Indian Journals of Science and Technology, Vol 9(19), DOI: 10.17485/ijst/2016/v9i19/92034, May 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

Dimensionality Reduction based on Hubness Property using Feature Weighting Method for Clustering

A. Jenneth[1] and K. Thangavel[2]

[1]Computer Science Department, Sri Krishna Arts and Science College, Sugunapuram, Kuniamuthur, Coimbatore - 641008, Tamil Nadu, India; jenneth_ismail@yahoo.com
[2]Department of Computer Science, Periyar University, Periyar Palkalai Nagar, Salem - 636011, Tamil Nadu, India; drktvelu@yahoo.com

Abstract

Grouping of high dimensional information is an imperative exploration subject in the information mining, in light of the fact that the genuine datasets frequently have high dimensional components. The objective of the clustering is to group the features which should be similar to each other. Many text mining approaches are optimized to mine the sparse data which incurs high computation cost. In this paper, we process a novel technique named as affine subspace clustering which incorporates the Hubness property to handle the local feature relevance value and Curse of dimensionality. The Hubness property reduces the discrimination problem in the cluster formation and used as clustering method with effects relevant to cluster structures. Rather than endeavoring to keep away from the scourge of dimensionality by watching a lower dimensional component subspace, we use substantial dimensionality by exploiting downward closure property and outlier detection in the k nearest neighbor list. Additionally we combine Feature weighting method to minimize the average inside cluster scattering and augment the average between cluster scatterings along all the element spaces. The experimental results prove that proposed system yields the good performance in numerous settings, especially within the sight of huge amounts of commotion. The proposed techniques are optimized for the most part to detect the cluster center accuracy and extended properly to handle clusters of random sizes. Average inside cluster scattering is minimized and average between-cluster scattering is expanded along all the element spaces.

Keywords: Clustering, Curse of Dimensionality, Dynamic Centroid

1. Introduction

The objective of the Clustering is to establish the useful groups of similar objects in the high dimensional information. In general High Dimensional information arises normally in numerous areas and poses substantial difficulties in the conventional clustering algorithms, both as far as proficiency and effectiveness[1]. Clustering the high dimensional data is difficult task, various clump algorithms are projected, which might be classified into four groups: partitioned off, hierarchical, density primarily based, and mathematical space primarily based algorithms. Mathematical space clump formula[2,3] works by establishing a random bunches in some glower dimensional expulsion of the first knowledge, and are usually most well-liked once addressing knowledge that square measure high dimensional[4,5]. This is primarily attributable to 2 persistent impacts: the unfilled house development and convergence of separations. The previous alludes to the actual certainty that everyone eminent dimensional knowledge sets have a tendency to be slight, as a consequence of the amount of focuses expected to speak to any dissemination becomes exponentially with the amount of measurements. This ends up in dangerous thickness gauges

*Author for correspondence

for high-dimensional knowledge, inflicting challenges for thickness based methodologies in terms of curse of spatiality. The concentration of the gap is critical attribute of high dimensional knowledge representations separations between knowledgetend to become tougher to differentiate the information into cluster as spatiality will increase, which might cause issues with distance-based algorithms[6-9]. The troubles in managing the high dimensional learning zone are universal and proliferating. Be that as it may, not all wonders emerge in the area unit essentially damaging to cluster strategies. We'll demonstrate during this paper hubness, that will be that the propensity of some information focuses in high-dimensional learning sets to happen much all the more frequently in K-closest neighbour arrangements of option focuses than the rest of the focuses from the set, will if frankly be utilized for cluster. This has not been antecedent tried to the simplest of our information. In an exceedingly restricted sense, hubs in graphs are accustomed represent typical word meanings in[10], that weren't used for knowledge cluster. An identical line of analysis has known essential super molecules as hubs within inside the converse closest neighbour topology of protein collaboration networks[11]. We have focused on investigating the capability of victimization centre focuses in cluster by coming up with hubness-mindful cluster calculations and testing them in an exceedingly high-dimensional settings. In addition, we have a tendency to propose 3 new bunch algorithms and valuate their execution in numerous high-dimensional group undertakings. We have a tendency to compare the algorithms with a baseline progressive model based technique (K-means[12]), yet as thickness based methodologies. The analysis demonstrates that projected calculations often provide enhancements in cluster quality and homogeneity. The correlation with kernel K means[13] uncovers that part based expansions of the underlying methodologies ought to even be thought-about within the future. Our present centre was absolutely on appropriately picking bunch models, with the anticipated routes streamlined for investigator work near group focuses. The rest of the paper is sorted out as takes after: In section-2, we have shown the related work about the Hubness based grouping. Segment 3 talks about proposed framework Feature weighting, while Section 4 investigations the execution of the framework. Ends with conclusion at Section 5.

