MAFCA-S: A Subspace Clustering Technique for High Dimensional Dataset

Dharmveer Singh Rajput
drdharmveer16382@gmail.com
Jaypee Institute of Information Technology, Noida

***Abstract*- Existing clustering methods are not able to determine efficient clustering of big dataset which has large number of dimensions, because first, it may contain many irrelevant dimensions and second, some time different clusters may exists in the different subsets of high dimensional dataset i.e. Subspace Clustering. The first problem can be solve by using feature selection approach on high dimensional dataset then create clusters of the reduced dataset. However, they do not deal with the concept of subspace clustering. In this paper, we propose MAFCA-S (Median absolute deviation And most Frequent value based high dimensional data Clustering Algorithm for Subspace) which extends the idea of feature selection based clustering method MAFCA (Median Absolute deviation and most Frequent value based high dimensional data Clustering Algorithm) to subspace clustering and works well with the high dimensional dataset consisting of attributes in continuous variable domain. The experimental results show that MAFCA-S performs better as compare to traditionally subspace as well as feature selection based clustering methods.**

***Keywords*- Projected clustering, subspace clustering, dimension reduction, feature selection and high dimensional dataset**

I. INTRODUCTION

Data clustering is a process to determine similar groups of instance in the dataset where the instances having similar features belong to one group and the instances having different features belong to another group. A group of similar instances is known as a *cluster* and the concept of similarity among the instances is defined by some similarity measure / distance measure [9]. The conventional clustering methods are broadly classified as *partitioning methods, hierarchical methods, density based methods, grid based methods* and *model based methods.* All these clustering methods, irrespective of the category they belong to, uses distance, e.g., Euclidean distance, Manhattan distance, as a similarity measure to group the instances. As the distance becomes meaningless in large dimensions, i.e., the instances are almost equidistant in sufficiently high dimensions, these methods do not efficiently work in high dimensional dataset. Apart from this, high dimensional datasets usually contain large number of insignificant dimensions which hide clusters in the sea of noise [14].

Dimension reduction techniques give a solution to minimize the problems of high dimensionality. These techniques are divided as *feature extraction techniques* and *feature selection techniques*. Feature extraction techniques such as principal component analysis, singular value decomposition, factor analysis, linear Discriminant analysis and multi dimensional scaling perform some linear transformation on the high dimensional dataset to map it into a low dimensional dataset. These techniques efficiently reduce the effect of high dimensional dataset however it fail to eliminate the ill effects of the insignificant dimensions as they maintain the original relative distance among the instances [5]. Feature selection techniques such as filter method and wrapper method select subset of significant dimensions from the high dimensional dataset based on some statistical measures i.e. MAFCA [15]. The feature selection based clustering methods determine convenient clusters in high dimensional dataset however all the clusters are formed only on the selected relevant dimensions whereas different clusters may exist in different subset of dimensions in high dimensional dataset, i.e., different dimensions may be relevant for different clusters. It arise the need of *subspace clustering*.

It is well explained by an example due to [14]. Figure 1 shows a dataset consisting of four hundred instances in three dimensions. It comprises four clusters of hundred instances each; two clusters belong in dimensions *a* and *b* whereas the other two clusters belong in dimensions b and *c*. The conventional clustering techniques do not determine all the four clusters separately as one dimension is insignificant. Feature extraction techniques also do not help to determine all the four clusters as they maintain the relative distance among the instances, hence it does not eliminate the bed effects of the insignificant dimensions. Feature selection techniques applied to project the dataset onto any of a single dimension also do not help as none of the projection clearly identifies all the four clusters separately (refer, figure 2). A projection of the dataset onto the dimensions *a* and b clearly identifies only *red* and *green* clusters (refer, figure 3(a)), a projection onto dimension *b* and *c* clearly identify only *blue* and *purple* clusters (refer, figure 3(b)) whereas a projection onto the dimension *a* and *c* does not completely separate the clusters (refer, figure 3(c)). Thus, we see that none of the combination of selected dimensions describes all the four clusters separately as the clusters belong to different subsets of dimensions.

