IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 2, Ver. VIII (Mar-Apr. 2014), PP 22-30
www.iosrjournals.org

A Study on Clustering High Dimensional Data Using Hubness Phenomenon

V.Suganthi[1,] S.Tamilarasi[2]
II M.E(CSE) ,Shivani Engineering College Tiruchirappalli
Assistant Professor ,Shivani Engineering College Tiruchirappalli

Abstract: *Data mining is the non-trivial process of extracting information from the very large database. In recent years, data repository has a high dimensional data, which makes a complete search in most of the data mining problems leads computationally infeasible. To eradicate this problem clustering plays a vital role in handling low dimensional data and high dimensional data. Low dimensional data makes a task very simple and easy to cluster. High dimensional data is a crucial fact to cluster and it has to resolve using hubness phenomenon. Here hubness refers a data point which may frequently occurr among the groups. The existing system has to determine and manage the hyperspheric clusters. To overcome this problem, the proposed system is to use hub based clustering technique to improve the quality of cluster in terms of effectiveness and accuracy, and to avoid only detecting hyper- spherical cluster.*

Index terms: *Clustering, Nearest neighbour, Hubs, High dimensional data, biological data.*

I. Introduction

Generally, the process of analyzing data from different outlooks and succinct it into useful information is called as data mining. It is also known as data or knowledge discovery. It allows users to analyze the data from many different dimensions or angles, categorize it, and summarize the relationships identified. Officially, data mining is the method of finding the association or correlations or patterns among masses of fields in enormous interactive databases. Any details, numbers, or text that can be administered by a computer are termed as data, which includes operational data, non-operational data and meta data. The result of patterns, associations, or relationships among all this data can afford information. Then the facts can be converted into knowledge about historical patterns and future trends.

Data mining functionalities are characterization and discrimination, classification and prediction, cluster analysis, outlier analysis, trend and evolution analysis. Generally the low dimensional data is very simple and easy to cluster using clustering algorithms. Clustering is the crucial task to cluster the high dimensional data. The data items are collected authorizing to logical relationships or end user likings, which can be defined by the cluster. Theoretically, a cluster is collection of items which are *related* between them and are *unrelated* to the items belonging to other clusters. Unsupervised learning problem is significant role in clustering; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabelled data.

The determination of the intrinsic grouping in a set of unlabelled data can be considered the most important goal of clustering. Clustering algorithm has several requirements. Some of the main requirements such as, that should satisfy its scalability, dealing with different types of features, find out clusters with random shape, minimal necessities for domain knowledge to regulate input parameters, capacity to deal with noise and outliers, inattentiveness to order of input records, high dimensionality, interpretability and usability. There are a number of problems with clustering. They are dealing with large number of dimensions and large number of data items can be challenging because of time complexity, the result of the clustering algorithm that in many cases can be arbitrary itself. Clustering algorithm can be applied in many fields such as marketing, biology, earthquake studies, insurance, city - planning, www etc.

TABLE 1: SAMPLE DATASET OF PHARAMOCOLOGICAL CLASS LIST

Drugs	**Indication**	**Dosage**
Zaleplon	Insomnia	Adult : 10mg before bedtime Max 20mg daily Elderly : >65yr:5 mg before bedtime. Max : 10mg daily
Paroxetine	Depression	Adult : 20mg daily increase gradually max:50mg/day Elderly : initially 10mg daily, increase if needed max:40mg daily
Globac liqd zydus (alidac)	Prophylaxis	Adult : 15-30ml daily before meals Child : 2-5yrs : 5ml bid; 5- 12yrs:10ml bid
Granisetron	Nausea & vomiting associated w/cancer chemotherapy	Adult : 1-2mg w/in 1hr before the start of chemotherapy Child : 1mth- 12yr; 20 mcg/kg w/in 1hr before chemotherapy
Salmeterol	Chronic asthma; chronic obstructive pulmonary	Adult : As metered-dose aerosol: 2 inhalations of 25mcg bid up to 100mcg. Child : As metered-dose aerosol: >=4 yr; 50 mcg bid
Salbutamol	Acute bronchospasm	Adult : 2-4 mg 3- 4 times daily. Child : 1mth-2 y: 100 mcg/kg, 2- 6yr : 1- 2mg, >6yr : 2mg. Doses to be taken 3-4 times daily. Elderly : initially 2mg 3-4 times daily
Ketotifen	Allergic conditions; asthma prophylaxis	Adult : 1mg bid up to 2 mg bid, as needed. Alternatively, 0.5-1 mg at night for the 1st few days of treatment to minimise drowsiness.