2. Related Works

2.1 Hubness based Clustering

Hubness has as of late been set up as a majority property of K-closest neighbour (K-NN) charts acquired from high-dimensional data utilizing a separation live, with attributes and impacts significant to the group structure of information, also as bunch calculations. The Hubness property is showed with expanding (inborn) information spatial property. The appropriation of information purpose in-degrees, i.e. the measure of times focuses appear among the k closest neighbours of option focuses inside the information, turns out to be extremely inclined.

2.2 Density based Clustering

Density based agglomeration ways typically believe this type of density estimation[14-16]. The density based algorithm is based on the implicit assumption that clusters are formed by separating high-thickness locales from one another by low-thickness areas.

In high-dimensional zones this is normally frequently extreme to assess, on account of data being terribly distributed. The problem of selecting the right neighbourhood size occurs conjointly at this point, subsequent to each modest and tremendous estimations of k will bring about issues for thickness basically based approaches[17]. Imposing k-closest neighbour consistency in calculations like K-Means was conjointly explored[18]. The foremost of the mill utilization of k-closest neighbour records; however is to build a K-NN graph[19] and downsize the matter to it of chart agglomeration. Results and uses of Hubness are additional completely investigated in different connected fields: classification[20-23] and data reduction[23]. In this paper, we have introduced the feature weighting technique to interface with the methodology of utilizing center points as group models and/or managing focuses amid model pursuit.

3. Proposed Model

Because of the illustrated challenges with applying thickness based and remove based grouping approaches in the high-dimensional case, an alternate class of strategies is normally utilized for high-dimensional information bunching. Here the notion is to study the group on a

reduced dimensional complex and to consequently identify a legitimate projection of the actual information.

3.1 Hubness Process

Hubness is a side of the scourge of spatial property touching on nearest neighbors that has recently involves attention, in contrast to the abundant mentioned distance concentration development. It was resolved that Hubs won't not group well utilizing exploitation typical model based agglomeration calculations, since they not exclusively have a tendency to be near points happiness to constant cluster (i.e., have low intra cluster distance) however conjointly have a tendency to be close directs appointed toward various bunches (low between group separation). Thus Hubs can be seen as (restricting) counterparts of exceptions that have high between group separation and in addition high intra bunch separation.

3.2 Outlier Detection in Cluster Formation

The convergence of separations empowers one to see unimodal high-dimensional information as lying around on a hyper sphere focused at the information dispersion mean. A low-hubness score demonstrates that some extent is on the normal unapproachable from the rest of the focuses related in this manner no doubt an anomaly. In high-dimensional territories, be that as it may, low-hubness parts are expected to happen by the frightfully way of those regions and knowledge appropriations. These learning focuses can bring about a mean increment in intra cluster separation. It completely was conjointly appeared for some agglomeration calculations that centers don't bunch very much contrasted with the rest of the focuses. This is frequently inferable from the genuine certainty that a few center points are actually close focuses in a few bunches. Subsequently, they cause a reduction in intercluster separation.