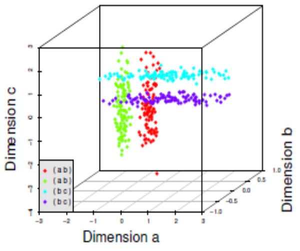
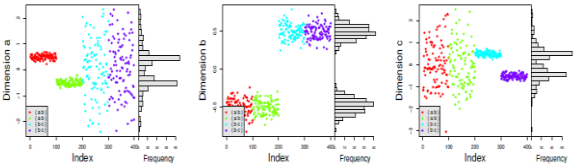
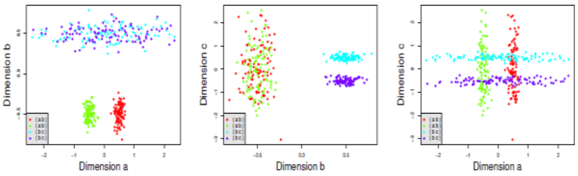


Figure 1: Sample dataset [14]



(a) Dimension *a* (b) Dimension *b* (c) Dimension *c*

Figure 2: Sample data plotted in one dimension [14]



(a) Dims *a* & *b* (b) Dims *b* & *c* (c) Dims *a* & *c*

Figure 3: Sample data plotted in two dimensions [14]

Subspace clustering is a appropriate approach to identify clusters in different subsets of dimensions in high dimensional datasets. These methods are broadly classified as *Bottom-up subspace clustering methods* and *Top-down (projected) subspace clustering methods* [14]. The earlier methods, e.g., *CLIQUE* [3], *ENCLUS* [6], combine the concepts of grid based clustering methods and density based clustering methods to obtain subspace clusters. Though these methods successfully obtain subspace clusters of different shape and varying size of subspace, they are sensitive to input parameters, e.g., proper tuning of grid size and density threshold. The latter methods, e.g., *PROCLUS* [1], *ORCLUS* [2], and *PCKA* [4] combine the concepts of partitioning clustering methods and feature selection methods to identify the subspace (projected) clusters. The time and space complexity of these methods are comparatively less as they use samples of the original dataset. However, they produce hyper-spherical clusters of fixed size subspace and take number of subspace clusters *K* as an input parameter.

In this paper, we propose a novel and efficient top-down subspace clustering method MAFCA-S (Median absolute deviation And most Frequent value based high dimensional data Clustering Algorithm for Subspace), which extends the idea of finding clusters in the reduced dataset of MAFCA [15] to the projected clustering and works well with the high dimensional dataset consisting of attributes in continuous variable domain. It works in four phases: *Sampling phase, Initialization phase, Dimension selection phase and Refinement phase*. We test the performance of our propose method MAFCA-S on three real and two synthetic datasets and compare the results with MAFCA, PROCLUS and PCKA. We use three well-known subspace clusters quality measures *Jagota index (Q), SCQE* and *Sum of Squared Error* (SSE) and Student's t-test too to verify the results statistically. The results and quality measures indicate that the MAFCA-S is superior to MAFCA and its competitors in the domain of subspace (projected) clustering.

The rest of the paper is organized as follows. Section II provides a brief literature survey of clustering of high dimensional dataset. Section III details our proposed top-down subspace clustering method MAFCA-S and Section IV provides the results and discussion of MAFCA-S with other methods. Finally, we conclude in Section V and provide suggestions for future research.

II. LITERATURE REVIEW

Subspace clustering is an extension of feature selection based clustering to high dimensional dataset. Here, it is intended to identify relevant dimensions for each meaningful cluster as different subset of dimensions may be relevant for different clusters in high dimensions. Subspace clustering methods are broadly classified as bottom-up subspace clustering methods and top-down subspace clustering methods [11], [13], and [14].