Clustering the high-dimensional data can be defined by the cluster analysis of data with wherever from a little dozen to many thousands of dimensions. Such high-dimensional data spaces are often met in areas such as medicine, where DNA microarray technology can produce a large number of measurements at once, and the clustering of text documents, where, if a word-frequency vector is used. The problem needs to overcome for clustering in high dimensional data is the curse of dimensionality. Multiple dimensions are hard to think in, impossible to visualize, and, due to the exponential growth of the number of possible values with each dimension, complete enumeration of all subspaces becomes intractable with increasing dimensionality. This problem is known as the curse of dimensionality.

Difficulty in distinguishing between related and unrelated points is, however, not the only aspect of the dimensionality curse which also problems in k-nearest neighbor based implication[21]. The phenomenon of hubness has been referred and used it as actually highly unfavourable. Formation of hubs by, very frequent neighbor points which take over among all the occurrences in the k-neighbor sets of intrinsically high dimensional data. Most other points either never appear as neighbors or appear so very rarely. They are mentioned to as anti-hubs. It is usually of a geometric nature and does not reflect the semantics of the data. In other words hubness has an impact on the forming of the shared neighbor similarity scores, which 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.

When compared with low dimensional data, high dimensional data is very difficult to cluster. Recently cluster is useful for bio-informatics and many other research areas. Clustering is needed in many number of domains, like the biological and medical applicaton such as microarray data analysis, analysis of drug activity and analysis of nutrition data.

Table 1 illustrates the sample real world dataset as pharmacological class list for clustering. The datasets contains the generic name, oral administration as its indication and dosage for different systems. Section 2 specifies the existing method that can be used for clustering the high dimensional data. The proposed method uses hub based clustering algorithm for this real world data sets for cluster.

II. Literature Survey

The various clustering algorithm has been used for predicting the high dimensional data. The high dimensional data has regularly affected by the curse of dimensionality. This leads to the greater impacts on the density based and subspace clustering algorithm. The effects of high dimensional data can be tried to resolved using hubness phenomenon. The several different approaches used to evaluate the high dimensional data can be discussed below:

A. Subspace clustering algorithm

The author [7] has proposed to detect the cluster and compared with three different prototype models such as cell based, density based and clustering oriented approach.

The author[7] proposed cell based approaches pursuit for sets of fixed or variable grid cells. It employs several approaches and all are based on a cell approximation of the data space. First typical model in this approach for clustering was presented by CLIQUE. The CLIQUE algorithm can be divided into the following three steps: (1) Find even units and identify subspaces containing clusters. (2) Identify clusters in the selected subspace. (3) Generate minimal description for the clusters. Another algorithm be SCHISM, which improves the cluster definition by variable threshold fit to the subspace aspects. Efficient subspace clustering is mainly based upon monotonicity property in pruning subspace. This pruning technique can be used in CLIQUE algorithm to remove noise and to improve the efficiency of the cluster quality. Both CLIQUE and SCHISM approach, they sum the number of objects within the cell and to compute whether this cell is a part of subspace cluster or not. Moreover, the result cluster is highly dependent on their cell properties but the result of the computation be much more efficient. Typically this methods independent on the number of data objects, yet dependent on grid size.

