M-DENCLUE DETECTION IN HIGH DIMENSIONAL NON-LINEAR DATA IN CLUSTERING TECHNIQUES

R.NANDHAKUMAR[1] and Dr.ANTONY SELVADOSS THANAMANI[2]

[1.]Assistant Professor, Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi-642001, India

[2.]Associate Professor & Head, Department of Computer Science,
Nallamuthu Gounder Mahalingam College, Pollachi-642001, India

ABSTRACT

Clustering is a technique in data mining which deals with huge amount of data. Clustering is intended to help a user in discovering and understanding the natural structure in a data set and abstract the meaning of large dataset. It is the task of partitioning objects of a data set into distinct groups such that two objects from one cluster are similar to each other, whereas two objects from distinct clusters are dissimilar. Clustering is unsupervised learning in which we are not provided with classes, where we can place the data objects.

With the advent growth of high dimensional data such as microarray gene expression data, and grouping high dimensional data into clusters will encounter the similarity between the objects in the full dimensional space is often invalid because it contains different types of data. The process of grouping into high dimensional data into clusters is not accurate and perhaps not up to the level of expectation when the dimension of the dataset is high.

Keywords

High dimensional non-linear data, Clustering, DNA Micro array, Noise, Density based, Grid based.

1. INTRODUCTION

Clustering is unsupervised learning in which we are not provided with classes, where we can place the data objects. Clustering is beneficial over classification because cost for labelling is reduced. Clustering has applications in molecular biology, astronomy, geography, customer relation management, text mining, web mining, etc. Clustering can be used to predict customer buying patterns based on their profiles to which cluster they belong.

A cluster defined as a dense component, where it can grow in any direction that density leads. There are two approaches. The first approach is a density to a training data point like DBSCAN and OPTICS. The second approach is a density to a data point in the attribute space uses a density function like DENCLUE.

Curse of Dimensionality - Dimensionality curse is one of the major problems faced by high dimensional data. In high dimensional space the points are more scattered or sparse and all points are almost equidistant from each other. Clustering approaches become ineffective to analyse the data due to this.

Noise- The noise present in real applications often hides the clusters to be selected from clustering algorithm and the problem is worsened in high dimensional data, where the number of errors increases linearly with dimensionality.

DENCLUE algorithm and OPTICS algorithm comes under the density based clustering technique, where as CLIQUE algorithm comes under the Grid based clustering technique. Clustering in High Dimensional Non-Linear data spaces is a recurrent problem in many domains. It affects time complexity, space complexity, Data Size Adaptability and Precision Value of clustering methods.

2. RELATED WORKS

Clustering high dimensional data is definitely a problem for clustering techniques. This issue offers been studied thoroughly and there are numerous solutions, every befitting various kinds of high dimensional data and data exploration methods. There are numerous potential applications like bioinformatics, text message gold mining with large dimensional info where subspace clustering, forecasted clustering methods may help to discover patterns skipped by current clustering strategy. To be able to gain conceptual clearness of the domain name under research various content articles, books, websites plus some of other personal reviews had been examined. The review provides been carried out by directing on the main element subject of clustering substantial dimensional info.

Maithri. C and Chandramouli. H presented a short a comparison of the prevailing methods that were primarily concentrating in clustering on high dimensional data. The primary objective of the study newspaper is to show the potency of excessive dimensional info evaluation and various algorithm inside the prediction procedure for Data exploration. The overall performance issues of the info clustering in great dimensional data, additionally it is essential to study problems like dimensionality decrease, redundancy elimination, subspace clustering, co-clustering and info Labeling intended for clusters are to analyzed and increased.

SunitaJahirabadkar, ParagKulkarni presented an assessment of various denseness centered subspace clustering codes as well as a comparative graph concentrating on their particular distinguishing features such as for example overlapping / non overlapping, axis similar / randomly oriented and so forth. Charles Bouveyron, Stephane Girard, et aje., presents a clustering strategy which estimates the precise subspace and the inbuilt dimension of every class. Their particular strategy gets used to the Gaussian combination unit framework to high-dimensional data and estimations the guidelines which greatest fit the info. We get yourself a robust clustering technique known as Large Dimensional Data Clustering. They used Great Dimensional Data Clustering to find items in organic pictures within a probabilistic platform. Experiments on a lately proposed data source demonstrate the potency of our clustering way for category localization.

