VOL. 11, NO. 1, JANUARY 2016 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CLUSTERING OF OUTSIDERS IN HIGH DIMENSIONAL DATA WITH SELF-ORGANIZING MAPPING

S. Gayathri[1], M. Mary Metilda[2] and Sanjaibabu Srinivasan[1]
[1]Bharathiar University, Coimbatore, India
[2]Queen Marys College, Chennai, India
E-Mail: gay_sri123@yahoo.com

ABSTRACT

Many real-world problems compact with clustering of high-dimensional data, such as images, videos, text and web documents, DNA microarray data, and etc., frequently, such outside elements are clustered in the high dimensional data not addressed. In this paper, we propose an algorithm, called Local Adaptive Receptive Field Dimension Selective Self-Organizing Map of outsider in high dimensional data (LARFDSSOMOH), to cluster the data points that are in a high-dimensional subspaces and also cluster the outsiders (outside elements) that are available in the high dimensional data. The proposed mapping scheme enhances the system efficiency, by providing better quality of clustering when compared to its conventional counterpart. Finally, we explain the capability of the proposed algorithm through experiments on unnatural data as well as the natural data.

Keywords: outsider, high dimensional data, cluster, self-organizing map.

INTRODUCTION

Recent days, clustering play a vital role in data mining, which means a separation of data into groups of the same objects. It is done through the dataset, which includes data of high, medium or low dimension. Clustering is used in various fields' viz., statistics, pattern recognition and machine learning. The dataset is a collection of relative sets of information composed of individual elements. There are lots of packages for doing cluster in various dataset (Mazel. J. 2011). Various algorithms are implemented to provide efficient clustering. But clustering of high dimensional data is not possible in the traditional algorithm.

High-Dimensional data are pervasive in many areas of machine learning, signal and image processing, computer vision, pattern recognition, etc. For occurrence, image includes with billions of pixels, videos can have millions of frames, text and web documents are combined with hundreds of thousands of structures etc. Subspace clustering is an addition of traditional clustering that search to nd groups in various subspaces within a dataset (Mazel. J., 2011).

Applications of subspace clustering include unsupervised approach to detect and characterize network anomalies in this application, the subspace clustering combined with inter clustering and blindly find irrelevant things in the traffic flows. In existing mechanism involves Self Organizing Map of subspace clustering in low dimensional data. Which provides more restriction of doing clustering? Self-organizing Map is the type of artificial neural network that is skilled and used for non-monitored learning for produce low dimensional data. Self-organizing map is implemented in various concepts such as incremental clustering (Milano. M., 2004). In these, Self-Organizing map deals between the unbalanced and new coming data in the incremental algorithm. In existing work, researchers focused on clustering quality and computational cost. But they did not implement the clustering of outsiders in the high dimensional data. To overcome this problem, a novel system is proposed to implement the concept to cluster the outsiders in the high dimensional data.

In this paper, a local adaptive receptive field dimensional selective self-organizing map of Outsider in High dimensional data (LARFDSSOMOH) is implemented to cluster the outside elements in the high dimensional data. The proposed algorithm deals in four phase's viz., self -organizing map in high dimensional data, combination phase, clustering of high dimensional data and outsider in the high dimensions.

RELATED WORK

The existing Dimension Selective Self-Organizing Maps (DSSOM) With Time-Varying Structure for Subspace and Projected Clustering (Bassani.F., 2014), only concentrate on the low dimensional data as well as provides less efficiency and cluster quality. LARFDSSOM abbreviates Local Adaptive Receptive Field Dimension Selective Self-Organizing Map and it is a self-Organizing Map with a time-varying design depends on Dimension Selective elements. This algorithm provide many phases to determine the existing work, they are arrangement of LARFDSSOM, convergence of LARFDSSOM, finally in clusters of LARFDSSOM. Here, they did not concentrate on the high dimensional data as well as the outside elements in the high dimensional data. This paper mainly focuses on clustering of outsiders in the high dimensional data using self-organizing map. In this proposed system, we have implemented an algorithm to Cluster the high dimensional data and outside elements with Self

Organizing Mapping. The proposed clustering technique has determined by various steps, whichthey are, Establishment of LARFDSSOMOH, Conjunction in LARFDSSOMOH, and Clustering in LARFDSSOMOH and finally Outsider in High Dimensional data. The proposed clustering scheme resulted in enhanced cluster quality, system efficiency and thereby enhanced effectiveness of the system.