The author[7] recommended density based approaches are defined with respect to Euclidean distance and minimum points. The “closeness” is defined in terms of a distance metric, such as Euclidean distance. First typical model in this approach for clustering was SUBCLU. SUBCLU is an extension of DBSCAN algorithm. SUBCLU uses monotonicity property, so that it reduces the search space by pruning technique like in CLIQUE.

For each computation process, SUBCLU uses the original data and require large database scans. Due to large database scan, the computation result be in efficient. Generally finding a meaningful cluster within the neighbourhood range is a major challenging task. This task be overcome by the method FIRES. It uses 1d histogram information and experimental to adapt a neighbourhood range to its subspace dimensionality.

Another approach in density based approach is INSCY. It is an extension of SUBCLU. It removes the superfluous low dimensional clusters. Therefore, INSCY attains an efficient computation. This results in an efficient computation. This method may filter out the outliers.

The author[7] offered clustering oriented approaches mainly concentrate on the clustering result. First approach in this model is PROCLUS[2]. PROCLUS extends the k-means algorithm. Datas are collected and stored in the database. These datas are separated into k clusters with normal dimension. Another approach in this model is STATPC. Typical function of STATPC is to determine the best non- unneeded clustering and it defines the properties of cluster. Its main objective is to improve the quality of clustering.

These clustering oriented approaches promptly control the clustering result. And it do not control the individual cluster. Both the cell based and density based approaches stipulate a cluster definition and they does not provide any improvement process to select cluster. But the clustering oriented approach provide optimization process to the overall clustering.

The overall quality of the cluster can be judge by the measure called accuracy. The author has specified this accuracy by

$$\text{Accuracy} = \frac{\text{Correctly Predicted Objects}}{\text{All Objects}}$$

Gene expressions are popular in medical sciences research and development. The author has adopted the real world data sets as gene expressions and provided a standard level set of results on a large variety of real world and synthetic data sets.

Table 2 illustrates the properties of each approaches and comparisons are made by these methods can be evaluated.

The author has been evaluated the results, which can be made by different measures such as RNIA, CE (clustering error), accuracy, entropy, coverage, number of clusters, runtime, average dimension and the results have represented by graphical representations.

The author has explored the important properties for each models and compared them in evaluations get highlighted.

TABLE 2: CHARACTERISTICS OF THREE PROTOTYPE MODELS

Prototype Model	**Approach**	**Properties**
Cell based	CLIQUE	Fixed threshold, fixed grid, pruning by monotonicity property
Cell based	SCHISM	Enhances CLIQUE by variable threshold
Density based	SUBCLU	Fixed density threshold, pruning by monotonicity
Density based	FIRES	Variable density threshold, using 1d histograms for approximate pruning
Density based	INSCY	Variable density threshold, reducing result size by redundancy elimination
Clustering oriented	PROCLUS	Fixed result size, iteratively improving result like k-means, partitioning
Clustering oriented	STATPC	Statistical tests, reducing result size by redundancy elimination

B. Density based clustering algorithm

The author[11] used SUBCLU algorithm for identifying the clusters in high dimensional data. The author has taken a gene expression as data set and compared the evaluation results for generating all the clusters.

One of the approach for subspace clustering is CLIQUE[1]. It is a simple grid-based method for finding density based clusters in subspaces. CLIQUE partitions each dimension into non overlapping intervals, thereby partitioning the entire embedding space of the data objects into cells. It uses a density threshold to identify dense cells and sparse ones. A cell is dense if the number of objects mapped to it exceeds the density threshold. All datas which are not dense are eliminated. CLIQUE uses pruning technique to eliminate the redundant data. It is used to genereate the minimal cluster report.

Another approach for subspace clustering is ENCLUS[2] also handled by the author[2]. It is based on computation of a discrete random variable. It uses monotonicity property and it is analogous to CLIQUE. It is based on bottom-up approach.