A. Bottom-Up Subspace Clustering

CLIQUE [3] regarded as the first bottom-up subspace clustering method, which combines the concepts of grid based and density based clustering methods to identify dense subspaces. The subspaces having density above a given threshold are selected while rest are pruned. Further, subspace clusters are obtained in dense subspaces using Disjunctive Normal Form (DNF) expression. Though it finds clusters of varying subspace and different shapes, it occasionally removes some small but important clusters in pruning stage. Besides, quality of the obtained results highly depends on input parameters grid size and density threshold. ENCLUS [6] borrows heavily from the CLIQUE; it determines the relevant subspaces using entropy measure instead of direct computation of density and coverage. However, it suffers from the same problems as CLIQUE. Chu et al. [7] introduce DENCOS which tackles the issue of diversity divergence. The authors use DFP - tree for threshold values of all dense subspaces which automatically vary at every stage. This automatic varying threshold value gives efficient dense subspace cluster.

B. Top-Down Subspace Clustering

PROCLUS [1] is regarded as the first top-down subspace clustering method. It obtains projected clusters in three phases: initialization phase, iteration phase and cluster refinement phase. The initialization phase selects a sample dataset from the full dataset and finds a set of potential medoids which are far apart from one another in the sample dataset using greedy approach. The iteration phase, assign points to the medoids using average Manhattan segmental distance, then determine the subspace for each medoids and remove the bed medoids, which has minimum deviation points. The cluster refinement phase reassigns the objects to medoids and removes the outliers. Though it is fast because of sampling and robust to the outliers, it is sensitive to its input parameters and is biased towards the clusters that are hyper spherical in shape. ORCLUS [2] is an extension of PROCLUS. It works in three phases: assign clusters phase, subspace determination phase and merge phase. The assign cluster phase partitions the dataset into predefined *K* groups by assigning the objects to their nearest cluster centre. Then, relevant subspace is defined in the subspace determination phase, using smallest Eigen value. Finally, the merge phase combines the nearest clusters which have similar direction. However, it suffers from the same problems as PROCLUS and sometime removes small but meaningful clusters. Bouguessa and Wang [4] introduce PCKA which identifies dense regions of each dimension then forms the clusters which have sufficient density in their subspace. Wang et al. [16] propose K-subspace clustering model which uses the distance minimization function and significant Eigen values to obtain the clusters of different shape such as line, plane and ball. Kumar and Puri [12] modify the Gustafson Kessel objective function for the projective clustering so that the relevant subspace for each cluster is automatically identified. It enhances the efficiency of clustering by simultaneously pruning the irrelevant subspaces. Gunnemann et al. [8] propose ASCLU (Alternate Subspace CLUstering) which considers subspace clusters as its input and obtains new subspace clusters using different parameters. However, it does not always produce better results.

III. PROPOSE TECHNIQUE

This section describes our propose method MAFCA-S which is an extension of MAFCA to the subspace clustering.

A. MAFCA

Rajput et al. [15] introduce feature selection based MAFCA (Median Absolute deviation And most Frequent value based high dimensional data Clustering Algorithm) for clustering high dimensional dataset. It consists of three phases: *selection of relevant features, identification of effective initial clusters centres* and *refinement phase*. In first phase, selection of relevant features, it computes the median absolute deviation (MAD) of each dimension and selects *d* dimensions which have minimum values of MAD. These selected dimensions are considered as relevant dimensions in the high dimensional dataset. In second phase, identification of effective initial clusters centres, it sorts the reduced dataset based on the dimension having minimum value of MAD and creates *k* equal partitions of the sorted reduced dataset. Further, it selects tuples consisting of most frequent values (MODEs) of dimensions in each partition. These tuples act as initial clusters centres. Finally in refinement phase, it performs an iterative process to obtain *k* clusters in the reduced dataset. The MAFCA performs efficiently if all the clusters exist in the same subset of relevant dimensions; however, it fails to obtain meaningful clusters if the clusters exist in subsets of different dimensions.