SUBSPACE CLUSTERING OF HIGH DIMENSIONAL DATA

Subspace clustering (Han Hu., 2014) (Adler. A., 2015) is the advance version of traditional clustering mechanism that attempts to find clusters in various subspaces. Frequently in high dimensional data, many elements are non-relative that makes the existing system to be noisier. This subspace clustering provides two major concepts viz., top down approach and bottom up approach. These two approaches are mainly designed for subspace clustering of high dimensional data, which defines clustering of data from a few dozen to many thousands of data. Subspace clustering has been examined additionally since traditional clustering algorithms often unsuccessful to find proper clusters in high-dimensional data.

LARFDSSOMOH

LARFDSSOMOH stands for Local Adaptive Receptive Feld Dimensional Selective Self-Organizing Map of outsiders in high dimensional data. In existing schemes, they implemented the time varying structure of LARFDSSOM, that focused on to enhance the quality of clustering properties and computational cost (Bassani. F., 2014) But this proposed LARFDSSOMOH, will focus on the outside elements in the high dimensional data, in which cluster the data from a few dozen to the many millions. Here it also concentrates on the outside elements when doing cluster. As mentioned earlier, the proposed LARFDSSOMOH mainly focus on the four processes, which provides the following, Establishment of LARFDSSOMOH, Conjunction of LARFDSSOMOH, Clustering of LARFDSSOMOH, Outlier in High dimensional data. In Establishment of LARFDSSOMOH, the nodes are formed into a cluster from arbitrarily chosen input. The winner node is always active, and if it is below the threshold value, then the new node will be inserted. The nearest node is formed by similar types of subset. The establishment and competition can be repeated by a limited number of times, if the node does not win then it is removed. In conjunction phase, all the nodes are inserted and deleted when it needed. It is similar that of the establishment phase. Here, if there is no insertion then the node will start decreasing. The clustering mechanism will start after finishing conjunction phase. In this clustering phase, it groups the outside elements in the high dimensional data. Finally, LARFDSSOMOH consists outsider (Outside elements) phase.

ALGORITHM

Algorithm 1: Establishment in high dimensional data
Input: a (r), e (a, e(x), β,q, N majority N (maj), Major Component maj comp, m (p), t majority t (maj)
begin the map with one node with C (j) initialized at the first input pattern, R (j) ←0, R (k)← 1 and win (sn)← 0;
begin the variable n wins ← 1;
for t←0 to t (maj) do
provide a arbitrarily chosen input pattern X to the map;
calculate the activation of all nodes (2);
find the winner s with large activation (as) (equ1);
if a(s)< a(r) and N< N maj then
create new node j and set: w(j) ←X, R (j)← 0 and
win S(n) ← l(p) x n win (s);
connect j to the other nodes as per equ (7)
else
update the distance vector s (s) of the winner node and its nearest (5);
update the relevance vector w(s) of the winner node and its neighbors (6);
update the weight vector C (s) of the winner and of its nearest (4);
Set winS (s) ← win S (s) +1;
If nwin (s) = maj (comp) then
delete nodes with win s(n) < l (p) x maj (comp);
update the connections of the remaining nodes as per (7)
reset the number of wins of the remaining nodes;
wins (n)← 0;
nwin (s)←0;
end
n wins ← n wins+1;
end
proceed conjunction algorithm (alg 2)
Algorithm 2: Conjunction in high dimensional data
while true do
N majority N maj;
N maj←N
delete nodes with win s (n) < l (p)x maj comp;
if (N= N maj) then give back
Update connections of all nodes;
win s(n) ←0;
for t← 0 t maj do
Provide a arbitrarily chosen input pattern X to the map;
Compute the activation of all nodes (2);
find the winners with the highest activation (1);
update the distance vector s (s) of winner node and of its nearest (5);
update the relevance vectors R (j) of the winner node and of its nearest (6);
update the weight vector C (s) of the winner and of its nearest (4);

set win s(n) ← win s(n)+1;
end
end
Algorithm 3: clustering with LARFDSSOM in
High Dimensional data
Input: High Dimensional data HD (data)
Output: clustering of HD (data)
initiate
a) for each input pattern y in the high dimensional dataset [alg4] do
b) Provide y to the map in HD (data)
c) Compute activation of all nodes in HD (data) [2]
d) if a(c) >= a(d) then
e) repeat
assign y to the cluster with winner node s;
if estimate cluster then
break;
find the nearest node in the map with majority activation a(c) irrelative to the past winners
until |a(c) < a(d)
else
assign y to the outsider in HD (data) [Algorithm 4]
find the clustering of outsider in HD (data) [Algorithm 4]
end
end
Algorithm 4: outsider in High Dimensional data