The major shortcomings of all these methods are, they depend on grid size. As an alternative of using grids, take on the concept of density connectivity. Generally clustering algorithm fade to produce a meaningful cluster. In order to create a meaningful datasets, the role of SUBCLU algorithm in subspaces. It has several qualities. SUBCLU algorithm has the ability to identify arbitrarily shaped cluster in subspace. It uses bottom-up strategy to remain efficient. It has well defined cluster concept. In contrast to CLIQUE, SUBCLU does not use any pruning technique.

The main strategy behind the density based clustering methods, which can discover clusters of non-spherical shape. The density based clustering methods uses DBSCAN [3] , DENCLUE[4] and OPTICS[5] for clustering.

Using a Minimum point, this specifies the density threshold of dense region in DBSCAN for finding arbitrary shaped cluster very effective. But it is not suitable for real world, high dimensional data. The edifice of the OPTICS is very similar to the DBSCAN. It does not explicitly produce a data set clustering. It does not require any specific density threshold. The author[15][6] offered DBSCAN and the author[14][6] recommended OPTICS. In DBSCAN and OPTICS, density is calculated by counting the number of objects in a neighbourhood and it can be highly sensitive. DENCLUS is based on a set of density distribution functions. It uses kernel density estimation and can effectively reduce the influence of noise. Moreover, DENCLUE is invariant against noise. All these density based methods may filter out the outliers.

The SUBCLU algorithm first develop all 1-dimensional clusters by using DBSCAN. Pruning technique is applied to the resulting cluster, to check whether this cluster is in part of it or not in the subspaces. This technique is used to reduce the clusters in the subspaces. Finally develop and list the k+1 dimensional clusters. The advantage of using SUBCLU algorithm is to minimize the cost of the runs of DBSCAN.

To evaluate the efficiency and scalability of SUBCLU the author compared it with CLIQUE. SUBCLU clustering algorithm can be applied to the following gene expression data that yields a meaningful clusters.

Table 3: Contents of two sample clusters in different subspace

Gene name	**Function**
Cluster 1 (subspace 90,110,130,190)	
RPC40	Subunit of RNA pol I and III, builds complex with cdc60
CDC60	tRNA synthesase, builds complex with RPC40
FRS1	tRNA synthesase
DOM34	Protein synthesase, mitotic cell cycle
CK41	Mitotic cell cycle control
CPA1	Control of translation
MIP6	RNA binding activity, mitotic cell cycle
Cluster 2(subspace 90,110,130,190)	
STE12	Transcription factor
CDC27	Regulation of cell cycle, possible STE12-site
EMP47	Golgi membrane protein, possible STE12-site
XBP1	Transcription factor

Using the datasets in table 3, to computed the scalabilty and accuracy of SUBCLU. In subspace of this data set, it may establish many attractive clusters by the time slots as 90,110,130,190.

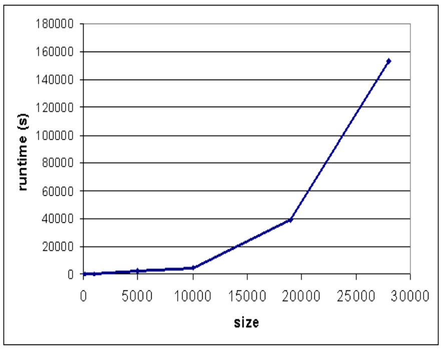


Figure 1: scalabilty of SUBCLU against the size of the data set

[1] Clustering In QUEst
[2] Entropy-based CLUStering
[3] Density Based Spatial Clustering of Applications with Noise
[4] DENsity based CLUstEring
[5] Ordering Points To Identify the Clustering Structure

