ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 5, Issue 12, December 2016**

# Clustering of High-Dimensional Data Using Hubness

Dinesh Vardam[1], Vikas Thombre[2]

P.G. Student, Department of Computer Engineering, SKNSITS, Lonavala, Maharashtra, India[1]

Associate Professor, Department of Computer Engineering, SKNSITS, Lonavala, Maharashtra, India[2]

**ABSTRACT**:Cluster analysis or clustering is the task of grouping a set of objects in such a way that objectsin the same group are more similar to each other than to those in other groups .Clustering ofthe images is one of the interested area in data mining and image processing. Traditional algorithms of image clustering propose algorithm with lower dimensional feature i.e. they considerless number of features which will become difficult task and inefficient algorithm. Clustering becomes difficult due to the increasing sparsity of such data, as well as the increasing difficulty indistinguishing distances between data points. Here we take a novel perspective on the problem ofclustering high-dimensional data. Instead of attempting to avoid the curse of dimensionality byobserving a lower-dimensional feature subspace, we embrace dimensionality by taking advantageof some inherently high-dimensional phenomena. More specifically, we show that hubness, i.e.,the tendency of high-dimensional data to contain points (hubs) that frequently occur in k-nearestneighbor lists of other points, can be successfully exploited in clustering. So, we use Hub conceptwith multiple feature extraction for image clustering.Image segmentation is the classification of an image into different groups.

**KEYWORDS**: Clustering, Hubness, Hub score.

## I. INTRODUCTION

High dimensional data processing is always challenging task, since data become sparse as increase in dimensionality, this cause uneven data processing. Most real time data sets are high-dimensional, but traditional algorithms are designed to work on lower dimensionality. Also, efficiency and performance of traditional algorithms get reduce as increase in dimensionality.

Hubness is a recently described aspect of the well-known curse of dimensionality. It isan intrinsic property of high-dimensional data, where hubs emerge as centers of inuence ink-nearest neighbor (KNN) topologies. They often exhibit a detrimental inuence on the learningprocess. As most data of practical concern is high-dimensional, this is an important issue. High-dimensional data are by their very nature often difficult to handle by conventional machinelearning algorithms, which is usually characterized as an aspect of the curse of dimensionality.However, it was shown that some of the arising high-dimensional phenomena can be exploitedto increase algorithm accuracy.

High dimensional data are ubiquitous in modern applications. They arise naturally when dealing with text, images, audio, data streams, medical records, etc. The impact of this highdimensionality is many fold. It is a well-known fact that many machine-learning algorithmsare plagued by what is usually termed the curse of dimensionality. This comprises a set of

Properties which tend to become more pronounced as the dimensionality of the data increases.First and foremost is an unavoidable sparsity of data. In higher-dimensional spaces all datais sparse, meaning that there is not enough data to make reliable density estimates. Anothermitigating inuence comes from the concentration of distances, which has been thoroughlyexplored in the past. Namely, all the distances between data points generated from thesame distribution tend to become increasingly more similar to one another as new dimensionsare added. Luckily, this does not affect multiple-distribution data as much. The questionof whether the very concept of nearest neighbors is meaningful in high-dimensional data sets Admittedly, there are some difficulties, but nearest neighbor methods remain popular, bothfor

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 5, Issue 12, December 2016**

classification and clustering. To deal with these problems regarding high dimensional data clustering, we are used concept
of Hub. Hubs are the data points in high dimensional data having tendency of more frequentlyoccurrence of KNN list of other data points. This property can be used in clustering processof high dimensional data.Clustering of the images is one of the interested area in data mining and image processing.Traditional algorithms of image clustering propose algorithm with lower dimensional feature i.e.they consider less number of features which will become difficult task and inefficient algorithm.So, we use Hub concept with multiple feature extraction for image clustering.
Clustering is nothing but Finding groups of objects such that the objects in a group willbe similar to one another and different from the objects in other groups. Three types ofclustering techniques are used. Unsupervised, semi unsupervised, supervised clustering. In case ofunsupervised clustering all the data points are known. So very to cluster.
In case of supervisedclustering data points are not known but total supervision is required for clustering of datapoints.
Paper is organized as follows. Section II describes hubness concept and how to use it in clustering also why clustering of high dimensional data is difficult one. Flow of the system and algorithms are given in Section III. Section IV presents experimental results analysis. Finally, Section V presents conclusion.