Input: Number of nearness determined, High Dimensional data HD (data), P, M, number of outsiders, Major distance Major (dist), Minority min
Output: outsider is HD data
Assume: Coterminous (o, R, P) give back P elements in HD (data), Major (dist) (o, R) give back Major (dist) between o, R in HD (data)
initiate
a) C(min) ← 0
b) HD (data)(0) ← null
c) For all segmentation s in M do
d) For all elements e in s do
Nearest ← null
{Finding for nearest in HD (data)}
{high dimensional data}
for n=j , n greater than 0 then increment n do
end for
for all segmentation n in M, beginning by segmentation s do, v in t do, v ≠ 0 do
if |nearest (0) | >= P and Cp (min)
Majority (dist)(0, nearest(0)) then
break
end if
end for
O= outsider in HD data
end for
end for
end

The above mentioned algorithms depict the methods of four phases of LARFDSSOMOH. In Establishment phase, nodes form the cluster is chosen randomly as input. The winner node is always an active mode. As mentioned earlier, if the winner is below the threshold value, then the node is removed. Then the nearest node formed by a similar type of data is chosen. Next in the Conjunction phase the same procedure followed. In this phase also, if the node will not win, then it is removed. Also in the conjunction phase all the nodes are inserted and deleted when it needed. In conjunction, if there is no insertion, then the nodes will start decreasing. Then in the clustering phase all the nodes are clustered in high dimensional data based on the distance factor. Finally, in Outsider phase, all the nodes are cluster as well as the outsider node.

In this establishment of LARFDSSOMOH, the nodes are from a clutch when the inputs are chosen as random manner. In this phase, the winner node is always in active mode. Here the new node will insert when the active node below the threshold range. The neighbor node is implemented by the same type of subset elements. The enactment and competition can be recurring by a limited amount of time duration. At the end the node will delete when it does not attain the winning target.

CONJUNCTION OF LARFDSSOMOH

In this phase, all nodes are inserted and removed depends on the requirement. Here, it is same that of the establishment phase, which it is used to form the clutch, defines the winner node, and implementation of a neighbor node. The nodes are decreased when there is no insertion of elements. Finally the clustering or clutching mechanism will start after that of conjunction phase.

CLUSTERING OF OUTSIDER IN HIGH DIMENSIONAL DATA

In this phase, combines each and every outsider element in the high dimensional data; Here the outsider, which means outside of elements are in the high dimensional data or uncover elements in the high dimensional data. In this phase of clutching of data in high dimensional, start clutch from few dozen and to the various millions of data. Aggregate of this phase, it cluster all elements of outsiders in high dimensional data.

MATHEMATICAL CALCULATION

In this insertion node, first the highest activation value can be calculated by,

ac (D (w) (x, c (j)), W(j)) → input pattern
R(x) → highest activation value
R (x) →argmaj(j) [mn(D(w) (x(i), P(j)) W(j)] → (1)

The Circular basis function can be determined by,

mn (D(w) (x,P(j)), w(j)) = 1/ 1+M/ (|| s ||2 + €) → (2)

R → relative vector
D (w) (x, P (j)) → distance of weighted component
€ → indicates less value

D (w) (x, nj)= $\sqrt{\sum_{i=1}^{y} wji\ (x\ (i) - nji)2}$ → (3)

Then the activation threshold in LARFSOMHO
a(t) → threshold value here, if the threshold value below then insert new node

To update winner and nearest node by using following,

n (j) (z+1) = nj(z) + l(x-nj(z)) → (4)
l= l(b)
l=l(x)

Updating via the input moving average of the distance between input and current center vector

n (j) = s (j)
s(j) (z+1) = (1- l (β) s (j) (z) + l(β) (| z- n(j) (z)|) → (5)

Relative vector calculated by an inverse logistic function in high dimensional data,

w (ji) = $\begin{cases} 1/\ 1 + \ \exp\ (A - \ sji/s\ (B - C)) \\ 1 \quad ; otherwise \end{cases}$ → (6)

here, A= s (ji) mean
B= s (ji) maj
C= s (ji) min

A, B, C are maximum, minimum and mean component of distance vector.

