K-Means Based Clustering In High Dimensional Data

Sindhupriya. R, Ignatius Selvarani. X

Abstract— **There is an increase in the high dimensional data naturally in many domains and usually introduce a great challenge for traditional data mining techniques in terms of effectiveness and efficiency. The main difficult in clustering is due to increase sparsity of such data, as well as increase difficult in distinguishing distances between data points. In this paper they take a novel perspective on the problem of clustering high-dimensional data. As an alternative attempting to avoid the curse of dimensionality by observing a lower-dimensional feature subspace, they embrace dimensionality by taking advantage of inherently high-dimensional phenomena. They show that hubness is the tendency of high-dimensional data to contain points that frequently occur in k-nearest neighbor lists of further points, can exploited in clustering. They check our hypothesis by demonstrating that hubness is a good measure of point centrality with a high-dimensional data cluster, and suggest 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. Based on the Experimental method demonstrate good performance of our algorithms in various position, particularly in the large quantities of noise. They suggest this technique is mostly tailored for detecting approximately hyper spherical clusters and need to be extended in order to properly handle clusters of arbitrary shapes.**

Index Terms— curse of dimensionality, hub based clustering, nearest neighbor

INTRODUCTION

Clustering is the process of grouping objects into classes of same objects. The objective of clustering is to determine the intrinsic grouping in a set of unlabeled data. There are four groups in clustering algorithms namely separation, hierarchical, density based, and subspace. The incentive for this partiality that observation having more dimensions usually leads to the so-called curse of dimensionality this showing many standard machine-learning algorithms becomes impaired. This is due to two common effects: the empty space phenomenon and concentration of distances. High dimensional data sets show a tendency of sparse, because the number of points required to represent any distribution grows exponentially with the number of dimensions. This shows a bad density approximate for high-dimensional data and to create strain for density-based approaches and concluding counterintuitive property of high-dimensional data descript all distances between data points become harder to discriminate as dimensionality increases, this can leads to problems with distance-based algorithm. The problem of high dimensional data is omnipresent and abundant. In this paper that 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

The proposed concept in this paper is hubness based clustering. If hubness is observed as a kind of local centrality measure, it may be able to done hubness for clustering in many ways. In order to test this theory, 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

I. RELATED WORK

In this section, we discuss related works on clustering high dimensional data. **[1]**High dimensional data is an challenge for clustering algorithms because of the implicit sparsity of the points. They produce a new concept of using extended cluster feature vectors to make the algorithm scalable for high databases. The lengths of the time and space requirements of the algorithm are adjustable, and are possibly to trade off with better validity. The problem of clustering data points is examine by a set of points in multidimensional space and a separation of the points into clusters so that the points within every cluster are near to one another. In this paper, they survey very general concept of using arbitrarily projected subspaces for searching clusters. Concentrated dimensional clusters cannot be found in very high dimensional data sets, this technique finds hidden subspaces in which points cluster able because of correlations between the dimensions. The goal of eliminating *the sparse* subspaces for every cluster, and projecting the points into those subspaces in which the greatest similarity occurs is a much generalized notion of clustering of which the full dimensional case is a special one. In future research shows

Manuscript received Jan, 2014.

Sindhupriya R*, Department of Computer Science and Engineering, Sri Vidya College of Engineering and Technology, Virudhunagar, India.*

Ignatius Selvarani. X, *Department of Computer Science and Engineering, Sri Vidya College of Engineering and Technology, Virudhunagar, India,*

how to use this technique for effective high dimensional data visualization. [2] An application domains like molecular biology and geography process a tremendous amount of data which can no longer without the help of efficient and effective data mining methods. They introduce a pre-processing step for traditional clustering algorithms, discover all interesting sub- spaces of high-dimensional data containing clusters quality criterion for the interestingness of a subspace and pro- pose an efficient algorithm called RIS (Ranking Interesting Subspaces) to examine all such subspaces.