## II. RELATED WORK

Clustering is a process of grouping similar data elements together so that they possess similarfeature 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 theentire image set collection The features on which clustering being performed is depends on datatype. This feature acts like dimension of that data. Several application domains such asmolecular biology and geography produce a tremendous amount of data which can no longer be managed without the help of efficient and effective data mining methods. However, traditional clustering algorithms often fail to detect meaningful clusters because most real-world data setsare characterized by a high dimensional, inherently sparse data space.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 featurespace. This comprises a set of properties which tends to become more pronounced as data dimensionality increases. The main cause of all is unavoidable sparseness of data. In highdimensionality, 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. This hardnessincreases with dimensionality since the distance between two points in space become larger andlarger. The data domains like images are most interested data for clustering purpose. Also, imageclustering is essential for purposes like image retrieval, image classification, object detectionetc. The local features are of intermediate complexity, which means that they are distinctive enough to determine likely matches in a large database of features. 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 algorithmto clustering of high dimensional data like image data. An image data will be represented usingits different features. These features usually contain information extracted from color, texture, edges and any property which we feel important for image clustering. The methods like k-meanand KNN are much powerful and widely used in classification methods. But these methodsreduce their performance as increase in dimensionality of data. There are some more difficultiesin dealing with high-dimensional data are omnipresent and abundant. Same as other thesetwo is also due to sparseness of data point. Hubs appear as a consequence of the geometry ofhigh-dimensional space, and the behavior of data distributions within them. Hubness is thetendency of some data points in high-dimensional data sets to occur much more frequently ink-nearest-neighbor lists of other points than the rest of the points from the set, can in fact beused for clustering. It is a high dimensional phenomenon which concern k-nearest-neighborset. Denoted by Nk(x) the number of k occurrences of x i.e. the number of times x appears ink-nearest neighbor list of other points in data. The distribution of Nk (x) exhibits significantskew in high dimensional cases, skew which increases with intrinsic dimensionality of the data.This leads to

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

the emergence of hubs, inuential points which affect the reasoning procedureof nearest-neighbor based methods for many data points. Fig. shows illustrative example bywhich we get idea of Hub in data.
The red dashed circle marks the centroid (C), yellow dotted circle the medoid (M), and greencircles denote two elements of highest hubness (H1;H2), for neighborhood size 3. Through imageclustering, we aim at developing techniques that support effective searching and browsing oflarge image digital libraries based on automatically derived image features. Recent methodsprovide 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 fromthe query image in terms of semantics due to the semantic gap. We will tackle the problemby using hub-based high-dimensional image clustering technique.

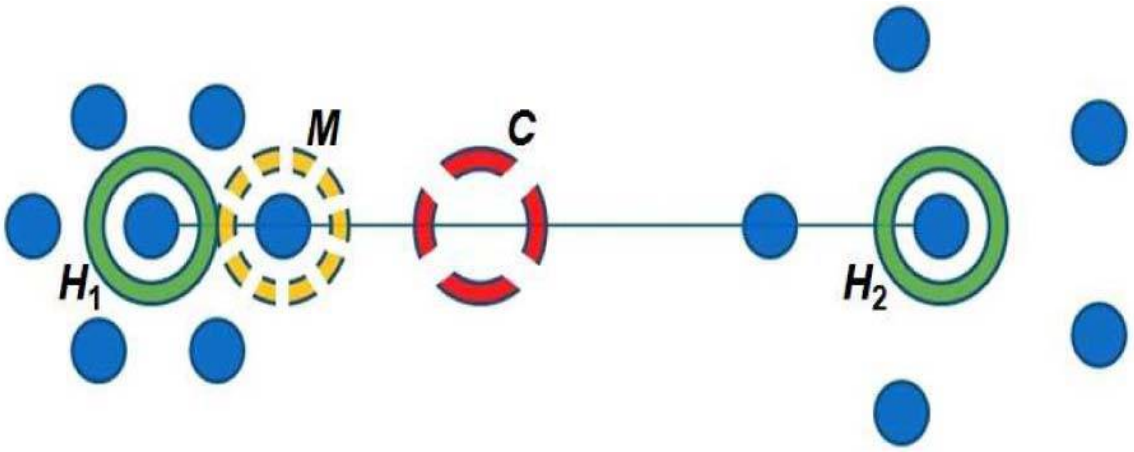



Fig (a.)Hubness

## III. FLOW AND ALGORITHM

In this proposed technique after loading of data base images clusters are formed by applyingfeature extraction followed by HPC

- 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
- 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.
- 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 findout the center images of these clusters so that query images features are only compared to cluster as center images, which cluster as center image having more similarity with queryimage, that whole cluster images are compared to query image. So there is no need tocompare query image to database as all images.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

