*International Journal of Engineering & Technology*, 7 (4) (2018) 4157-4163

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*
*doi: 10.14419/ijet. v7i4.20398*
***Research paper***

# A new generic interpretation of enhanced subspace clustering in high dimensional data

**J. Rama Devi [1] *, Dr. M.Venkateswara Rao [2]**

[1] *Research Scholar, Department of Computer Science and Engineering, GITAM Institute of Technology, Visakhapatnam, India*
[2] *Professor, Department of Information Technology, GITAM Institute of Technology, Visakhapatnam, India*
**Corresponding author E-mail:k.ramakarthik@gmail.com*

## Abstract

The prominent challenging task in data mining is to find the high dimensional data point clusters. In particular, the subspace clustering methods can be understood well in high dimensional data mining process. However the traditional subspace clustering techniques failed to find significance and quality of clusters that are present in the identified subspaces in growing the number of dimensions in large data. Most of the conventional clustering algorithms used bottom up search method and took multiple database scans to lead inefficiency. This research paper focuses a new enhanced subspace clustering scheme called ENSUBCLU, which overcomes the inefficiency from traditional subspace clustering techniques. Initially ENSUBCLU model was found in the dense units for each one dimensional projection of a given dataset. After that, applied subspace steering scheme to identify the promising subspaces and their combinations of common points among one dimensional subspaces. This model finds all interesting combinational dense core regions, from all lower dimensions of dense units. This lead to the reduction of subspace processing and obtain high quality subspace clusters and eliminates the redundant subspace clusters using hashing technique. Finally this model scales well with increasing attributes. ENSUBCLU model presents an empirical study on various synthetic, real world datasets and find the maximal subspace clusters in more improved manner than existing algorithms. It can even tackle many application areas like social networking, computer vision, bio-informatics, financial and sales analysis maintaining the high dimensional data.

***Keywords***: *Enhanced Subspace Clustering; Dense Core Region; High Dimensional Data Mining; Hashing; Subspace Clustering Algorithms.*

## 1. Introduction

Clustering is a challenging data mining task, which finds similar dense regions in high dimensional data. Conventional clustering techniques produce dense clusters in full original dimensional space[1-2]. But, increasing of dimensionality of data leads to some dimensions to become irrelevant for some clusters. These traditional clustering algorithms are inefficient for the growth of large data. This curse of dimensionality directs some efficient scalable problems in high dimensional data.[3-4]
Feature selection and feature creation are handles the curse of dimensionality. Principal Component Analysis (PCA) [5] Singular Value Decomposition (SVD)[6] are dimensionality reduction methods. These techniques are not support the high dimensional data and not consider some attributes perfectly in the transformed reduction dimension space, that leads to loss of data. These existing clustering methods generate irrelevant clusters in subsets of original space. Subspace clustering tackles this problem and process the clusters in all the identified subspaces of high dimensional data. Detecting clusters in the subspaces that permits quality clustering of the data which identifies the points than the full dimensional space called subspace clustering.
The definition of a subspace cluster is subspace of the feature space[8]. That is $C = (O_{DB}, S_D)$ where $O_{DB} \subset DB$, $S_D \subset D$ in this, C is defined as a subspace cluster, $O_{DB}$ is a set of objects in specified database DB and S is a subspace set of the dimension space D. This research paper also describes various types of subspace clustering algorithms which handles the high dimensional data. And describes existing enhanced subspace clustering algorithms[7, 8] which improves the significance of the clusters in the given datasets. Specifically, for the density based clustering paradigm like Dbscan [9] and extended to density parameters based on subspace clustering algorithms [10] .The neighborhood dense points around each object computed in each density based subspace approach. Various subsets of the entire dimension space clusters are embedded in high dimensional data [11]. Growing of dimensions is not easy to handle. $2^n$-1 subspaces are required for N –dimensional data. Most state-of-the-art subspace clustering algorithms were approached. Apriori bottom up search methods are expensive. It takes more number of database scans, generation of redundant subspaces and lack of scalability.
This research paper propose a New Enhanced subspace clustering model that detects pure maximal subspace clusters satisfying with user deterministic density parameters epsilon ε, minpoints μ and discretization point $d_\alpha$. It also reduces the number of database scans using new subspace steering scheme to identify promising subspaces. By allocating hash mapping collision procedure to eliminate the redundant subspaces
This paper is coordinated into a few portions. Section 1 is introduction and Section 2 will present current literature on conventional subspace clustering approaches. Section 3 will discuss the proposed work methodology in new generic enhanced subspace clustering algorithm. The last section 4 will discuss the proposed work experimental results comparing with conventional algorithms and followed final concluding remarks and discuss further directions.