In general, most of the clustering algorithms cannot create correct results because of the inherent sparsity of the data space. High dimensional data does not cluster large distance. But clusters in lower dimensional subspaces are easily use. In this paper, they present a preprocessing step for clustering high-dimensional data. Based on a quality criterion for the interestingness of a subspace, they presented an efficient algorithm to compute all interesting subspaces containing dense regions of arbitrary shape and size called RIS. Technique of random sampling can be applied to RIS in order to boost-up the runtime of the algorithm significantly with a small quality of loss. Analysis shows that RIS can be successfully applied to high-dimensional real-world data, e.g. gene expression data in order to find co-regulated genes .[3] Many application domains like molecular biology and geography generate a tremendous amount of data which can no longer without the help of efficient and effective data mining methods. Karin kailing et al present SUBCLU (density-connected Subspace Clustering), an effective and efficient approach to the subspace clustering problem.

By using the concept of density-connectivity the algorithm DBSCAN, SUBCLU is based on a formal clustering notion, different to existing grid-based approaches, SUBCLU discover arbitrarily shaped and positioned clusters in subspaces. First approaches to subspace clustering is CLIQUE. CLIQUE is a grid-based algorithm. The curse of dimensionality for data mining tasks like clustering methods to decrease the dimensionality of the data space. Dimensionality reduction has some disadvantage: First, the change attributes have no intuitive meaning. The resulting clusters are hard to interpret. Second, dimensionality reduction does not admit the desired results (e.g. [AGGR98] present an example where PCA/KLT does not reduce the dimensionality). Third, using dimensionality reduction method, the data is clustered only in a certain subspace. The clustered object dissimilar in varying subspaces is lost. The concept of density-connectivity introduce in to the subspace clustering problem. This has some advantages:

- SUBCLU detects arbitrarily shaped and positioned clusters in subspaces.
- Different to CLIQUE and its successors, the clustering concept is well defined.
- SUBCLU do not use pruning heuristics such as CLIQUE, it gives each subspace the same clusters as if DBSCAN is applied to this subspace. The future work is the development of a logical index structure for limited range queries (range queries in arbitrary subspaces of the original data space). The inverted files uses SUBCLU are less powerful the more dimensions are applicable to the range query, a finer index support improve the efficiency of SUBCLU [4] Clustering high dimensional data is appear to research field. Analysis and contrast is challenged by three problems. First, No ground truth that relate the "true" clusters in real world data. Second, a massive diversity of different measures is used that reflect evaluation aspects of the clustering result. In this paper, they introduce a systematic approach to evaluate the great paradigms in a usual framework. Clustering is a normal data mining task for unmanned grouping object [14].

Cluster recognition is based on similarity between objects and distance functions. In high dimensional spaces, effects attributed to the "curse of dimensionality" are known to break traditional clustering algorithms [9].Three problems persist. First, No ground truth that relate the true clusters in real world data. Second, a massive diversity of different measure is used that reflects evaluation aspects of the clustering result. Finally, authors have limited their survey to their recommend paradigm only, while paying other paradigms no observation. In a systematic analysis they used various quality measures and provide an outcome for a wide range of synthetic and real world data. They deliver the first comparison of divergent paradigm properties in a thorough evaluation. they show that density-based approaches does not scale very high dimensional data, while clustering align approaches are overdone by noisy data resulting in low clustering quality.

Analysis constitutes an important basis for subspace clustering research. [5] In this paper, they provide a new probabilistic approach to k nearest neighbor classification, naive hubness Bayesian KNN (NHBNN), which employs hubness for computing class likelihood estimates. Demonstration show that NHBNN collate favorably to various variants of the KNN classifier, inclusive probabilistic KNN (PNN) which is often used as a probabilistic framework for Nearest Neighbor classification, signifying that nearest hub based NN is a promising possible framework for development probabilistic Nearest Neighbor algorithms. The idea is to examine if it is feasible to incorporate hubness data for Bayesian class prediction. Hubness is expound as prior data about an event of the part in k-neighbor sets and used to set up a Bayesian KNN model. The presented approach is basically different from related methods, probabilistic framework for using hub, and introduces the naive hubness Bayesian kNN classification algorithm (NHBNN).

One of the problems of standard KNN classifier is output does not meaningful probabilities related to class prediction. Hubness is an event to inherent high-dimensional data which has recently taken into serious analysis and has never used in a Bayesian framework. They have shown in this paper that taking point in hubness may be beneficial to nearest-neighbor methods and that it should be thoroughly explore. The presented algorithm varies in its conceiving greatly from the probabilistic KNN (PNN) and related methods. [6] High-dimensional data are difficult to handle by ordinary machine-learning algorithms, which is particularly characterized the curse of dimensional. In this paper they go additionally by holding the soft approach, and introduce many fuzzy measures for k-nearest neighbor classification, all are based on the hubness, which expose fuzziness of elements materialize in k-neighborhoods of another points.

