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

International Journal of Innovative Research in Science, Engineering and Technology

Volume 3, Special Issue 3, March 2014

2014 IEEE International Conference on Innovations in Engineering and Technology (ICIET'14)

On 21st&22ndMarch, Organized by

K.L.N. College of Engineering, Madurai, Tamil Nadu, India

Kullback-Leibler Divergence Measurement for Clustering Based On Probability Distribution Similarity

Priyadharshini.J, Akila Devi.S, Askerunisa.A

P.G Scholar, Vickram college of Engineering, Enathi, India.

Assistant Professor, Vickram college of Engineering, Enathi, India.

Head of the Dept., Vickram college of Engineering, Enathi, India.

Abstract-Clustering on Distribution measurement is an essential task in mining methodology. The previous methods extend traditional partitioning based clustering methods like k-means and density based clustering methods like DBSCAN rely on geometric measurements between objects. The probability distributions have not been considered in measuring distance similarity between objects. In this paper, objects are systematically modeled in discrete domains and the Kullback-Leibler Divergence is used to measure similarity between the probabilities of discrete values and integrate it into partitioning and density based clustering methods to cluster objects. Finally the resultant execution time and Noise Point Detection is calculated and it is compared for Partitioning Based Clustering Algorithm and Density Based Clustering Algorithm. The Partitioning and Density Based clustering using KL divergence have reduced the execution time to 68 sec and 22 Noise Points are detected. The efficiency of Distribution based measurement clustering is better than the Distance based measurement clustering.

Index Terms-Partitioning based clustering methods, Density based clustering method, Distribution based clustering, Kullback-Leibler divergence

I. INTRODUTION

Clustering is a grouping of similar objects based on their distance similarity measurement which is an "unsupervised classification". Finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups is called clustering. Basically clustering is categorized as inter cluster and intra cluster. The rule behind the cluster is "similarity between intra clustering is higher than the similarity between the inter clustering". In this paper, Distance Based Clustering Algorithms like Partitioning Based and Density Based are used and various Distribution based measurements are applied. In the Partitioning Based Clustering approach, the k-means method uses the distance to measure the similarity between two objects. In the Density Based Clustering approach, the objects in geometrically dense regions are grouped together as clusters and clusters are separated by sparse regions. The above Distance based approaches cannot distinguish the two sets of objects having different distributions. Hence Kullback-Leibler divergence measurement is used. In the Kullback-Leibler divergence, the objects Probability Mass Functions (PMF) over the entire data space are different and the difference between the PMF's can be captured using KL divergence.

II. LIMITATIONS OF EXISTING SYSTEM

In the Partitioning clustering approach, [2], [4], [6], [7], [9] which is an extension of the k-means method, the expected distance to measure the similarity between two objects, an object P and a cluster center C (which is a certain point) is calculated using equation 1

$$ED(P,C) = dist(P.C,C) + Var(P) \quad \ldots (1)$$

Where P-Object, Var(P)-variance of P and The distance measure dist is the square of Euclidean distance.

Here, only the centers of objects are taken into account in certain versions of the k-means method. As every object has the same center, the expected distance based approaches cannot distinguish the two sets of objects having different distributions.

In the Density based clustering approach, [1], [3], [10], [18], [19] the basic idea behind the algorithms does not change objects in geometrically dense regions are grouped together as clusters and clusters are separated by sparse regions. Generally the objects are heavily overlapped. There is no clear sparse region to separate objects into clusters. Therefore, the density based approach cannot separate the objects.

III. RELATED WORK

The Related work for the proposed paper comprises of the works that uses various clustering algorithms to implement the clustering of objects using Distribution measurement.

Alie.J et.al, [1] have used DBSCAN technique for clustering the spatial databases to identify arbitrary shape objects and to remove the noise during the clustering process.

Liping Jing.,et.al, [7] have used k-means Algorithm for clustering high dimensional objects in subspace. In high dimensional data, clusters of objects often exist in subspaces rather than in the entire space. The k-means clustering process uses the weight values calculated to identify the subsets of important dimensions that categorize clusters.