3.3 Centroid Selection for Outlier Elimination

Centroids rely on upon all present bunch parts, while center points depend absolutely on their close segments and, accordingly, convey restricted position information. We are going to mull over 2 styles of hubness underneath, to be specific worldwide hubness and neighborhood hubness. By and large, there are two sorts of subspace grouping approaches – those that attempt to locate a genuine formal component sub space, and those that reproduce the procedure via naturally doling out weights to highlights keeping in mind the end goal to build the impact of certain elements on the nearness measure and reduction the impact of others. We characterize neighborhood Hubness as a confinement of worldwide hubness on any given group, considered with regards to the present calculation cycle

3.4 Feature Weighting (Figure 1)

In general, Hubs emerge near centers of dense sub regions may recommend some type of a relationship amongst Hubness and therefore the thickness gauge at the decided data point. Marking noise influences the accuracy of the classification. One probable reason is that some viable components that ought to be given high weights are

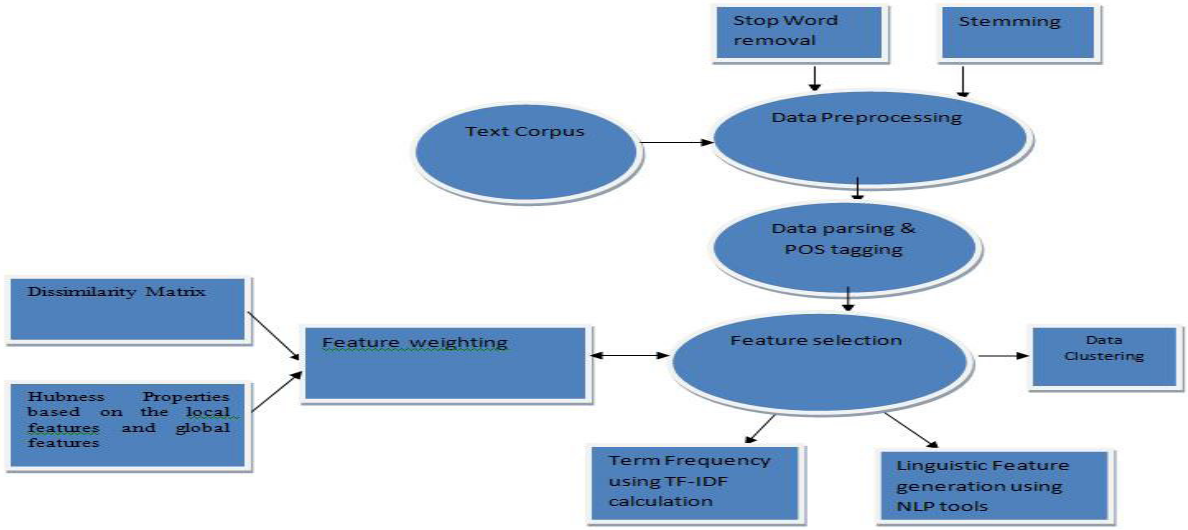

Figure 1. Feature weighting method for outlier elimination.

inhibited in the preparation stage because of the labeling mistakes. We grow computationally shoddy component weighting systems to neutralize such impact by propelling the heaviness of discriminative elements, so they would not be quelled and the examples with such structures would have higher opportunity to be accurately arranged. A basic approach to utilize center points for bunching is to utilize them as one would regularly utilize centroids. There are two principle objectives of building up this component weighting: (1) accurately anticipating the marks of information focuses and positioning them in light of forecast certainty, so that the in all probability blunders can be viably distinguished; (2) requesting a littler sum time on preparing, so that the spared time can be spent on redressing all the more naming mistakes. Along these lines we mean to manufacture a group that is both precise and time productive in terms of eliminating the outlier.