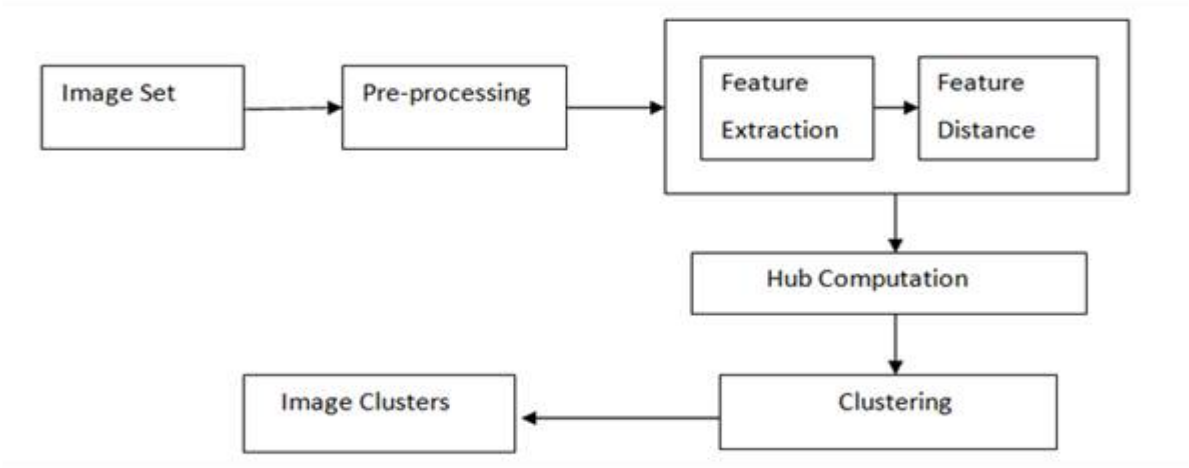



Fig (b) Flow of system

**Algorithm:**

Algorithm: Histogram Feature Extraction

Input: img: Image
Output: histogram [ ];
1. Initialize Histogram []:=0;
2. for each pixel of image
Follow step 3-4
3. C: = color value and pixel;
4. Histogram [c] ++;

Algorithm: Hubness Computation
Input: Data File
Output: hubscores [ ];
1. Load Data file
2. Initialize DistanceMatrix [] []:=0;
3. for each datapoint1 from Data
follow step 4
4. for each datapoint2 from Data
follow step 5
5. DistanceMatrix [] []:= ComputeDistance (dataPoint1, dataPoint2)
end loop
end loop
6. DistanceMatrix: = Sort Distance Matrix in descending order
7. HubnessScore []:= 0;
8. for each dataPoint from Data
follow step 9
9. HubnessScore: = count HubnessScore (dataPoint);
return HubnessScore;

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 5, Issue 12, December 2016

Algorithm: Compute Distance (dataPoint1, dataPoint2)
1. Set Distance: = 0;
2. for each Dimension d1 and d2 from dataPoint1 and dataPoint2
follow step3
3. Distance =Distance + (d1 +d2) ^2;
end loop
4. Distance= $\sqrt{Distance}$
5. return Distance;

Algorithm: Count HubnessScore (dataPoint)
1. Set Score: = 0;
2. for each dataPoint in Data
follow step 3
3. if dataPoint is in KNN list of dataPoint then
score++;
end if
end loop
4. return score;

## IV. EXPERIMENTAL RESULTS

| Data set. | No of classes | No of features | No of examples |
|---|---|---|---|
| digit-369 | 3 | 16 | 3185 |
| Ecoli | 5 | 7 | 327 |
| Glass | 6 | 9 | 214 |

Table(a): Characteristics of low dimensional data.

In this paper, we cluster the high dimensional data. As we know that high dimensional datameans data with large number of features. The traditional algorithm work on low features, butnot work well with high features. But nowadays most of real time data is high dimensional.i.e. images, audio, video.
This algorithm work well on image dataset i.e. data contains high features. Followingtable shows characteristics of low dimension and high dimension data set.

| Data set. | No of classes | No of features | No of examples |
|---|---|---|---|
| Image-1 | 3 | 256 | 156 |
| Image-2 | 2 | 256 | 92 |

Table(b):Characteristics of high dimensional data.