Grigorious,F,.Tzortzis, et.al, [8] have used Kernel K means Algorithm which was an extension of K means that depends on cluster initialization.

IV. PROPOSED SYSTEM

The difference between the distributions cannot be captured by geometric distances. In this paper, the Kullback-Leibler divergence is used to analyze the Probability Mass Functions (PMF) over the entire data space. The difference in the PMF value is captured by KL divergence. In general, KL divergence between two probability distributions is defined as follows. In the discrete case, let f and g are two probability mass functions in a discrete domain D with a finite number of values. The Kullback-Leibler Divergence between f and g is calculated using equation 2.

$$D(f||g) = \sum_{x \epsilon D} f(x) \log \frac{f(x)}{g(x)} \quad \ldots\ldots (2)$$

Where f(x) and g(x) – Probability Mass Function.

The probability of discrete values is integrated with Partitioning and Density Based Clustering Methods to cluster objects.

A. System Model

Discrete values in the dataset are preprocessed and the values are supplied to the Partitioning and Density Based Algorithm and their corresponding results are evaluated. The Probability Mass Function (PMF) value is calculated to calculating KL divergence measurement. Finally the values are again fed back to the Distance based algorithm and the results are compared.

1) Data Preprocessing

Real world databases are highly susceptible to noisy, missing, and inconsistent data due to their typically huge size and their origin from multiple, heterogeneous sources. Low quality data will lead to low quality mining results so the preprocessing techniques are used to improve the quality.

In this paper, Data Cleaning is applied to remove the noise in the weather dataset. The Kullback-Leibler Divergence supports only the positive attribute values in the weather dataset so the negative attribute values are considered as a noise and it has removed from the weather dataset.

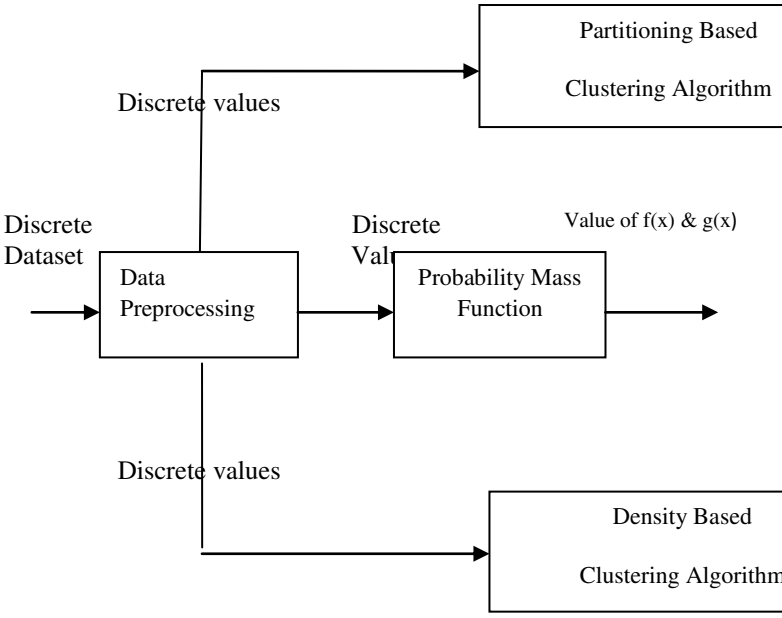


Fig1. Distance and Distribution Measurement Clustering Algorithms

Figure 1 describes how the Distance Based Clustering Algorithm like Partitioning and Density Based Clustering Algorithm works with the Distribution measurement of KL divergence.