B. MAFCA-S

Here, we present a top-down subspace clustering method MAFCA-S (Median absolute deviation And most Frequent value based high dimensional data Clustering Algorithm for Subspace), which extends the concept of MAFCA to the projected clustering and works well with the high dimensional dataset consisting of attributes in continuous variable domain. It identifies projected clusters in high dimensional dataset in four phases: *Sampling phase, Initialization phase, Dimension selection phase* and *Refinement phase*. The sampling phase initially transforms the high dimensional dataset into one dimensional dataset using principal component analysis (PCA) and sorts the transformed data in ascending order. Then, systematic sampling is applied to select *S* (sample size) objects from the high dimensional dataset. These selected objects constitute a representative sample of the high dimensional dataset. The initialization phase computes the median absolute deviation (MAD) of each dimension of the sample dataset, sorts it with respect to the dimension having the maximum MAD value, and partitions it into *k* equal parts. Further, tuples consisting of the most frequent values (MODEs) of each dimension in each partition is found which serve as initial clusters centres in the sample dataset. The dimension selection phase assign each object of sample dataset to its nearest initial cluster centre to form *k* initial clusters. Then *d* dimensions of every cluster are selected based on the minimum MAD values of the dimensions. Finally in refinement phase, it performs an iterative process on the full dataset using predefined relevant dimensions and initial clusters centres to obtain meaningful projected clusters. The sequential steps of MAFCA-S are shown below;

```
Input:
  N x D dataset: Here, N represents the number
  of objects and D denotes the number of
  dimensions in the dataset.
  K: The number of clusters
  S: Size of the sample
  d: The number of relevant dimensions of the
  cluster
Output:
  K projected clusters of dimensions d, where d
  < D.
Algorithm:
Sampling Phase
  1. Transform the dataset into one
     dimensional data using PCA.
  2. Sort the transformed one dimensional data
     in ascending order and select ‖N/2S,
     3N/2S, 5N/2S... (2S-1)N/2S‖objects from
     the dataset.
  3. The set of objects selected in step 2, in
     the full dimensional dataset, form sample
     dataset.
Initialization Phase
  4. Compute the Median Absolute Deviation
     (MAD) of each dimension of sample data
         $MAD = median(abs(X - median(X)))$
  5. Sort the sample dataset in ascending
     order indexing on the maximum Median
     Absolute Deviation dimension.
  6. Partition the sorted sample dataset into
     K equal partitions and compute Most
     Frequent Value (MODE) of each dimension
     in each partition. These values form
     tuples serve as initial clusters centers.
         $MODE = \mathrm{mod}\,e(X_i)$
Dimension Selection Phase
  7. Assign each object of the sample dataset
     to the nearest initial cluster center.
  8. Compute the Median Absolute Deviation
     (MAD) of each dimension in each cluster.
         $MAD = median(abs(X - median(X)))$
  9. Select the d dimensions having minimum
     Median Absolute Deviation in each
     cluster.
Refinement Phase
  10. Redefine initial clusters centres with
      respect to the selected (d) relevant
      dimensions in each cluster.
  11. Assign each object of full dataset to the
      nearest cluster centre.
  12. Re-compute the centre of every cluster.
  13. Repeat steps 11 – 12 until clusters
      centres stabilizes.
```

The sampling phase, i.e., the first three steps, uses PCA to transform full dimensional dataset into one dimensional dataset and apply systematic sampling on sorted transformed dataset to obtain a good representative sample of the original dataset. The sampling reduces time and space complexity of the method. Here, we use PCA for data transformation as PCA is very simple and effective data transformation method and use systematic sampling method to minimize the bias in the sample.

The initialization phase, i.e., steps 4 – 6, identify efficient initial clusters centres using the most frequent value (MODE). As MODE is the number that appears most often in the dataset, it is very robust to outliers. Though, some other methods such as arithmetic average, geometric mean, harmonic mean etc. are also available in the literature to determine the central tendency of the dataset, they are not as effective and robust as the MODE. Average is a simple and popular measure to identify the central location of dataset but it applies only on the normally distributed dataset and is very sensitive to the outliers as one bad data can move the average value away from the center of the rest of the data by an arbitrarily large distance. The geometric mean and harmonic mean are suitable for log normally distributed dataset and are also sensitive to the outliers. The MODE is the most frequently appeared value in the dataset, which changes only slightly if data has large perturbation to any value, hence it is more robust to outliers.