Algorithm – Affine Subspace Clustering

Initialize the cluster centre ()
Form clusters () based on cluster centre
For all data points
Set feature weights for each cluster
Normalize the features
Then form cluster based on the features weights

Specifically, there exist several knowledge focuses having low hubness scores making them unfortunate contender for bunch centers. Such focuses can have an occasional likelihood of being designated. As to high light this more, we prefer to use the sq. of the particular Hubness score rather than creating the possibilities specifically relative to Nk(x).

4. Experimental Analysis

We tested our methodology on different high-dimensional manufactured and certifiable information sets. There is no well-known understood strategy for picking the most straightforward K for finding neighbor sets, the matter being space particular and high dimensional. To see however the determination of K reflects on Hubness property in the feature weighting technique, we conduct an experiment with 500 text corpus to establish a clustering.

This section can discuss the rationale why feature weight rule can give higher performance contrasted with K-Means regarding intra-and intercluster distance expressed by the silhouette index.

We observe the intra and inter parts of the silhouette index, and compute a (dissimilarity with all other information within same cluster) and b (the most reduced normal disparityto any other cluster), and thereby arriving the silhouette index on a given information set. The model's capacity to separate at the component level can be further supported by utilizing the dispersal of highlight weights over various classes. The refinement of different classes can be utilized to further drive highlight bias scores separated to enhance the distinguishing proof of class particular elements within the sight of naming mistakes. Let nh be the of hubs designated. Next, we have a tendency to choose as outliers the NH points with the bottom events. At long last, we have a tendency to choose all remaining points as "regular" points. Figure 2 illustrates the line at the break-up of the silhouette index on the five hundred file as text corpus (we have detected similar trends with all alternative knowledge sets. It is seen that every one other data sets. It can be seen that all clustering methods perform roughly.

Because of the numerous lop-sidedness of the square hubness scores, including extra probabilistic iterations helps in achieving higher agglomeration, up to an explicit upland that's eventually reached. An equivalent form of the curve conjointly seems within the case of not taking the last, however the blunder minimizing design. Hubs usually have low b-values that cause them to cluster badly and negatively affect the cluster strategy. It absolutely was steered that they must be treated virtually as outliers. That's why it's encouraging to visualize that the projected clump ways cause clump configurations, wherever hubs have higher b-values than inside the instance of K-Means.

4.1 Silhouette Index

Silhouette analysis may be wont to study the separation distance between the ensuing clusters. The silhouette plot displays a live off however shut every purpose in one cluster is to points within the neighbor clusters and therefore provides some way to assess parameters like range of clusters visually. This live includes a vary of [-1, 1]

$$\text{Silhouette index s (i)} = \frac{\text{b(i)} - \text{a(i)}}{msx(a(i), b(i)}$$

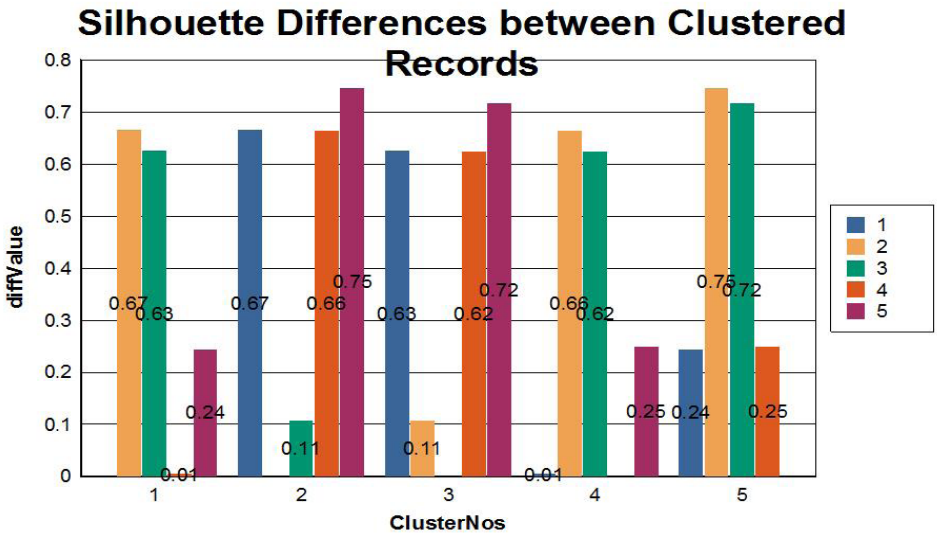

Figure 2. Silhouette difference between clustered records.