E. Kailing, L. P. Kriegel et approach., launched a SUBCLU (density- linked Subspace Clustering), a competent method of the subspace clustering issue. Applying the idea of thickness connection fundamental the formula DBSCAN, SUBCLU is founded on an official clustering idea. As opposed to existing grid-based techniques, SUBCLU will be able to detect randomly formed and positioned groupings in subspaces.

.

3. PROPOSED MODEL

Clustering or data grouping is the key technique of the data mining. It is an unsupervised learning task where one seeks to identify a finite set of categories termed clusters to describe the data. The grouping of data into clusters is based on the principle of maximizing the intra class similarity and minimizing the inter class similarity.

3.1. CLUSTERING TECHNIQUE USED

The grouping of data into clusters is based on the principle of maximizing the intra class similarity and minimizing the inter class similarity. A good clustering method will produce high quality clusters with high intra-class similarity - Similar to one another within the same cluster low inter-class similarity. The quality of a clustering method is also measured by its ability to discover some or all of the hidden patterns.

The objective of the clustering technique is to determine the intrinsic grouping in a set of unlabeled data. The similarity between data objects can be measured with the imposed distance values. Specifying the distance measures for the high dimensional data is becoming very trivial because it holds different data values in their corresponding attributes.

Data mining allows extracting data from the huge information and changing that data into a reasonable and important structure for additionally utilize. Data mining is a fundamental task during the time spent learning revelation from large information. Data mining is an advance mechanism that is very useful to mine the comprehensible knowledge, previously unknown, information from large amount of data stored in various formats, with the objectives of improving the decision of companies, organizations where the data would be collected.

3.2. HIGH-DIMENSIONAL DATA IN KNOWLEDGE DISCOVERY DATABASE

Clustering high dimensional data in a gene expression microarray data set, there could be tens or hundreds of dimensions, each of which corresponds to an experimental condition. Curse Dimensionality is a loose way of speaking about data separation in high dimensional space. The complexity of many existing data mining algorithms is exponential with respect to the number of dimensions. Each group is a dataset such that the similarity among the data inside the group is maximized and the similarity in outside group is minimized.

3.3 DENSITY BASED CLUSTERING ALGORITHM

Density based clustering algorithms are very popular in the applications of data mining. These approaches use a local cluster criterion and define clusters as the regions in the data space of higher density compared to the regions of noise points or border points. Density based clustering algorithms using the notion of DBSCAN, can find clusters of arbitrary size and shape. Density-based clustering can be seen as a non-parametric approach, where clusters are modeled as areas of high density. CLIQUE is the first grid based subspace clustering approach designed for high dimensional data. It detects subspaces of the highest dimensionalities.

3.4 EXECUTION PHASE

Stage 1: Applying DENCLUE, Optical technologies and Fanfare Algorithm upon High-dimensional Non- Linear Dataset.

Stage 2: Predicated on stage 1 effect, two methods (DENCLUE and OPTICS) had been chosen like a greatest formula. To resolve some disadvantages of the two algorithms several mathematical strategies such as for example curse of dimensionality, data redirecting, correlation, regular distribution and Darboux variate had been added with these algorithms. Finally, these modified algorithms had been applied about High dimensional nonlinear info set and result.

4. RESULT AND DISCUSSION

M-DENCLUE works on two stages as pre-processing stage and clustering stage. In pre-processing step, it creates a grid for the data by dividing the minimal bounding hyper-rectangle into d-dimensional hyper-rectangles with edge length 2σ. In the clustering stage, M-DENCLUE associates an "influence function" with each data point and the overall density of the dataset is modelled as the sum of influence functions associated with each point. The resulting general density function will have local peaks, i.e., local density maxima, and these local peaks can be used to define clusters.