The dimension selection phase, i.e., steps 7 – 9, find the relevant dimensions to every cluster based on the minimum median absolute deviation (MAD). The MAD is used to determine a relevant subset of dimensions so that distance between the data objects in deferent clusters is as large as possible while distance between the data objects in the same clusters is as small as possible. Though, some other methods such as range, standard deviation, variance, mean absolute deviation etc. are also available in the literature to determine the dispersion in the dimensions of the dataset. Range is very easy measure of dispersion but it is very sensitive to the outliers because it computes the difference between the maximum and minimum value of each dimension. The standard deviation and variance are also sensitive to the outliers in the presence of bad data. On the other hand, the mean absolute deviation is less sensitive to outliers, as it does not move quite as much as the standard deviation or variance in response to bad data. The median absolute deviation is more robust to outliers in comparison to range, standard deviation and variance.

Finally, the refinement phase, i.e., steps 10 – 13, uses K-means to obtain final projected clusters.

IV. EXPERIMENTAL RESULTS

In this section, we show the clustering results obtained by our proposed method MAFCA-S on three real and two synthetic datasets, and compare them with the results of other well - known top-down subspace clustering methods (PROCLUS and PCKA) and feature selection based non-subspace clustering method MAFCA. The obtained results are verified by three well-known subspace clustering quality measures Jagota index (Q), SCQE [10] and Sum of Square Error. The minimum value (Bold face) of all quality measures indicates a better quality of clusters. We also use Student's t-test to determine significant difference between clustering results, if required. The real and synthetic datasets and experiments with different input parameters to the clustering methods are described in the next subsections.

A. Description of Datasets

In this experiment, we use three real datasets Wisconsin Prognostic Breast Cancer[1] (WPBC), Heart Disease Data[2] (HDD) and Image Segmentation Data[3] (ISD), and two synthetic datasets Point20dCCNorms[4] (P20D) and Point70dCCNorms[5] (P70D). The description of all the datasets such number of objects and number of dimensions is presented in Table 1.

Table 1: Description of Datasets

Datasets	Objects	Dimensions
WPBC	198	34
HDD	303	14
ISD	2310	19
P20D	1,00,000	20
P70D	1,00,000	70

B. Subspace Clustering with Two Relevant Dimensions

In this experiment, we consider two relevant dimensions in each dataset for every subspace cluster, 10% sample size for every dataset and assume that WPBC, HDD, ISD, P20D and P70D datasets contain 10, 5, 7, 5 and 5 clusters respectively. The computed quality measure values for the obtained clusters are shown in Table 2. We observe that the proposed method MAFCA-S obtains better subspace clusters in most but not in all the cases. The MAFCA performs comparatively better in HDD and ISD datasets based on the (Q, SCQE, SSE) and Q indexes respectively. The PCKA performs comparatively better based on the WPBC and P20D datasets based on the Q and SCQE indexes respectively. Therefore, we apply Student's t-test for determining the significant difference between the clustering results of quality measure obtained by MAFCA, PROCLUS, PCKA and our proposed method MAFCA-S.

Table 2: Qualitative Results of Subspace Clustering with Two Relevant Dimensions

Data Set	Quality Measure	MAFCA	PROCLUS	PCKA	MAFCA-S
WPBC	Q	1183.26	538.41	**419.34**	438.26
	SCQE	9.36	8.21	7.39	**6.27**
	SSE	2738263	15629.79	24534.84	**10837.62**
HDD	Q	**2.41**	128.48	234.19	28.73
	SCQE	**5.32**	9.05	8.14	6.45
	SSE	**82.71**	27475.8	81142.86	427.74
ISD	Q	**1.34**	124.62	127.07	8.38
	SCQE	7.57	9.70	8.00	**6.32**
	SSE	100.162	236385	8879403	**85.29**
P20D	Q	832.43	706.23	971.65	**523.47**
	SCQE	6.83	7.95	**4.38**	5.31
	SSE	5378327	2760834	6551567	**952834**
P70D	Q	573.37	751.3	699.1	**488.35**
	SCQE	6.48	8.40	8.14	**3.52**
	SSE	6379426	8308540	5497792	**3842348**