Table 1. Silhouette index for K-Means clustering

Cluster no 1	Cluster no 2	Difference Value
1	2	0.01108
1	3	0.00560
1	4	0.01659
1	5	0.01650
2	1	0.01108
2	3	0.00551
2	4	0.00558
2	5	0.00548
3	1	0.00560
3	2	0.00551
3	4	0.01106
3	5	0.01096
4	1	0.01659

Silhouette Index Difference between Clusters (feature weighting)

Where b (i) be the most reduced normal disparity of I to any other cluster and a (i) be the average dissimilarity of with all other information within same cluster.

The Silhouette index for high dimensional data clustering using K-Means and feature weighting algorithm is explained in Table 1 and Table 2.

Table 2. Silhouette index for Feature weighting method.

Cluster no 1	Cluster no 2	Difference Value
1	2	0.47163
1	3	0.38969
1	4	0.14330
1	5	0.50231
2	1	0.47163
2	3	0.13426
2	4	0.38324
2	5	0.73703
3	1	0.38969
3	2	0.13426
3	4	0.28760
3	5	0.69625
4	1	0.14330

Silhouette Index Difference between Cluster (K-Means Clustering)

The improvements stem from a superior situation of center point focuses into bunches, which helps in expanding the between-group separation. Hence, it turns out to be more convenientto recognize close and far off focuses and to legitimately identify bunch limits.

5. Conclusion

In this work, we designed and implemented Feature weighting technique for data clustering in the subspace of the initial cluster. Initial Clustering is carried out with K-Means but which directed us to subspace formation due to curse of dimensionality. We have shown that mistreatment hubs to approximate native knowledge centers aren't solely a possible possibility, however additionally oft ends up in improvement over the centroid-based approach. The projected Feature coefficient technique had proved to be additional durable than the K-Means on each artificial and true learning knowledge, still as within the nearness of large amountsof by artificial means introduced noise. This first analysis suggests that mistreatment hubs each as cluster models focuses directing the centroid-based hunt might be a promising new arrangement in agglomeration high-dimensional and shouting information. Also, international hubness estimates are usually to be most popular with relevance the native ones. Hub-based calculations arecomposed particularly for prime dimensional knowledge. This is often an uncommon property, since the performance of most traditional agglomeration calculations disintegrates with a rise of spatial property. Hubness, on the opposite hand, may be a property of in and of itself high-dimensional knowledge and are required to supply change by giving higher inter cluster distance, i.e., higher cluster partition.