To update a nearest node by the following,

i and j nodes are in the high dimensional data =
$\begin{cases} connected\ if\ ||a(i) - a(j)|| < n\sqrt{m} \\ disconnected\ ; otherwise \end{cases}$ → (7)
m→ input dimensions

To detect outlier data, by the following
OHD (maj) = $(a(\max) - a(\min))/(s(n) - a(\min) + n))\{> (\left(\frac{2}{n}\right) + w(i))$ majority

$\leq ((\frac{2}{n}) + w(i))\}$→ (8)

OHD (min) = a(max)- a (min)/ a (max) * n- s (n) { ≤ $((\frac{2}{n}) + w(i) > ((\frac{2}{n})) + w(i));\ \}$ → (9)

From (8) and (9) equation shows the minimum and maximum outliers in the high dimensional data.

CASE STUDY ON IMPLEMENTATION OF PROPOSED ALGORITHM

To verify the efficiency of the proposed algorithm, an educational system model has been implemented. The information about students in an university are taken into account. The university data set is the high dimensional data set, as illustrated in Figure-1. The data set consists of type of admission viz., regular, part time student, good character student, bad character, not interested candidate, money oriented problem, quit without any reason and academic percentage. This case study is implemented with the Self Organizing Map and to cluster the outsiders in the high dimensional data. From these parameters clustering of outside elements in high dimensional data can be achieved.

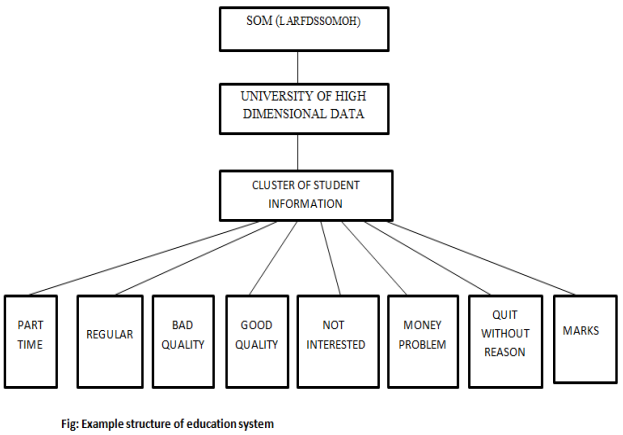
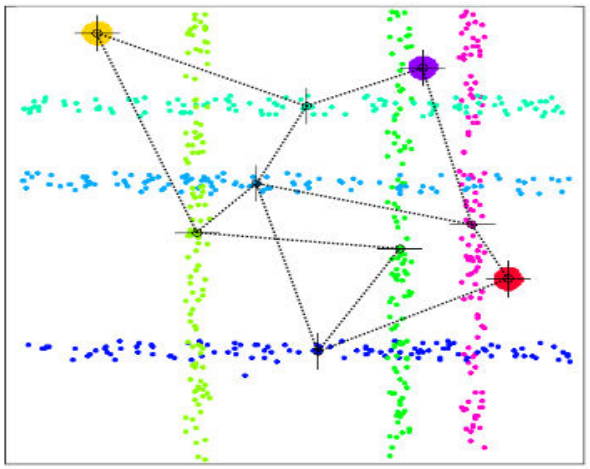
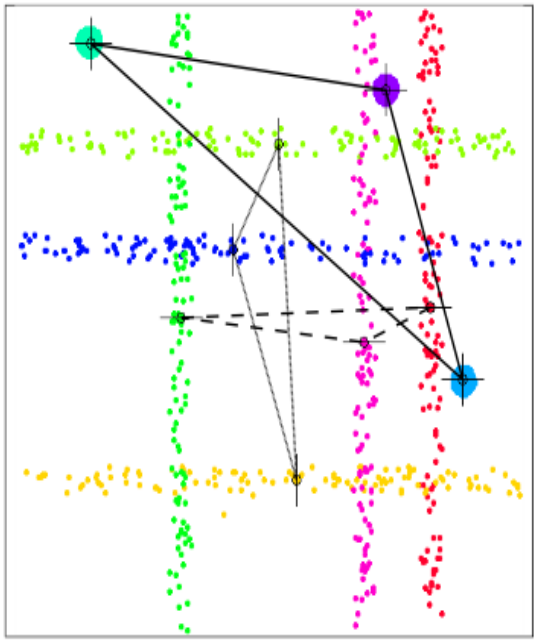