ISSN: 2278 – 1323

the goal is to enlarge the class-non specific crisp k Nearest Neighbor weighting scheme[5].

Fuzzy nearest-neighbor classification provide better assured calculate the proposed labels, which leads to likely easier interpretability of the product by experts working on the problem and this is the reason they decided to enlarge the previous crisp hubness-based approach into a fuzzy counterpart. many hybrid fuzzy functions were tested and analyzed.[7] Distance concentration is an occurrence the difference between the nearest and the farthest neighboring points disappear in the data dimensionality get larger. it has the advantage of the resulting testing procedure makes no assumptions on the data distribution. Moreover, It should be recollect a tighter distribution free bound cannot be procure without introducing hypothesis. The problem of identifying specific data distributions of a big enough interest that allow tighter bounds to be derived is non-trivial. The event of the distance absorption asymptotically in the limit of infinite dimensions. This enables evaluation of a given distance in a given data model, and allows us to recognize conditions on the data distribution that matter.

II. HUBNESS BASED CLUSTERING

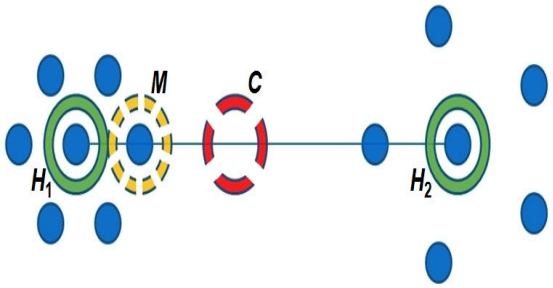

Fig. 1. Hub Based Clustering

From the Fig.1 hubness is observed as a kind of local centrality measure, it may be able to done hubness for clustering in many ways. In order to test this theory, 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 sub-clusters 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 Centroids-based technique. For this reason, the algorithm look like K means, 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 actually speed up the convergence of the algorithms, leading straight to the promising regions in the data space. Centroids depend on all current cluster elements, while hubs depend mostly on their neighboring elements and therefore carry localized centrality information. We will consider two types of hubness namely global hubness and local hubness. We define local hubness as a restriction of global hubness on any given cluster, considered in the context of the current algorithm iteration.

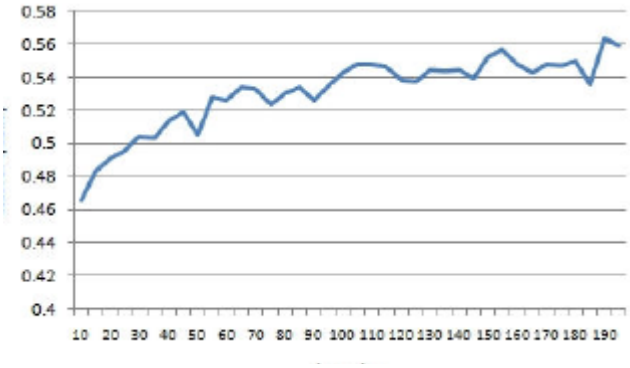

Fig. 2. Quality of Clustering for Various Durations

From the Fig.2 Quality of Clustering for various duration in Hubness proportional clustering. The silhouette index estimates the cluster quality. The hubness proportional clustering algorithm explain the search between the data space based on hubness as a kind of a local centrality evaluate, probabilistic iteration help in achieving better clustering

III. PROBLEM DESCRIPTION

The methods tailored for detecting approximately hyper spherical clusters, clustering methods lead to cluster configurations where hubs have higher b-values than in the case of K-means the proposed algorithms represent only one possible approach to using hubness for improving high dimensional data clustering. To properly handle clusters of arbitrary shapes shows excellent quality because it uses robust methods in order to measure distances between clusters. The real world data sets evaluated its accuracy in comparison to subspace selection proposed work. The proposed algorithms are exploiting hubness for data clustering. The proposed forward search algorithms in none of the correct sub spaces were found.

IV. SYSTEM PROCESS

From the Fig.2 User can take low dimensional data in previous stage clustering with data no losses should occur. Maximum no of data will not used in low dimensional data Clustering with wrong data enabling default data accuracy displayed.

