

# Hubness Implementation for High Dimensional Data Clustering Using Image Feature Extraction

Ms. Sumare Sneha Prakash<sup>1</sup> Mr. Khandagale Hridayanath P.<sup>2</sup>

<sup>1</sup>Student <sup>2</sup>Assistant professor

<sup>1,2</sup>Department of Technology

<sup>1,2</sup>Shivaji University, Vidyanagar, Kolhapur 416 004

**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.

**Key words:** Hubness, image feature extraction, high-dimensionality, hub-computation

## I. INTRODUCTION

Clustering is a process of grouping similar data elements together so that they possess 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 world 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 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  $N_k(x)$  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  $N_k(x)$  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.

## II. LITERATURE REVIEW

Concept of hubness has not being given much attention in data clustering techniques. In 2006, Thanh N. Tran, Ron Wehrens, 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 Mladenec, 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].

Also, Nenad Tomasev, Raluca Brehar, Dunja Mladenec and Sergiu Nedeveschi 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 [9]. The most typical usage of k-nearest-neighbor lists, however, is to construct a k-NN graph [10] and reduce the problem to that of graph clustering.

### III. PROPOSED WORK

#### A. Implementation Of HBIC:

In this proposed technique after loading of data base images clusters are formed by applying feature extraction followed by HPC [4].

##### 1) Pre-Processing:

It is done over all data set to reduce noisy data and improve the quality of data set. It is done before all the processes in the proposed system.

##### 2) Feature Extraction:

Features are information carry by image which is used for mining purpose. Feature extraction process finds the highlighted information from image and using these information we can compare between two images. Features are of two types: low level and high level.

##### 3) Clustering:

Firstly, distance between different features of all images will be calculated. This distance will treat as data points on which further processing will performed. Clustering is performed to group the similar images having similar color feature in image database for fast image retrieval and then hub computation algorithm is applied to find out the center images of these clusters so that query images features are only compared to cluster's center images, which cluster's center image having more similarity with query image, that whole cluster images are compared to query image. So there is no need to compare query image to database's all images [11].

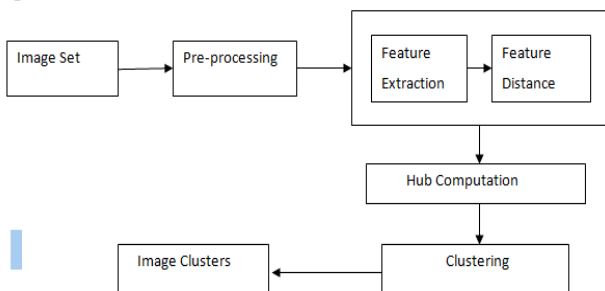


Fig. 1: Hub Based Image Clustering (HBIC)

#### B. Implementation Of HBIR:

In retrieval process features of input query image is extracted using feature extraction module and compared with only clustered image's feature. Using these feature comparison distance between all images and query image is computed.

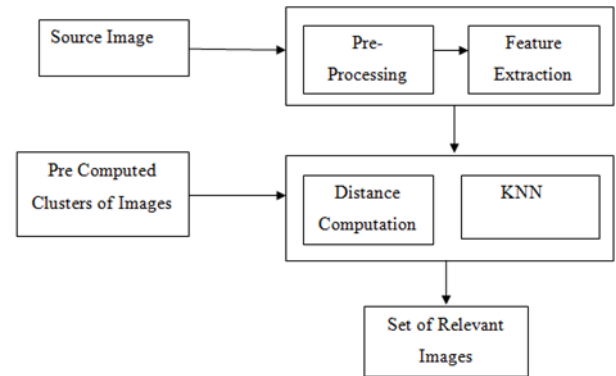


Fig. 2: Hub Based Image Retrieval (HBIR)

#### C. Hub Computation Process:

- (1) Compute Feature Distance Set  $IRd = \{x_1, x_2, \dots\}$  be a real word multi-dimensional data points.
- (2) Initialize hubness scores of all data points to zero
- (3) Compute k-nearest-neighbor list  $D_k(x)$  for all data point
- (4) According to  $D_k(x)$ , calculate hubness scores for each all  $x_i \in D_k(x)$ ;  
 $N_k(x_i) = |D_k(x)|$
- (5) Compute  $H(x) = \{x_i | N_k(x_i) > k\}$ ; where  $H(x)$  is set of hub point.