2) Measurement of Kullback-Leibler Divergence for Partitioning Based Clustering Approach

The Kullback-Leibler Divergence (KL divergence for short) also known as Information entropy

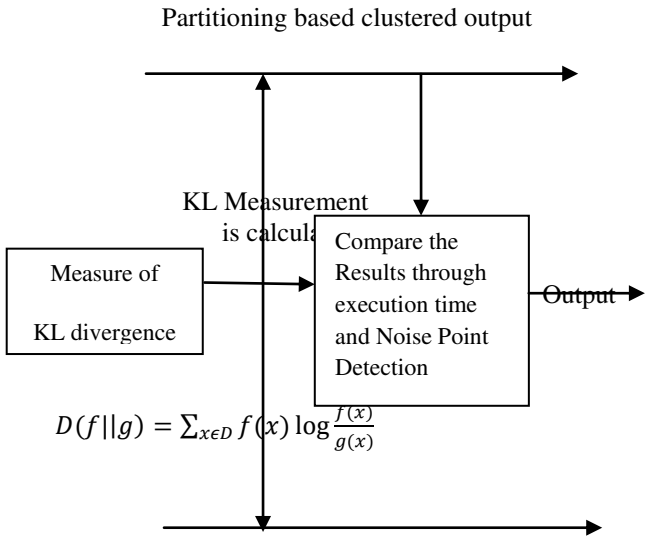
or Relative entropy. In the discrete case, let f and g are two probability mass functions in a discrete domain D with a finite number of values. The Kullback-Leibler divergence between f and g is calculated using (2) specified in proposed system.

The Kullback-Leibler divergence is used to measure similarity between the probabilities of discrete values. The Partitioning based clustering [18], [19], [20], [21], [22], [24] is calculated using (1). Finally, the KL divergence value is then integrated with the Partitioning based clustering approach.

3) Measurement of Kullback-Leibler Divergence for Density
Based Clustering Approach

Density Based Spatial Clustering of Applications with Noise (DBSCAN) Algorithm [11], [12] ,13], [14], [16], [17], [25] locates regions of high density that are separated from the regions of low density. DBSCAN is a center based approach for clustering in which density is estimated for a particular point in the dataset by counting the number of points within the specified radius, r, of that point.

The center based approach consists of points such as, Core points- These points are in the interior of the dense region. Border points- These points are not the core points, but fall within the neighborhood of the core point. Noise points- A noise point is a point that is neither a core point nor a border point.



DBSCAN ALGORITHM:

Input: The dataset D
Parameter:
∑ **Neighborhood –** Objects within a radius of r from an object.
Minpts - Minimum number of objects within the radius r.
Output: Region of high density is separated from region of low density.
Method:
If p>= minpts within the radius r
The point p is a core point
Else if p<minpts or p=neighborhood (core point)
The point p is a border point
Else
The point p is a noise point.

The Kullback-Leibler divergence between f and g is calculated using (2) and the value of KL divergence is then integrated with the density based clustering approach.

4) PERFORMANCE METRICS

The execution time of the Partitioning Based clustering using KL divergence reduced to 68 sec and Density Based clustering using KL-Divergence reduced to 65sec, 22 Noise points are detected. The execution time is calculated using Time series function. The Noise Points are calculated using equation 3.

Noise Point= [Total number of points] - [{Total number of Core point} + {Total number of Border point}]

….. (3)

V. IMPLEMENTATION

In this section we first present a Data set used then we present Measurement of KL divergence and we present our evaluation results.

A. Data set

In our experiment we use discrete data set and weather data set from National Climate Data Center (NCDC) which has 72 records, 18 Attributes and several fields like 1.Statecode 2.Division 3.year 4.Month 5.precipitation 6.Temperature 7.Palmer Drought Severity Index (PDSI) provides the rainfall information 8.CDD and HDD represents the Temperature model 9.SP provides a spring value.

B. Experimental setup

1) Hardware

Intel core 2 Duo Processor T7500, HDD with 80GB, RAM with 512MB was used

2) Software

MATLAB tool was used to implement Distance and Distribution based clustering. In this paper MATLAB with 7.5.0.342(R2007b) which was released at August 15, 2007 is used.

C. Measurement of KL divergence:

For a Random variable x={2,6} two distributions f(x) and g(x) with f(2)=x-1,f(6)=x and g(2)=y-1,g(6)=y was assumed. If the probability value using Divergence measurement wants to be calculated then

The KL divergence formula is:

$$D(f||g)=\sum_{x \in D} f(x) \log \frac{f(x)}{g(x)}$$

Given x= {2, 6} and the distribution are f(2)=x-1,f(6)=x and g(2)=y-1,g(6)=y.