6. Reference

1. J. Han and M. Kamber. Morgan Kaufmann: Data Mining: Concepts and Techniques, Second ed. 2006.
2. Aggarwal CC, Yu PS. Finding Generalized Projected Clusters in High Dimensional Spaces. Proc. 26th ACM SIGMOD Int'l Conf. Management of Data. 2000; p. 70-81.
3. Kailing K, Kriegel HP, Kroger P, Wanka S. Ranking Interesting Subspaces for Clustering High Dimensional Data. Proc. Seventh European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD). 2003; p. 241-52.
4. Kailing K, Kriegel HP, Kroger P. Density-Connected Subspace Clustering for High-Dimensional Data. Proc. Fourth SIAM Int'l Conf. Data Mining (SDM), 2004; p. 246-57.
5. Muller E, Gunnemann S, Assent I, Seidl T. Evaluating Clustering in Subspace Projections of High Dimensional Data. Proc. VLDB Endowment. 2009; 2:1270-81.
6. Aggarwal CC, Hinneburg A, Keim DA. On the Surprising Behavior of Distance Metrics in High Dimensional Spaces. Proc. Eighth Int'l Conf. Database Theory (ICDT). 2001; p. 420-34.
7. Francois D, Wertz V, Verleysen M. The Concentration of Fractional Distances. IEEE Trans. Knowledge and Data Eng. 2007 July; 19(7):873-86.
8. Durrant RJ, Kaban A. When Is 'Nearest Neighbour' Meaningful: A Converse Theorem and Implications. J. Complexity. 2009; 25(4):385-97.
9. Kaban A. Non-Parametric Detection of Meaningless Distances in High Dimensional Data. Statistics and Computing. 2012; 22(2):375-85.
10. Agirre E, Martınez D, de Lacalle OL, Soroa A. Two Graph-Based Algorithms for State-of-the-Art WSD. Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP). 2006; p. 585-93.
11. Ning K, Ng H, Srihari S, Leong H, Nesvizhskii A. Examination of the Relationship between Essential Genes in PPI Network and Hub Proteins in Reverse Nearest Neighbor Topology. BMC Bioinformatics. 2010; 11:1-14.
12. Arthur D, Vassilvitskii S. K-Means++: The Advantages of Careful Seeding, Proc. 18th Ann. ACM-SIAM Symp. Discrete Algorithms (SODA). 2007; p. 1027-35.
13. Dhillon IS, Guan Y, Kulis B. Kernel K-Means: Spectral Clustering and Normalized Cuts. Proc. 10th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining. 2004; p. 551-56.
14. Tran TN, Wehrens R, Buydens LMC Knn. Density-Based Clustering for High Dimensional Multispectral Images. Proc. Second GRSS/ISPRS Joint Workshop Remote Sensing and Data Fusion over Urban Areas. 2003; p. 147-51.
15. Bicici E, Yuret D. Locally Scaled Density Based Clustering. Proc. Eighth Int'l Conf. Adaptive and Natural Computing Algorithms (ICANNGA), Part I. 2007; p. 739-48.
16. Zhang C, Zhang X, Zhang MQ, Li Y. Neighbor Number, Valley Seeking and Clustering. Pattern Recognition Letters. 2007; 28(2):173-80.
17. Hader S, Hamprecht FA. Efficient Density Clustering Using Basin Spanning Trees. Proc. 26th Ann. Conf. Gesellschaftfur Klassifikation. 2003; p. 39-48.
18. Ding C, He X. K-Nearest-Neighbor Consistency in Data Clustering: Incorporating Local Information into Global Optimization. Proc. ACM Symp. Applied Computing (SAC). 2004; p. 584-89.5
19. Chang CT, Lai JZC, Jeng MD. Fast Agglomerative Clustering Using Information of k-Nearest Neighbors. Pattern Recognition. 2010; 43(12):3958-68.
20. Tomasev N, Radovanovic M, Mladenic D, Ivanovic M. Hubness-Based Fuzzy Measures for High-Dimensional k-Nearest Neighbor Classification. Proc. Seventh Int'l

Conf. Machine Learning and Data Mining (MLDM). 2011; p. 16-30.
21. Tomasev N, Radovanovic M, Mladenic D, Ivanovic M. A Probabilistic Approach to Nearest-Neighbor Classification: Naive Hubness Bayesian kNN. Proc. 20th ACM Int'l Conf. Information and Knowledge Management (CIKM). 2011; p. 2173-76.
22. Radovanovic M, Nanopoulos A, Ivanovic M. Time-Series Classification in Many Intrinsic Dimensions. Proc. 10th SIAM Int'l Conf. Data Mining (SDM). 2010; p. 677-88.
23. Radovanovic M, Nanopoulos A, Ivanovic M. Hubs in Space: Popular Nearest Neighbours in High-Dimensional Data. J. Machine Learning Research. 2010; 11:2487-531.