M-DENCLUE uses influence functions. Influence of each data point can be modelled as mathematical function. The resulting function is called Influence Function. Influence function illustrates the impact of data point within its neighbourhood.

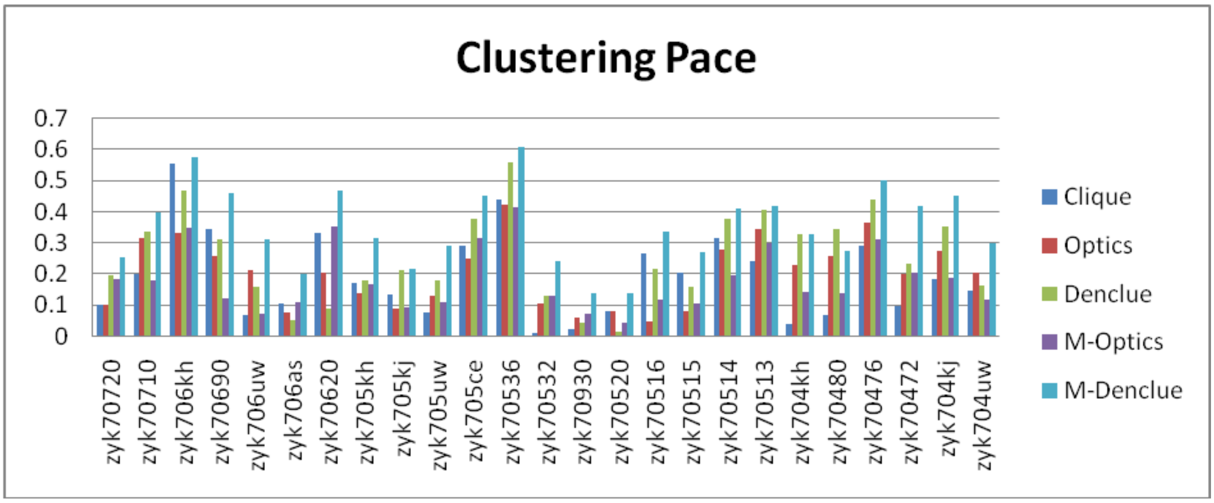
The M-OPTICS algorithm creates an ordering of the objects in a database, M-OPTICS additionally storing the core-distance and a suitable reachability distance for each object. An algorithm was proposed to extract clusters based on the ordering information produced by M-OPTICS and once the order and the reachability distances are computed, we can extract the clusters for any clustering distance.

The work is illustrated through graphs with the help of - DNA microarray Data.

Clustering Pace on DNA Microarray data set CLIQUE, OPTICS, DENCLUE, M-OPTICS and M-DENCLUE Algorithms.

	CLIQUE	OPTICS	DENCLUE	M-OPTICS	M-DENCLUE
zyk70720	0.101391	0.10273	0.199473	0.183355	0.25517338
zyk70710	0.200161	0.317208	0.337633	0.182342	0.39739786
zyk706kh	0.555159	0.333244	0.469775	0.351924	0.57830717
zyk70690	0.347334	0.260886	0.31413	0.124695	0.4606677
zyk706uw	0.071067	0.214153	0.159508	0.072177	0.31296296
zyk706as	0.108637	0.077021	0.05461	0.109321	0.20170116
zyk70620	0.335354	0.206745	0.090335	0.352806	0.47094082
zyk705kh	0.174004	0.139521	0.181694	0.170664	0.31579885
zyk705kj	0.137659	0.091897	0.213333	0.09523	0.22
zyk705uw	0.079098	0.131335	0.179458	0.112586	0.29094511
zyk705ce	0.29313	0.252728	0.378045	0.317399	0.45156635
zyk70536	0.441155	0.425971	0.558058	0.414736	0.6110013
zyk70532	0.011645	0.106578	0.132544	0.129983	0.24398184
zyk70930	0.025174	0.060603	0.044888	0.075757	0.13996714
zyk70520	0.081876	0.080965	0.015625	0.046639	0.140625
zyk70516	0.269663	0.047389	0.219693	0.121003	0.33900181
zyk70515	0.205245	0.0817	0.159397	0.105742	0.26987437
zyk70514	0.315193	0.281765	0.378375	0.19719	0.410431
zyk70513	0.242647	0.345973	0.40807	0.300967	0.41970008
zyk704kh	0.042641	0.230318	0.330698	0.144264	0.32906407
zyk70480	0.068823	0.258209	0.346408	0.140234	0.27737333
zyk70476	0.291449	0.367389	0.43933	0.311252	0.50195872
zyk70472	0.099653	0.202734	0.234231	0.207511	0.42033378
zyk704kj	0.185863	0.275163	0.355582	0.191243	0.45221191
zyk704uw	0.146397	0.206896	0.166375	0.117798	0.30009238