The computed t-values of MAFCA-S:MAFCA, MAFCA-S:PROCLUS and MAFCA-S:PCKA are shown in Table 3. In this case, the degree of freedom is 28 as the sum of elements in two groups, i.e., $n_1 + n_2$, is 30. The two tailed alpha level is set to 0.05 as a "rule of thumb". The critical value in the student's t-test table based on these parameters is 2.048. The computed t-values are very small in comparison to the t-value in critical table and negative. Therefore, we claim that our proposed technique MAFCA-S produces better subspace clustering in comparison to other competitive methods.

Table 3: Computed t-values (d = 2)

Methods	Computed Value of t-test
MAFCA-S:MAFCA	-1.06
MAFCA-S:PROCLUS	-0.69
MAFCA-S:PCKA	-1.34

C. Subspace Clustering with Five Relevant Dimensions

In this experiment, we assume five relevant dimensions in each dataset for every subspace cluster, sample size is 10% for each dataset and assume that WPBC, HDD, ISD, P20D and P70D

[1] http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wpbc.data
[2] http://archive.ics.uci.edu/ml/machine-learning-databases/heart-disease/processed.cleveland.data
[3] http://archive.ics.uci.edu/ml/machine-learning-databases/image/segmentation.data
[4] http://uisacad2.uis.edu/dstar/data/clusteringdata.html
[5] http://uisacad2.uis.edu/dstar/data/clusteringdata.html

datasets contain number of clusters 4, 3, 5, 7 and 7 respectively. The computed quality measure values for the obtained clusters are shown in Table 4.Here also, we observe that the proposed method MAFCA-S obtains better subspace clusters in most but not in all the cases. The MAFCA performs comparatively better in HDD, ISD and P20D datasets based on the Q, SCQE and SCQE indexes respectively. The PCKA performs better in ISD and P70D datasets based on the Q index.

Table 4: Qualitative Results of Subspace Clustering with Five Relevant Dimensions

Data Set	Quality Measure	MAFCA	PROCLUS	PCKA	MAFCA-S
WPBC	Q	683	457	485	**386**
	SCQE	9.35	6.84	7.36	**5.26**
	SSE	826741	910507	950945	**693485**
HDD	Q	**638**	678	934	655
	SCQE	8.37	9.75	9.57	**4.92**
	SSE	42831	74381	65505	**21573**
ISD	Q	37.25	31.18	**24.98**	28.49
	SCQE	**6.73**	9.15	9.59	7.38
	SSE	974628	823538	3171712	**583542**
P20D	Q	851	971	950	**782**
	SCQE	**6.28**	7.09	6.46	6.93
	SSE	4735932	4984898	5853490	**2948575**
P70D	Q	3492	5472	**1493**	2184
	SCQE	53.48	35.74	47.33	**32.59**
	SSE	4723853	7572688	5678371	**3948576**

Therefore, we apply Student's t-test for determining the significant difference between the clustering results of quality measure obtained by MAFCA, PROCLUS, PCKA and our proposed method MAFCA-S. The computed t-values of MAFCA-S: MAFCA, MAFCA-S: PROCLUS and MAFCA-S: PCKA are shown in Table 5. In this case, the conditions are same as previous subsection. Here also, the computed t-values are very small in comparison to the t-value in critical table and negative. Therefore, we claim that our proposed technique MAFCA-S produces better subspace clustering in comparison to other competitive methods.

Table 5: Computed t-values (d = 5)

Methods	Computed Value of t-test
MAFCA-S:MAFCA	-0.39
MAFCA-S:PROCLUS	-0.62
MAFCA-S:PCKA	-0.80

V. CONCLUSIONS

Subspace clustering is a prominent approach to obtain meaningful clusters in different subsets of relevant dimensions. The MAFCA does not address the issue of subspace clustering. In this paper, we proposed a top-down subspace clustering method MAFCA-S, which extends the idea of MAFCA to subspace clustering and works well with the high dimensional dataset consisting of attributes in continuous variable domain. The experiments performed on various dataset verify the claim.