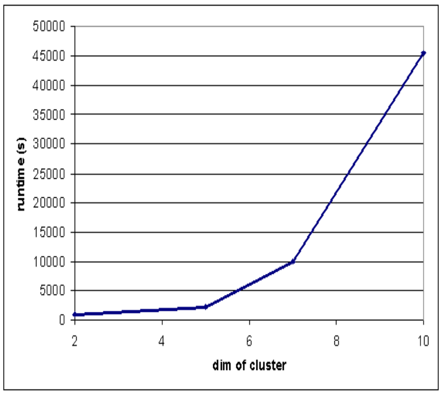


Figure 2: scalabilty of SUBCLU against the maximum dimensionality of the subspace cluster

Figure 1&2 performs multiple range queries in arbitrary subspaces with respect to the size of the data set and dimensionality of the cluster with scalability.

Therefore, the author presented the experimental and theoretical reults to compute the accuracy and scalability of SUBCLU.

C. Hubness based clustering algorithm

The author [17] has been proposed Hubness Information k- Nearest Neighbor (HIKNN) for managing high dimensional data. HIKNN algorithm was compared with other previous hubness based algorithm.

Hubs, is a data point that frequently occurred in k-nearest neighbor list and rarely occurring points or may outliers are called as antihubs. The search for nearest neighbour is a very critical aspect in clustering algorithm. The k-nearest neighbor [12] [21] [2] algorithm is the basic method for simple to find the nearest neighbor. It is broadly used as a classification method and very straightforward. The phenomenon of hubness is normally associated with concentration of distances.

Hubness aware approaches have three algorithms such as hw-kNN, h-FNN, NHBNN. Hubs can be categorized into two types. First one is good hubs and another one is bad hubs. This classification can be based on the number of label matches and mismatches in the k-occurrences. First approach is hw- kNN. This method reduces the impact of bad hubs and it is very simple to implement. Bad hubness can be identified by its weight. If a point shows a bad hubness, give its vote as lesser weight.

Second approach is h-FNN. This algorithm combines weight with fuzzy votes. It uses a threshold parameter. To determined the antihubs by using the threshold parameter. One major drawback in this algorithm as it has not explain a clear way of managing with antihubs.

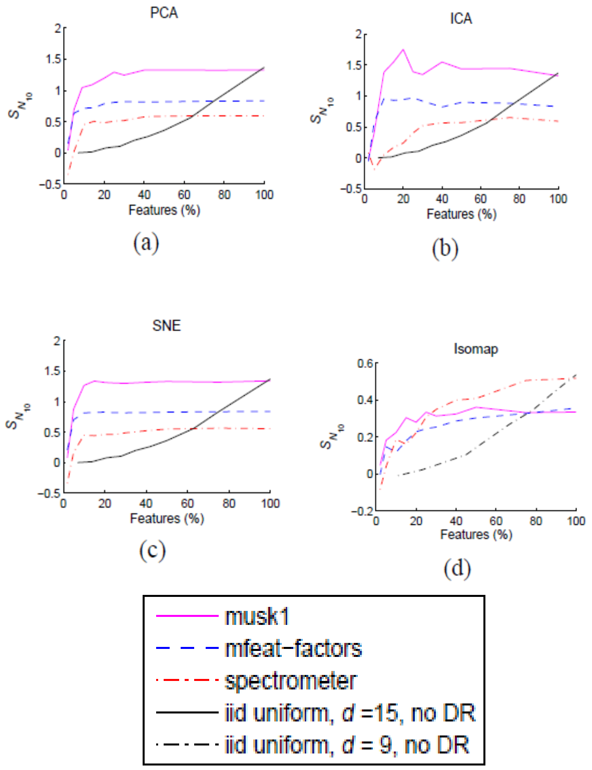
Third approach is NHBNN. This algorithm uses the Naïve Bayes rule to allowed for further optimization. It also have not give a detailed description of managing with antihubs. Both h- Fnn and NHBNN does not handle with antihubs. In High dimensional data, almost most of the points may belongs to either hubs or to antihubs but very few points may neither belongs to hubs nor to antihubs. These points have not take consideration into the previous algorithm. The following information based voting procedure taken these points into consideration.