#### D. Hubness Probabilistic Clustering (HPC) Process:

- (1) Take  $H(x) = \{x_i | N_k(x_i) > k\}$ ; set of hub points from data.
- (2) Perform clustering according to k-nearest-neighbor for each  $H(x)$
- (3) Re-Compute cluster centroid according to hub point and cluster members
- (4) If centroid changes then go to step 2;
- (5) Return Clusters

#### E. Hub – Based Image Retrieval Process:

- (1) Extract Features of source image
- (2) Compute distance to cluster for source images
- (3) Perform KNN to find relevant image
- (4) Return set of relevant images

### IV. SYSTEM ANALYSIS

In fig. 3 we choose the directory in which our data set images are stored. Once you choose your dataset this screen shows preview of all images in normal, resize and in grey scale manner.

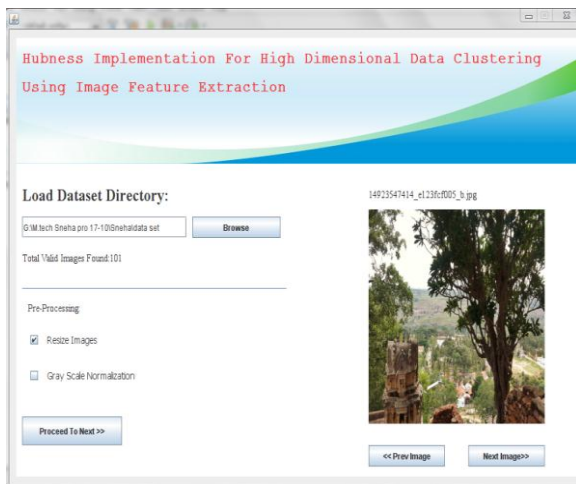


Fig. 3: Data set selection

In fig 4 shows screen after KNN computation for  $k=5$  and Hubness score computation based on KNN list of each data point. Using these hubscore we have to cluster image dataset.

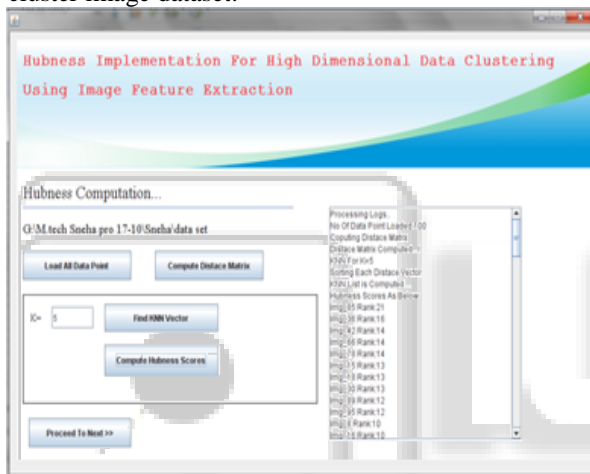


Fig. 4: Hub score computation

## V. CONCLUSION AND FUTURE WORK

Using hubness for image 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. 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. Hub-based algorithms are designed specifically for high dimensional data like images.

This is an unusual property, since the performance of most standard clustering algorithms deteriorates with an increase of dimensionality. The proposed algorithms represent only one possible approach to using hubness for improving image clustering.

## 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 9<sup>th</sup> 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. Nedeveschi, "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] V.S.V.S Murthy et al. "Content based image retrieval using Hierarchical and K-means clustering techniques" IJEST Vol. 2(3), 2010, pp. 209-212.