Figure-1. Case Study - Information about Students

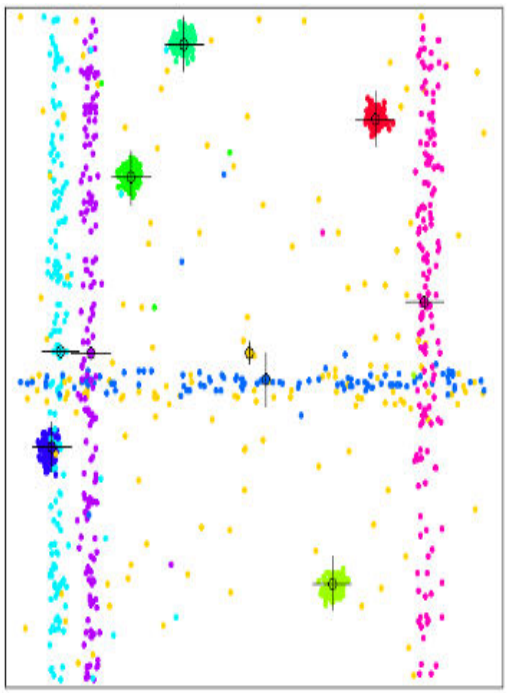
IX. Performance analysis.



(a) Formation of first nodes



(b) Formation of relative and neighbourhood nodes



(c) Formation of long distance clustering

Figure-2. Formation of nodes.

Figure-2 depicts the formation of nodes by DSSSOM method. Figure-2a, implicit the formation of first nodes and Figure-2b, illustrates the formation of relative and neighborhood nodes. The formation of long distance node is shown in Figure-2c. Once the nodes are formed, then clustering the nodes are done based on the data set. Figure-3 depicts the method of clustering the inside elements.

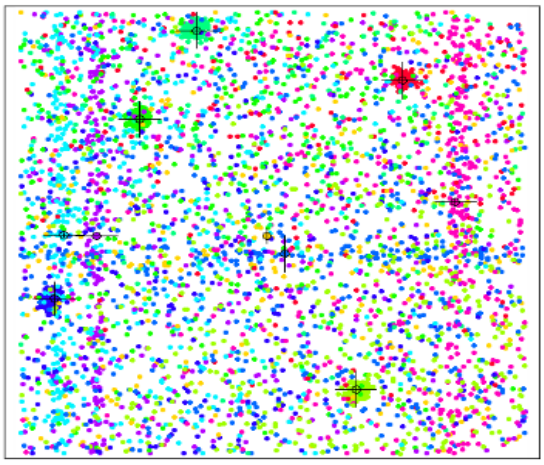
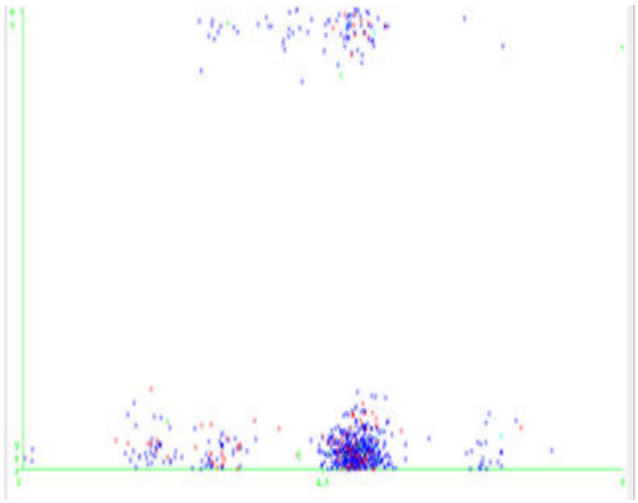
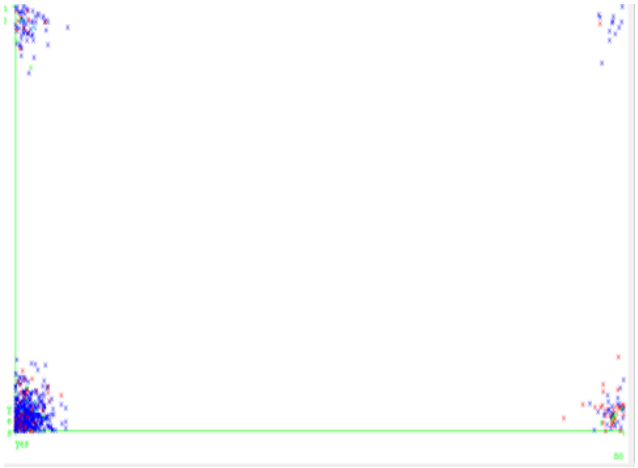


Figure-3. Clustering of inside elements.