Experimental Clustering Pace

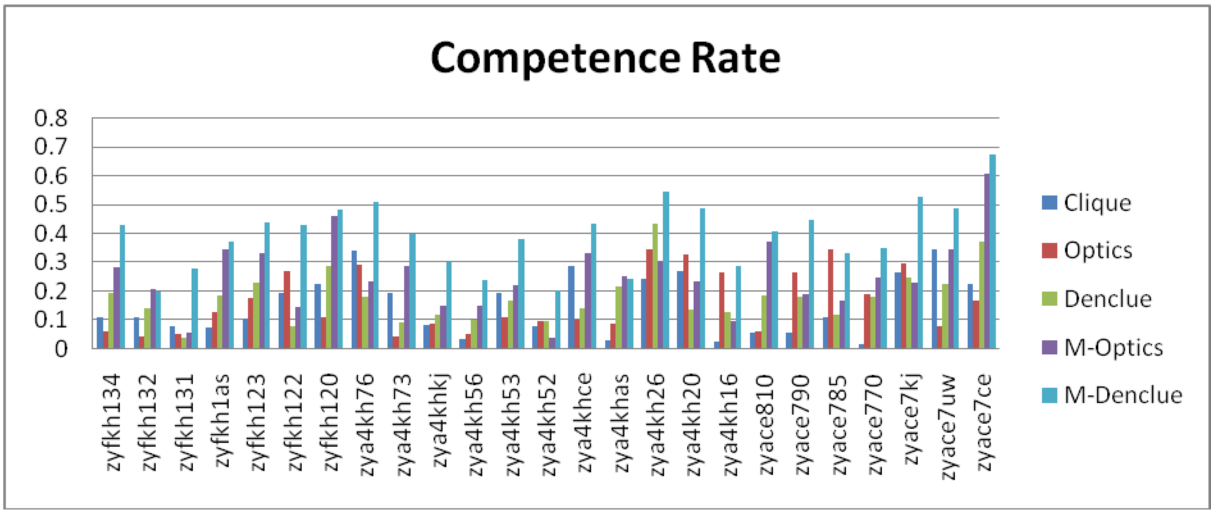


Experimental Inference-Competence Rate

	CLIQUE	OPTICS	DENCLUE	M-OPTICS	M-DENCLUE
zyfkh134	0.10726	0.059277	0.190335	0.281155	0.42707339
zyfkh132	0.108917	0.041342	0.140705	0.204966	0.19507669
zyfkh131	0.077543	0.047218	0.036129	0.052473	0.27699428
zyfkh1as	0.070764	0.123056	0.185223	0.346021	0.37236181
zyfkh123	0.104919	0.175736	0.229547	0.331158	0.43615085
zyfkh122	0.192935	0.266635	0.076078	0.143506	0.43048422
zyfkh120	0.222095	0.108282	0.285689	0.461416	0.48158565
zya4kh76	0.337714	0.292634	0.180675	0.230791	0.50999927
zya4kh73	0.191967	0.0418	0.091544	0.284445	0.39554934
zya4khkj	0.078303	0.08539	0.114213	0.146326	0.29810327
zya4kh56	0.030084	0.047161	0.100025	0.14649	0.23644031
zya4kh53	0.193828	0.108925	0.164135	0.217014	0.37879008
zya4kh52	0.077756	0.093267	0.091911	0.034201	0.20067229
zya4khce	0.286268	0.09681	0.140163	0.332161	0.43385884
zya4khas	0.028523	0.083555	0.216592	0.252168	0.24257063
zya4kh26	0.241483	0.342935	0.43169	0.302444	0.54731229
zya4kh20	0.269746	0.325518	0.132449	0.230701	0.48624276
zya4kh16	0.022534	0.261976	0.127218	0.092992	0.28767229