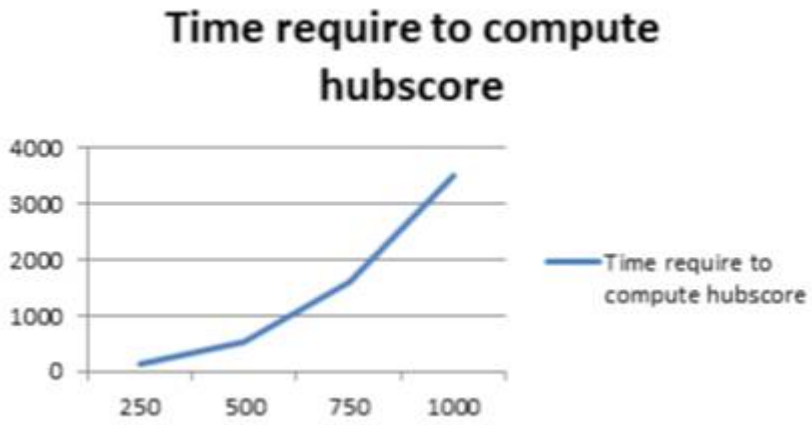
For our experiments we assume images are data points with 256 dimensions represent 256bean histogram of image. Each image from data set resized to same size for make ease ofanalysis. Then we analyze time require to compute hubness score of each data point for varyingdata size and varying image size (width and height). In this whole experiment, the time requirefor feature extraction will be ignore since for a particular image size feature extraction time will remain approximately same. Once features get extracted we store it into a excel file. This excel file will used for furthercomputing process.

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 5, Issue 12, December 2016**




Fig(c): Number of images to time

The following Table represents the relation between numbers of imagestaken as input and time (in Millis) to execute full process of hub computation from computingdistance matrix to computing .The number of dimensions used in this experiment is 256 beans histogram. And number ofimages varies from 250 to 1000 images. In this chart we can see that effect of number of datapoints in data set on time to compute k hub points or compute hubness score of data points.

As increase in data point time to compute hubness score is also increases rapidly.

| Number of Images | Time To Compute Hub For 256 Dimensions |
|---|---|
| 250 | 136 |
| 500 | 536 |
| 750 | 1594 |
| 1000 | 3482 |

Table(c): Number images to time for 256 images

## V. CONCLUSION

Finally, we conclude that it is easy to use hub concept for clustering of high dimensional datalike image. In clustering process hub plays role of centroid or medoid of other algorithms. Thisis the new method which improves algorithm performance and execution time.Hub concept is mainly designed for high dimensional data type like image. These data typescreate large empty space between data points. So, only hubs can able to find Centre of data.In case of centroid based approaches, this task is inconvenient.

## VI. ACKNOWLEDGMENT

We would like to thank IJIRSET for giving such wonderful and useful platform for the PG students to publish their research work. Also would like to thanks to our guide & respected teachers for their constant support and motivation for us. Our sincere thanks to SKN Sinhgad Institute of Technology and Science for providing a strong platform to develop our skill and capabilities.

## REFERENCES

[1] Sicheng Xiong, Javad Azimi, and Xiaoli Z. Fern," Active Learning of Constraints for Semi-Supervised Clustering"
[2] J.Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong, Locality-constrained linear coding for image classification, in Proc. IEEE Conf.Comput. Vis. Pattern Recognition., Jun. 2010,

ISSN(Online) : 2319-8753
ISSN (Print) : 2347-6710

International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 5, Issue 12, December 2016**

[3] S. Basu, A. Banerjee, and R. Mooney, Active Semi- Supervision for Pairwise ConstrainedClustering, Proc. SIAM Intl Conf. Data Mining, pp. 333-344, 2004.
[4] Nenad Tomasev , Milo Radovanovi , Dunja Mladeni , and Mirjana Ivanovi ,The Role ofHubness in Clustering High-Dimensional Data
[5] R. Huang and W. Lam, Semi-Supervised Document Clustering via Active Learning withPairwise Constraints, Proc. Intl Conf.Date Mining, pp. 517-522, 2007
[6] D. Cohn, Z. Ghahramani, and M. Jordan, Active Learning with Statistical Models, Artificial Intelligence Research, vol. 4, pp. 129-145, 1996.
[7] L. Breiman, Random Forests, Machine learning, vol. 45, no. 1, pp. 5-32, 2001
[8] C. Galleguillos, A. Rabinovich, and S. Belongie, "Object categorization using co-occurrence, location and appearance," in *Proc. IEEE Conf.Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.
[9] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 35–367, Feb. 2011.
[10] P. Perona and J. Malik, "Scale-space and edge detection using an isotropic diffusion," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 12, no. 7, pp. 629–639, Jul. 1990.
[11] M. M. Cheng, G. X. Zhang, N. J. Mitra, X. L. Huang, and S. M. Hu, "Global contrast based salient region detection," in *Proc. IEEE Conf.Comput. Vis. Pattern Recognit.*, Dec. 2011, pp. 409–416.
[12] T. Judd, K. A. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, Feb. 2009, pp. 2106–2113.