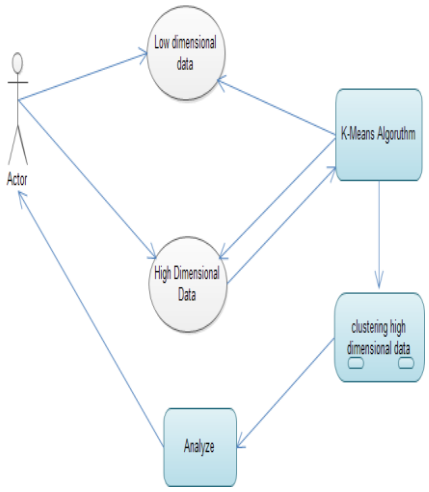

Fig. 3 System Architecture

In this stage user can use high dimensional data with maximum no of data .multiple rows and columns used for high dimensional data clustering into real data dimensional using k means algorithm K means algorithm using clustering with real data depends upon density based approach to find nearest neighbor node. Clustering of high dimensional data to transfer analyze process to find out how much amount of noises occur in data. User get this searching data entity value (real data value) how much data relation between the maximum distances of data noises occur, it should become data losses. Analyze process noises data neglecting into high dimensional data process .To finalize clustering of real data into without losses to get this user real data

A. Relation of Hubs Phenomenon

As the dimensionality of data increases, the distribution of k-occurrences becomes considerably skewed as a phenomenon, appears in high-dimensional data as an inherent property of high dimensionality, and is not an arte fact of finite samples or a peculiarity of some specific data sets.

B Scalability for Very Large Databases

The merged cluster may be calculated easily. In high-dimensional spaces, however, low-hubness elements are expected to occur by the very nature of these spaces and data distributions. This is due to the fact that some hubs are actually close to points in different clusters.

C Clustering Of Real World Data

Real-world data is usually much more complex and difficult to cluster; therefore such tests are of a higher practical significance. A single data set was clustered for many different K-s (number of clusters), to see if there is any difference when the number of clusters is varied. Results were compared for various predefined numbers of clusters in algorithm calls. K-means achieved the best overall cluster quality on this data set.

D Synthetic Data Evaluation

comparing the resulting clustering quality, we used mainly as an unsupervised measure of configuration validity most of the generated data sets we also report the normalized frequency with which the algorithms were able to find these perfect configuration

V. RESULT ANALYSIS

This section describes experiments on various high-dimensional synthetic and real-world data sets. The algorithm suggested in this paper is not very sensitive to changes of k, with no distinct monotonicity in score i.e. the clustering quality does not rise with rising k or vice versa. Results were compared for various predefined numbers of clusters in algorithm; Real-world data often contains noisy or erroneous values due to the nature of the data collecting procedure. It is natural to conclude that hub-based algorithms will be more robust with respect to noise. In the case of KM, all of the instances within the current cluster directly determine the location of the Centroids in the next iteration. K-means++ will be used as the main baseline for contrast, it is suitable for determining the practicality of using hubness to evaluate local centrality of points.

	GKH		GHPC		KM	
	Silhouette	Entropy	Silhouette	Entropy	Silhouette	Entropy
Avg. total	0.78	0.35	**0.83**	**0.28**	0.70	0.62
Avg. noise 10–50%	0.77	0.37	**0.82**	**0.29**	0.68	0.66
Avg. noise 30–50%	0.73	0.42	**0.80**	**0.32**	0.66	0.71

Table. 1 Cluster Quality at Various Noise Level

VI. CONCLUSION

Using hubs to inexact local data centers is not a practical option, but also normally leads to improvement over the Centroids-based approach. In our test GHPC (Global Hubness-Proportional Clustering) had a best performance in various test settings, on synthetic and real-world data, as well as in the existence of high levels of artificially initiate noise. The proposed algorithm represents using hubness for improving high dimensional data clustering. In future they do Kernel mappings and shared-neighbor clustering. They would like to explore methods for using hubs to automatically determine the number of clusters in the data

.