However, top-down subspace clustering methods are sensitive to the input parameter *k* (the number of clusters) and *d* (the number of relevant dimensions). Our proposed method is no exception to it. In our future work, we intend to develop a subspace clustering method, which minimizes the user's intervention and determines the number of subspace clusters and size of their subspace on its own.

REFERENCES

[1] Aggarwal CC, Wolf JL, Yu PS, Procopiuc C, and Park JS, "Fast Algorithms for Projected Clustering", Int. Conf. on Management of Data, ACM, pp. 61-72, 1999.

[2] Aggarwal CC and Yu PS, "Finding Generalized Projected Clusters in High Dimensional Spaces", Int. Conf. on Management of Data, ACM, pp. 70-81, 2000.

[3] Agrawal R, Gehrke J, Gunopulos D, and Raghavan P., "Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications", Int. Conf. on Management of Data, ACM Press, pp. 94-105, 1998.

[4] Bouguessa M and Wang S, "Mining Projected Clusters in High-Dimensional Spaces", IEEE Transactions on Knowledge and Data Engineering, Vol. 21, No. 4, pp. 507-522, 2009.

[5] Cai D, He X, Wu and Han J, "Non-negative Matrix Factorization on Mani-fold", In 8th IEEE Int. Conf. on Data Mining, IEEE Press, New York, pp. 63-72, 2008.

[6] Cheng, C. H. Fu, A. W. C. and Zhang, Y., "Entropy Based Subspace Clustering for Mining Numerical Data", Int. Conf. on Knowledge Discovery and Data Mining, ACM Press, pp. 84-93, 1999.

[7] Chu, Y. H. Huang, J. W. Chuang, K. T. Yang, D. N. and Chen, M. S., "Density Conscious Subspace Clustering for High Dimensional Data", IEEE Transaction on Knowledge and Data Engineering, Vol. 22, Issue 1, pp. 16-30, 2010.

[8] Gunnemann S, Farber I, Muller E and Seidl T, "ASCLU: Alternative Subspace Clustering", 16th Int. Conf. on Knowledge Discovery and Data Mining, ACM, 2010.

[9] Han J, and Kamber M, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, Second Edition, Chapter 8, pp. 335-393, 2001.

[10] Kaczmar, U. M. and Hurej, A., "Evaluation of Subspace Clustering Quality", Lecturer Notes in Artificial Intelligence 5271, pp. 400–407, Springer-Verlag Berlin Heidelberg, 2008.

[11] Kriegel HP, Kroger P and Zimek A, "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering", ACM Transactions on Knowledge Discovery from Data, Vol. 3, Issue 1, pp. 1-58, 2009.

[12] Kumar N and Puri C, "Projected Gustafson Kessel Clustering", 12th Int. Conf. on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, Springer, LNCS Vol. 5908, pp. 431-438, 2009.

[13] Moise G, Zimek A, Kroger P, Kriegel HP and Sander J, "Subspace and Projected Clustering: Experimental

Evaluation and Analysis", Knowledge and Information System, Springer, Vol. 21, Issue 3, pp. 299-326, 2009.

[14] Parsons L, Haque E and Liu H, "Subspace Clustering for High Dimensional Data: A Review", ACM SIGKDD Explorations Newsletter – Special Issue on Learning from Imbalanced Datasets, Vol. 6, Issue 1, pp. 90 – 105, 2004.

[15] Rajput DS, Singh PK and Bhattacharya M, "Feature Selection with Efficient Initialization of Clusters Center For High Dimensional Data Clustering", IEEE International Conference on Communication System and Network Technology, pp. 293 – 297, 2011.

[16] Wang D, Ding C, and Li T, "K-Subspace Clustering", PKDD European Conference on Machine Learning and Knowledge Discovering in Databases: Part II, Springer, Vol. 5786, pp. 506–521, 2009.