HIKNN handles antihubs through information based structure. The overall occurrence of informativeness is taken into consideration. It had well generalized and may be over fitting on the dataset. It was parameter free. It has improved the overall classification accuracy.

D. Popular Nearest Neighbors in High Dimensional Data

The author [9] [16] [20] [22] [5] [18] has been performed a theoretical and experimental analysis of hubness and its triggres, techniques for clustering, classification and information retrieval. k-occurrence value can be computed by the position of a points in data space. If the dimensionality increases then the strong correlation appear, which indicates that the points closer to the mean and that points have a tendency to become a hubs. The mechanism of hubs can be applied to both unimodal and multimodal data distribution. The phenomenon of hubness has been usually related to the focus of distances. The ratio between the standard deviation and the mean can be represented by distance concenteration.

Hubness in real data necessarily takes two factor into an account. The first factor is dependent attributes and the second factor is many groups, that is real datasets are typically clustered. The author has been observed the following methods for reduce the dimensionality: (a) principal component analysis (PCA), (b) independent component
analysis (ICA), (c) stochastic neighbour embedding(SNE), (d)
isomap and (e) diffusion map.

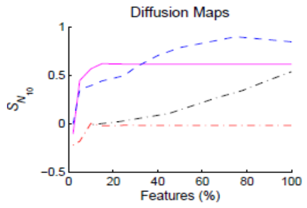


(a)

(b)

(c)

(d)



(e)

Figure 3: (a) principal component analysis-PCA (b) independent component analysis-ICA (c) stochastic neighbour embedding-SNE (d) isomap (e) diffusion map.

Figure 3(a-c) reduction of dimensionality may incur loss of valuable information. Figure 3(d,e) replace the original distance by the distances derived from a neighbourhood graph.

The author has been observed the interaction of k- occurrences with information provided by the labels based on good and bad k-occurrences. Good occurrences can be defined as the number of points from the dataset where the labels do match. Bad occurrences can be defined as the number of points from the dataset where the labels do not match. Bad hubs which affects the well-known classification. So that the author has been focused on classification algorithms such as k-nearest neighbour classifier, support vector machines and AdaBoost.

The first algorithm is k-nearest classifier[13]. This algorithm uses Euclidean distance. It is used to detect the nearest neighbors. Bad hubs are not suitable to carry more information than other points for k-NN classification because non-borderline regions of space can be utilized by each class.

The Second algorithm is support vector machines. It can also specified by the border between classes. In contrast to k-NN classification, bad hubs is much more important for support vector machines. This algorithm can be defined by smooth monotone function of Euclidean distance between points in the dataset and which contains data dependent constant. Finally, the third algorithm is boosting algorithms. This algorithm takes more concentration to the points in the trained data sets. It assigns the weight to the points and updating the weights to the individual points. This algorithm improves accuracy and make restricted to its outliers. It act as a good performance of the weighting scheme to both hubs and antihubs and distance based outliers.

Clustering and outlier detection are the tasks that takes a vital role in hubness phenomenon. The goals of distance based clustering algorithm are to minimize the intra cluster distance and maximize the inter cluster distance. Outliers cannot be cluster because all the points other than cluster group in the dataset have high intra cluster distance. The points that can have low k-occurrences which leads to increase the intra cluster distance. On the other hand, the points that can have high k-occurrences which leads to reduce the inter cluster distance. Hubs do not cluster well because it have low inter cluster distance.

The author has been explored the feature of the curse of dimensionality and that can be cleared using the hubness phenomenon. The author has demonstrated it through theoretical and experimental analysis including synthetic and real data sets.