REFERENCE

[1] C. C. Aggarwal and P. S. Yu, "Finding generalized projected clusters in high dimensional spaces," in Proc. 26th ACM SIGMOD Int. Conf. on Management of Data, 2000, pp. 70–81

[2] K. Kailing, H.-P. Kriegel, P. Kr ¨oger, and S. Wanka, "Ranking subspaces for clustering high dimensional data,"

ISSN: 2278 – 1323

in Proc. 7th European Conf. on Principles and Practice of Knowledge Discovery in Databases (PKDD), 2003, pp. 241–252.

[3] K. Kailing, H.-P.Kriegel, and P.Kr¨oger "Density-connected subspace clustering for high-dimensional data," in Proc. 4th SIAM Int. Conf. on Data Mining (SDM), 2004, pp. 246–257.

[4] E. M¨uller, S. G¨unnemann, I. Assent, and T. Seidl, "Evaluating clustering in subspace projections of high dimensional data,"Proceedings of the VLDB Endowment, vol. 2, pp. 1270 1281, 2009

[5] C. C. Aggarwal, A. Hinneburg, and D. A. Keim, "On the surprising behavior of distance metrics in high dimensional spaces," in Proc. 8th Int. Conf. on Database Theory (ICDT), 2001,pp. 420–434.

[6] D. Franc¸ois, V. Wertz, and M. Verleysen, "The concentration of fractional distances," IEEE Transactions on Knowledge and Data Engineering, vol. 19, no. 7, pp. 873–886, 2007.

[7]A. Kab´an, "Non-parametric detection of meaningless distances in high dimensional data," Statistics and Computing, vol. 22, no. 2, pp. 375–385, 2012

[8] I. S. Dhillon, Y. Guan, and B. Kulis, "Kernel k-means: spectral clustering and normalized cuts," in Proc. 10th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, 2004, pp. 551–

[9] E. Bic¸ici and D. Yuret, "Locally scaled density based clustering," in Proc. 8th Int. Conf. on Adaptive and Natural Computing Algorithms (ICANNGA), Part I, 2007, pp. 739–748.

[10] "A probabilistic approach to nearest-neighbor classification: Naive hubness bayesian kNN," in Proc. 20th ACM Int. Conf. on Information and Knowledge Management (CIKM), 2011, pp. 2173–2176.

[11] E. Agirre, D. Mart´ınez, O. L. de Lacalle, and A. Soroa, "Two graph-based algorithms for state-of-the-art WSD," in Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP), 2006, pp. 585–593.

[12] K. Ning, H. Ng, S. Srihari, H. Leong, and A. Nesvizhskii, "Examination of the relationship between essential genes in PPI network and hub proteins in reverse nearest neighbor topology," BMC Bioinformatics, vol. 11, pp. 1–14, 2010.

[13] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in Proc. 18th Annual ACM-SIAM Symposium on Discrete Algorithms (SODA), 2007, pp. 1027–1035.

[14] C. Zhang, X. Zhang, M. Q. Zhang, and Y. Li, "Neighbor number, valley seeking and clustering," Pattern Recognition Letters, vol. 28, no. 2, pp. 173–180, 2007.

[15] N. Tomasev, R. Brehar, D. Mladeni´c, and S. Nedevschi, "Theinfluence of hubness on nearest-neighbor methods in object recognition," in Proc. 7th IEEE Int. Conf. on Intelligent Computer Communication and Processing (ICCP), 2011, pp. 367–374.

[16] K. Buza, A. Nanopoulos, and L. Schmidt-Thieme, "INSIGHT: Efficient and effective instance selection for time-series classification,"in Proc. 15th Pacific-Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), Part II, 2011, pp. 149–160.

[17] A.Nanopoulos, M. Radovanovi´c, and M.Ivanovi´c, "How does high dimensionality affect collaborative filtering?" in Proc. 3rd ACM Conf. on Recommender Systems (RecSys), 2009, pp. 293–296.

[18] M.Radovanovi´c, A. Nanopoulos, and M.Ivanovi´c, "On the existence of obstinate results in vector space models," in Proc.33rd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2010, pp. 186–193.

[19] J. J. Aucouturier and F. Pachet, "Improving timbre similarity:How high is the sky?" Journal of Negative Results in Speech and Audio Sciences, vol. 1, 2004.

[20] D. Schnitzer, A. Flexer, M. Schedl, and G. Widmer, "Local and global scaling reduce hubs in space," Journal of Machine Learning Research, vol. 13, pp. 2871–2902, 2012.