(a) Part time Vs Regular



(b) Money Problem Vs Student

Figure-4. High dimensional clustering.

Figure-4 illustrates the analysis of information and how the students cluster is formed. The parameters taken up for study include part time students Vsregular and the money problem Vs students. The nearest clustering is formed after forming the high dimensional data of the student as regular or part time. Once that cluster is made, then with the students, the money problem is formed as a nearest cluster. The implementation of these graphs infers how efficiently the clustering of nodes, clustering of relative and nearest nodes, projection of nodes and clustering of nodes with distance are done with the proposed algorithm.

Table 1

Parameters of LARFDSSOM

NUMBER OF NODES	MINIMUM 2	MAXIMUM 18
Learning rate (A)	0.001	0.1
Learning rate decay (B)	0.01	2
Neighborhood (C)	0.001	0.99
Neighborhood Decay (D)	0.01	2
Relevance rate (E)	0.001	0.1
Min Relevance (F)	0	1
New winner (G)	0	1
Maximum Winner (H)	1	3
Relevance rate1 (I)	0.001	0.1
Relevance rate 2 (J)	0.8	0.99
Relevance rate 3 (K)	1.01	1.2

Table 2

LARFDSSOM - OUTSIDER IN HIGH DIMENSIONAL DATA

NUMBER OF NODES	MINIMUM 40	MAXIMUM 200
Learning rate (A)	7.42	49.45
Learning rate decay (B)	70.05	70.05
Nearest (C)	46.07	46.07
Nearest Decay (D)	70.05	70.05
Relevance rate (E)	46.7	46.7
Min Relevance (F)	9.47	9.47
New winner (G)	9.62	9.62
Maximum Winner (H)	25.40	25.40
Relevance rate1 (I)	7.42	49.45
Relevance rate 2 (J)	10.75	50.07
Relevance rate 3 (K)	**20.46**	**55.47**
Outlier threshold (L)	**1.46**	**9.75**

The Tables 1 and 2 infers the parameters of the LARFDSSOM and LARFDSSOMH model. The parameters tabulated illustrate that when the number of nodes is maximum in both the LARDFSSOM and LARFDSSOMH models, the learning rate increases with increase in the number of nodes. Similarly the relevance rate is also increased with increase in the number of nodes. These points infer that the efficiency of the proposed algorithm is much better compared to its conventional counterpart.

CONCLUSIONS

From this, Clustering of Outsiders in High Dimensional data with Self Organizing Mapping was studied and it mainly focuses on the Self Organizing Map for high dimensional data as well as the outsiders. The proposed LARFDSSOMOH model has been implemented in a case study. In this the novel algorithm has been created for clustering outside elements in the high dimensional data. The LARFDSSOMOH algorithm used to cluster the high dimensional data with high efficiency than the existing methods. The analysis of the same proves that the proposed model provides enhanced effectiveness of clustering. In future we have to increase the algorithm performance to enhance the system performance.

REFERENCES

Adler. A. 2015. Linear-Time Subspace Clustering via Bipartite Graph Modeling. 99: 1.

Bassani.F. 2014. Dimension Selective Self-Organizing Maps with Time-Varying Structure for Subspace and Projected Clustering. 26(3): 458-471.

Coelho. M. 2011. Subspace clustering in information retrieval in urban scene databases. pp. 173-180,

Gan. G. 2006. PARTCAT: A subspace clustering algorithm for high dimensional categorical data. pp. 4406-4412.

Han Hu. 2014. Exploiting Unsupervised and Supervised Constraints for Subspace Clustering. 37(8): 1542-1557.

Mazel. J. 2011. SubSpace clustering, Inter-Clustering Results Association and anomaly correlation of unsupervised network anomaly detection. pp. 1-8.

Milano M. 2004. Self Organizing Nets for optimization. 15(3): 758-756.

Moise. G. 2015. Finding non-redundant, statistically significant regions in high dimensional data: A novel approach to project and subspace clustering. pp. 533-541.

Vidal. R 2011. Subspace clustering. (28): 52-68.

Zhenya Zhang. 2008. Clustering aggregation based on genetic algorithms for documents clustering. 3156-3161.