III. Conclusions And Future Work

From the literature survey, we discussed about the existing clustering method for cluster the data. In recent trends, biological data as real world dataset for clustering becomes difficult because it is a very high dimensional data. To eradicate this problem, the proposed system use a hub based clustering algorithm to automatically determine the number of cluster from the biological data. The proposed method is design to find the spherical shaped clusters from the sample dataset as mentioned as earlier in the Table 1. By using a hub based clustering technique to improve the quality of cluster in terms of effectiveness and accuracy, and to avoid only detecting hyper-spherical cluster.

References

[1] C. Aggarwal, J. Wolf, P. Yu, C. Procopiuc, and J. Park. Fast algorithms for projected clustering. In SIGMOD, pages 61-72, 1999.

[2] C. Ding and X. He, “K-nearest-neighbor consistency in data clustering:Incorporating local information into global optimization,”in *Proc. ACM Symposium on Applied Computing(SAC)*, 2004, pp. 584–589.

[3] C.-H. Cheng, A.W.-C. Fu, and Y. Zhang. ”Entropy-Based Subspace Clustering for Mining Numerical Data”. In *Proc. ACM SIGKDD Int. Conf.*

[4] D. Francois, 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.

[5] 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, 201

[6] E. Bicici 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.

[7] E. Muller, S. Gunnemann, 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.

[8] G. LakshmiPriya, Shanmugasundaram Hariharan, "An Efficient Approach for Generating Frequent Patterns Without Candidate Generation", ICACCI'12, August 03-05 2012,CHENNAI, India. Copyright 2012 ACM 978-1-4503-1196-0/12/08$10.00

[9] Hubs in space: Popular nearest neighbors in highdimensional data,*Journal of Machine Learning Research*, vol 11, pp 2487–2531, 2010. [10] J. Hartigan. Clustering Algorithms. John Wiley & Sons,1975.

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

[12] K. Q. Weinberger, J. Blitzer, and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," in *In NIPS*. MIT Press, 2006.

[13] L. Wang, L. Khan, and B. Thuraisingham, "An effective evidence theory based k-nearest neighbor (knn) classification," in *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology -Volume 01*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 797–801.[Online]. Available: http://portal.acm.org/citation.cfm?id=1486927.1487026

[14] M Ankerst, Markus M Breunig, Hans-Peter Kriegel, and Jorg Sander. "OPTICS: Ordering Points to Identify the Clustering Structure". In *Proc. ACM SIGMOD Int. Conf. on Management of Data, Philadelphia, PA*, pages 49–60, 1999.

[15] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases. In KDD, pages 226-231, 1996.

[16] M. Radovanovic, A. Nanopoulos, and M. Ivanovic, "Nearest neighbors in highdimensional data: The emergence and influence of hubs," in *Proc. 26th Int. Conf. on Machine Learning (ICML)*, 2009, pp. 865–872.

[17] 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.

[18] N. Tomasev, M. Radovanovi´c, D. Mladenic, and M. Ivanovic, "Hubness-based fuzzy measures for high dimensional k-nearest neighbor classification," in *Machine Learning and Data Mining in Pattern Recognition, MLDM conference*, 2011.

[19] N. Tomasev, M. Radovanovic, D. Mladenic, and M. Ivanovic,"The role of hubness in clustering high-dimensional data," in *Proc. 15th Pacific- Asia Conf. on Knowledge Discovery and Data Mining (PAKDD), PartI*, 2011, pp. 183–195.

[20] N. Tomasev, R. Brehar, D. Mladenic, and S. Nedevschi, "The influence of hubness on nearest-neighbor methods in object recognition," in *IEEE Conference on Intelligent Computer Communication and Processing*, 2011.

[21] P. B. Callahan and S. Rao Kosaraju. A decomposition of multidimensional point sets with applications to *k*-nearest-neighbors and *n*-body potential fields. *Journal of the ACM*, 42(1):67–90, 1995.

[22] P. Indyk and R. Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the Thirtieth Annual ACM Symposium on Theory of Computing*, 1998.