On applying the given values in above formula:

We get,

$$D(f||g)= f(2) \log \frac{f(2)}{g(2)} + f(6) \log \frac{f(6)}{g(6)}$$

$$D(f||g) =(x-1)\log\left[\frac{(x-1)}{(y-1)}\right]+(x) \log \left[\frac{(x)}{(y)}\right]$$

$$= (1) \log \frac{(1)}{(5)} + (2) \log \frac{(2)}{(6)} = -1.653$$

$$D(g||f) =(y-1)\log\left[\frac{(y-1)}{(x-1)}\right]+ (y) \log\left[\frac{(y)}{(x)}\right]$$

$$= (5) \log \frac{(5)}{(1)} + (6) \log \frac{(6)}{(2)} = 6.357$$

The value of D(f||g)=-1.653 and D(g||f)=6.357 provides the KL divergence value.

D. Evaluation

The result analysis has been performed for distance and distribution based clustering algorithm based on the calculation of execution time and Noise Point Detection.

Table I describes the execution time of predefined data set and weather data set in partitioning and density based clustering algorithm without using KL divergence.

TABLE I
EXECUTION TIME WITHOUT USING KL DIVERGENCE

Clustering Algorithm	**Data set**	**Start Time**	**End Time**
Partitioning Based Clustering Algorithm without using KL divergence	Discrete Dataset	0 sec	19 sec
	Weather Dataset	0 sec	71 sec
Density Based Clustering Algorithm without using KL divergence	Discrete Dataset	0 sec	17 sec
	Weather Dataset	0 sec	68 sec

Table II describes the execution time of predefined data set and weather data set in partitioning and density based clustering algorithm using KL divergence

TABLE II
EXECUTION TIME WITH USING KL DIVERGENCE

Clustering Algorithm	Data set	Start Time	End Time
Partitioning Based Clustering Algorithm using KL divergence	Discrete Dataset Weather Dataset	0 sec 0 sec	15 sec 68 sec
Density Based Clustering Algorithm using KL divergence	Discrete Dataset Weather Dataset	0 sec 0 sec	13 sec 65 sec

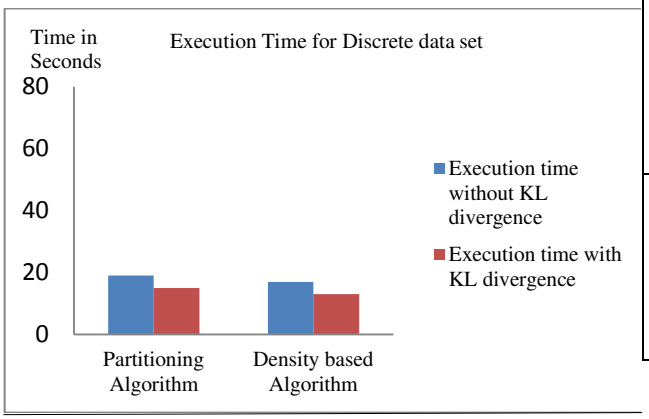


Fig 2 Execution Time for Discrete data set

Figure 2 describes the execution time for discrete data set in partitioning and density based clustering with and without using KL divergence Measurement.

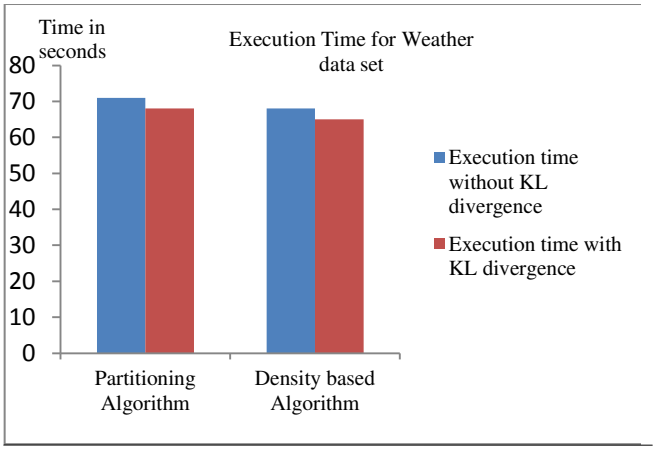


Fig 3 Execution Time for Weather data set

Figure 3 describes the execution time for weather data set in partitioning and density based clustering with and without using KL divergence Measurement.