zyace810	0.052984	0.058634	0.182239	0.369891	0.40614054
zyace790	0.053503	0.265784	0.179295	0.189185	0.44763072
zyace785	0.10673	0.344255	0.11689	0.165832	0.32922957
zyace770	0.013127	0.186011	0.180138	0.244415	0.34674772
zyace7kj	0.265722	0.295721	0.246047	0.227704	0.52521232
zyace7uw	0.343682	0.07658	0.224918	0.343935	0.48799682
zyace7ce	0.223578	0.164429	0.370004	0.608936	0.67430078

Experimental Competence Rate



5. CONCLUSION AND FUTURE WORK

DENCLU and OPTICS are density based clustering technique, where as CLIQUE comes under the grid-based clustering technique. Comparing all those finally conclude that DENCLUE is the best one. Since day to day life changes with digital world, to group particular data. DENCLUE helps in reducing noise.

From experimental results it has been found that large and dense data needs higher computational power. In future the problems encountered in the existing methods can be overcome by developing a hybrid based density algorithm.

REFERENCE

[1] Dr. Anjali B. Raut, "A Hybrid Framework using Fuzzy if-then rules for DBSCAN Algorithm", Advances in Wireless and Mobile Communications, ISSN 0973-6972 Volume 10, Number 5 (2017), pp. 933-942.

[2] Feng Cao, Weining Qian et al., "Density-Based Clustering over an Evolving Data Stream with Noise", Department of Computer Science and Engineering, Fudan University.

[3] Gaff, B. M., Sussman, H. E., & Geetter, J., “Privacy and big data”, Computer, 47(6), 7–9. doi:10.1109/mc.2014.161, 2014.

[4] Gosain Anjana & Chugh Nikita, “Privacy Preservation in Big Data”, International Journal of Computer Application, Vol. 100 No.17 August 2014.

[5] Hajar Rehioui, Abdellah Idrissi et al., "DENCLUE-IM: A New Approach for Big Data Clustering", The 7th International Conference on Ambient Systems, Networks and Technologies (ANT 2016), ScienceDirect, Procedia Computer Science 83 (2016) 560 – 567.

[6] Harsh Shah, Karan Napanda et al., "Density Based Clustering Algorithms", International Journal of Computer Sciences and Engineering, Volume-3, Issue-11, E-ISSN: 2347-2693.

[7] Harsh Shah, Karan Napanda, "Density Based Clustering Algorithms", International Journal of Computer Sciences and Engineering, Review Paper, Volume-3, Issue-11, E-ISSN: 2347-2693, 2015.

[8] Hadi Saboohi et al., "On Density-Based Data Streams Clustering Algorithms: A Survey", Journal of Computer Science and Technology 29(1): 116{141 Jan. 2014.

AUTHORS PROFILE

R.Nandhakumar ,Assistant Professor in Computer Science, Nallamuthu Gounder Mahalingam College, undergoing Ph.D in Data Mining. Published various papers under reputed journals. He is the coordinator in various activates in college like NAAC, ISO, etc..

Dr. Antony Selvadoss Thanamani, Associate Professor and Head, Research Department of Computer Science, He is Research supervisor for Ph.D. degree in Computer Science in the Bharathiar University, Dravidian University, etc.. He established Common Research centre, E-content studio, ISBN Nodal Agency at NGM College, Pollachi. He is Advisory

Committee Member in National Conference on Advanced Computing. Editor in International Journal of Advanced Scientific Research, India . Editorial Board Member in International Journal of Advanced Research in Computer and Communication Engineering, India.