[17].J.J. Aucouturier, "Ten Experiments on the Modelling of Polyphonic Timbre," PhD dissertation, Univ. of Paris 6, 2006.

[18].D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Local and Global Scaling Reduce Hubs in Space," J. Machine Learning Research, vol. 13, pp. 2871-2902, 2012.

[19].Y. Chen, J. Z. Wang, R. Krovetz, "Content-Based Image Retrieval by Clustering" MIR'03, November 7, 2003, Berkeley, California, USA.