Table III describes the Core Point Detection using Density based clustering algorithm without KL-Divergence and Density based clustering algorithm with KL divergence.

TABLE III
CORE POINT DETECTION

Clustering Algorithm	Data set	Core Point Detection
Density Based Clustering Algorithm	Discrete Dataset Weather Dataset	20 62
Density Based Clustering Algorithm using KL divergence Measurement	Discrete Dataset Weather Dataset	2 16

Table IV describes the Border Point Detection using Density Based Clustering Algorithm without KL-Divergence and Density Based Clustering Algorithm with KL divergence.

TABLE IV
BORDER POINT DETECTION

Clustering Algorithm	Data set	Border Point Detection
Density Based Clustering Algorithm	Discrete Dataset Weather Dataset	- 4
Density Based Clustering Algorithm using KL divergence Measurement	Discrete Dataset Weather Dataset	4 4

Table V describes the Noise Point Detection using Density Based Clustering Algorithm without KL-Divergence and Density Based Clustering Algorithm with KL divergence.

TABLE V
NOISE POINT DETECTION

Clustering Algorithm	Data set	Noise Point Detection
Density Based Clustering Algorithm	Discrete Dataset	-
	Weather Dataset	6
Density Based Clustering Algorithm using KL-Divergence Measurement	Discrete Dataset	14
	Weather Dataset	22

Data	Temperature	Precipitation	DBSCAN	DBSCAN Using KL Divergence
1	4.3000	36.600	Noise	
2	2.6700	43		Noise
3	4.5700	50.100		
4	6.2600	53.400		Noise
5	3.1700	65.600		
6	3.9900	70.300		
7	4.0100	72.300		Noise
8	3.4700	71.200		
9	4.1900	64.800	Noise	
10	2.9000	56.800		Noise

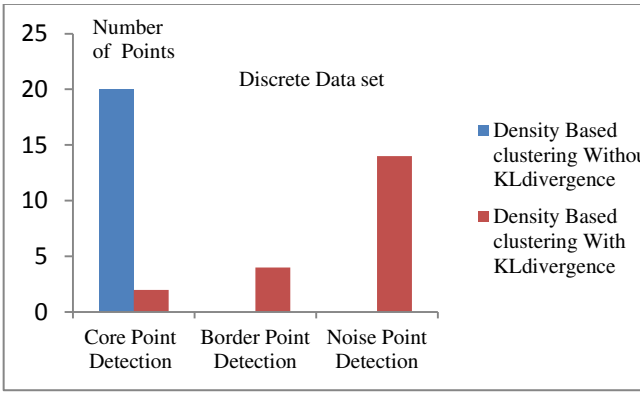


Fig 4 Core, Border, Noise Point Detection in Discrete Data set

Figure 4 describes the Core Point, Border Point, and Noise Point Detection for Discrete data set with and without using KL divergence Measurement.

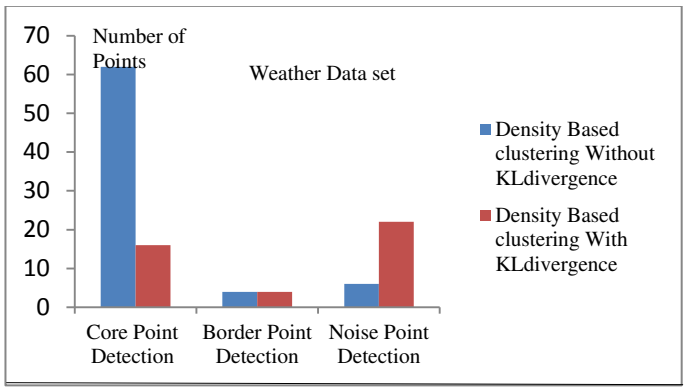


Fig 5 Core , Border, Noise Point Detection in weather Data set

Figure 5 describes the Core Point, Border Point, and Noise Point Detection for weather data set with and without using KL divergence Measurement.

Table VI shows that Noise Points for the first 10 points in weather Dataset using DBSCAN without KL divergence and DBSCAN with KL divergence

TABLE VI
NOISE POINTS

Based on the above result analysis the Partitioning and Density based clustering using KL divergence have reduced the execution time and increased the Noise Point Detection.

The Partitioning and Density based clustering using KL divergence have reduced the execution time to 68 sec and 22 Noise Points are detected.

VI. CONCLUSION & FUTURE WORK

In this paper, based on the similarity between the distributions the objects are clustered. The Kullback-Leibler divergence is used for identifying the similarity measurement between objects in discrete cases. KL divergence value is integrated into the partitioning and density based clustering methods. The extensive experiment confirms that KL divergence measurement is effective and efficient. The most important contribution of this paper is to introduce distribution difference as the similarity measure for data and the accuracy is evaluated based on execution time and Noise Point Detection. The Partitioning and Density based clustering using KL divergence have reduced the execution time to 68 sec and 22 Noise Points are detected.

In Future Hybrid clustering Algorithms which consist of Partitioning and Density based clustering will be implemented and their performances will be compared with the existing Partitioning and Density based clustering algorithm.

REFERENCES

[1] Alie J Sajid N.A," *Critical Analysis of DBSCAN Variations*", IEEE Transactions on information and Emerging Technology , Year: 2010, pages:258-269.

[2] Huang, J.Z. Yunming Ye, "*k Means: Automated Two-Level Variable weighting Clustering Algorithm for Multi View Data*", IEEE Transactions on Knowledge and Data Engineering, Volume: 25, Issue: 4, Publication Year:2013, Page(s):932-944

[3] Lian Duan , Deyi Xiong; Jun Lee; Feng Guo, "*A Local Density Based Spatial Clustering Algorithm with Noise*", IEEE Conference on Systems, Man, and Cybernetics,Volume:5, PublicationYear:2012, Page(s): 4061-4066 .

[4] Sulaiman S.N,Isa, N.A.M, "*Adaptive Fuzzy-K-means Clustering Algorithm for Image Segmentation*", IEEE Transactions on Consumer Electronics, Volume: 56, Issue: 4, Publication Year: 2010, Page(s): 2661-2668.

[5] Pei,jein,tao, " *Clustering Uncertain Data Based on Probability Distribution Similarity*", IEEE Transactions on knowledge and data Engineering, Volume: 25, issue_4, Publication Year: 2013, Page(s): 721 – 733.

[6] Bishnu,S.; Bhattacherjee, V," *Software Fault Prediction Using Quad Tree-Based K-Means Clustering Algorithm*", IEEE Transaction on Knowledge and Data Engineering, Volume: 24, Issue: 6; Publication Year: 2012, page(s):1146-1150.

[7] Liping Jing ; Ng, M.K. ; Huang , J.Z., "*An Entropy Weighting k-Means Algorithm for Subspace Clustering of High - Dimensional Sparse*", Volume: 19, Issue: 8, Publication Year: 2010, Page(s):1026-1041.

[8] Grigorios,F. Tzortzis and Aristidis C. Likas, "*The Global Kernel k-means Algorithm for Clustering in Feature Space*", IEEE Transaction on Neural networks, Volume: 20, Issue: 1, year: JULY 2011.

[9] Isa, N.A.M. ; Salamah, S.A. ; Ngah, U.K, "*Adaptive Fuzzy Moving K-means Clustering Algorithm for Image Segmentation*", IEEE Transactions on Consumer Electronics, Publication Year: 2009,Page(s):2145-2153.

[10] M.Ester,H.P.Kriegel.J," *A Density Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*", IEEE Conferences on Knowledge and Data Engineering, Publication Year: 2010

[11] Maji, S.K. ; Patra, P.K.*,"FDCA: A Fast density based Clustering algorithm for spatial database system*", In IEEE International Transactions on Computer and Communication Technology (ICCCT),Publication Year:2011, Page(s): 21 – 26

[12] Amini, A. ; Teh Ying Wah ,et.al.,"*A Study of density-grid Based clustering Algorithms on data streams*", In IEEE International Conferences on Fuzzy Systems and Knowledge Discovery (FSKD), Volume: 3 , Publication Year: 2011, Page(s): 1652 - 1656

[13] Siyuan Liu ; Yunhuai Liu ; Ni, L. et.al.,"*Detecting Crowdedness Spot in City Transportation*",IEEE Transactions on Vehicular Technology Volume: 62 ,Issue: 4 ,Publication Year: 2013 , Page(s): 1527 - 1539

[14] Xiaopeng Yu ; Deyi Zhou ; Yan Zhou, "*A new clustering Algorithm based on Distance and density*", In IEEE International Conference on Services Systems and Services Management,Volume: 2 , Publication Year: 005, Page(s):1016 – 1021.

[15] Osman, M.K. ; Mashor, M.Y, "*Performance comparison of Clustering algorithms for Tuberculosis Bacilli Segmentation*", IEEE Transactions on Computer, Information, Publication Year: 2012, Page(s):1-5..

[16] Wu Lingyu ; Gao Xuedong, " *A Density-based clustering Algorithm for Weighted Network with Attribute Information*", In IEEE International Conference on Advanced Computer Control, Publication Year: 2011 ,Page(s): 629 - 633

[17] Cheng-Fa Tsai ; Chun-Yi Sung ," *DBSCALE: An Efficient Density-based clustering algorithm for data Mining in large databases*", IEEE International Conference on Circuits, Communications, Volume: 1, Publication Year: 2010 , Page(s): 98 - 101

[18] Huang, X. ; Ye, Y.; Zhang, H*," Extensions of Kmeans-Type Algorithms: A New Clustering Framework by Integrating Intracluster Compactness and Intercluster Separation* ,In IEEE Transactions on Neural Networks and Learning Systems, Volume: PP , Issue: 99 Publication Year: 2013 .

[19] Yue Yang ; Zhuo Liu ; Jian-pei Zhang ; Jing Yang ,"*Dynamic density-based clustering algorithm over Uncertain data streams*", In IEEE International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),Publication Year: 2012, Page(s): 2664-2670

[20] Jinhua Xu; Hong Liu*," Web user clustering analysis based on KMeans Algorithm* ", IEEE International Conference on Information Networking and Automation (ICINA), Volume: 2, Publication Year: 2010 , Page(s): V2-6 - V2-9

[21] Yi Hong; Sam Kwong," *Learning Assignment Order of Instances for the Constrained K-Means Clustering Algorithm,*" In IEEE Transactions on Systems, Man, and Cybernetics , Volume: 39 , Issue: 2 ,Publication Year:2009, Page(s): 568 – 574

[22] Czink, N., Cera, P.," *Improving clustering performance Using multipath component Distance*," In IEEE Transactions on Knowledge and Engineering, Volume: 42 , issue:1, PublicationYear:2006

[23] Jie Cao; Zhiang Wu*," SAIL: Summation-bAsed Incremental Learning for Information-Theoretic Text Clustering"*, In IEEE Transactions on knowledge and Engineering Volume: 43 , Issue: 2 ,Publication Year: 2013, Page(s): 570 – 584

[24] Siddiqui, F.U. ; Isa, N.A.M.," *Enhanced moving K-means (EMKM) algorithm for image segmentation*" ,IEEE Transactions on Consumer Electronics, Volume: 57 , Issue: 2 , Publication Year: 2011 , Page(s): 833 – 841,

[25] Vijayalakshmi,S. Punithavalli, M., "*Improved varied Density based spatial clustering algorithm with noise*", IEEE International Conference on Computational Intelligence and Computing Research, Publication Year: 2010 , Page(s): 1 - 4 .