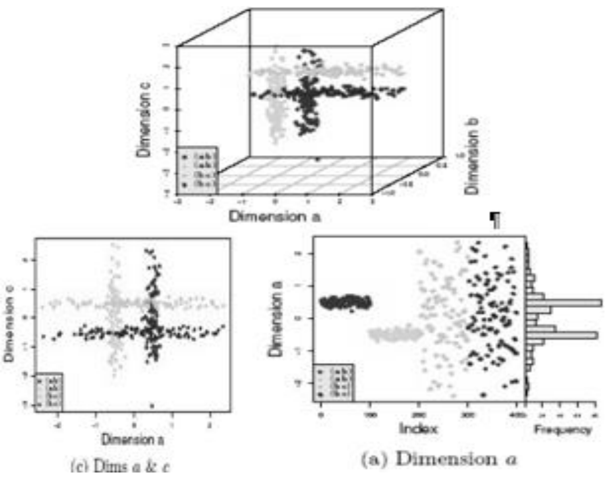


**Fig. 1:**Subspace Clustering In Some of the Subspaces [10].

## 2. Related background work

An era of subspace clustering categorized in different approaches. Bottom up subspace clustering approaches, top down subspace clustering approaches, specific soft subspace clustering techniques and enhanced subspace clustering approaches.

### 2.1. Bottom up- subspace clustering

Grid based, Density based are important subspace clustering algorithms. One important algorithm to grid based subspace clustering is CLIQUE(Clustering In Quest)[12], using an apriori principal for grid based subspace assessment in a bottom up way. In this epsilon ε, minpoints μ are the input parameters. Especially, the epsilon ε, grid based techniques based on the grid positions. It is observed that some clusters were missed due to unsatisfactorily shaped or oriented regions. Grid based subspace clustering algorithms. ENCLUS (Entropy based subspace Clustering) [13] and MAFIA(Merging of Adaptive Finite Intervals)[14] also miss the pure clusters and not get the useful information in subspaces of original space. SCHISM is another traditional enhanced grid clustering model for mining true clusters. Sequeirira and Zaki et al. suggested SCHISM algorithm [15] because it followed the depth first approach to get the maximal clusters. The convention of monotonicity was recycled to eliminate the search space, but this algorithm does not reach the scalability with respect to high dimensional data. Another traditional density subspace clustering methods are SUBCLU(Density Connected Subspace Clustering)[9], FIRES(Filter Refinement Subspace Clustering)[16], DENCOS(Density Conscious Subspace Clustering)[17]and INSCY(Indexing Subspace Clusters with In-Process Removal of Redundancy)[18] are density subspace approaches. Density based subspace clustering SUBCLU is follows APR based method which is represented in Figure 2. The defined input parameters are ε epsilon and μminimum points. It has generated number of trivial clusters and maintains at high run time. Kriegt et al. suggested FIRES (Filter Refine subspace clustering) which is a Fast algorithm to find relevant subspace clusters directly in one – dimensional clusters. This approach is also used a Apriori bottom up search method to find approximation clusters, but it does not associate Level by level subspaces. FIRES unable to locate all the private subspaces clusters and evaluates wrongly and gives resemblance of subspace clusters. One important index data structure is indexing subspace clusters is called INSCY. In this a novel index structures were used for efficient clusters and remove redundancy clusters but obtained multiple database scans.

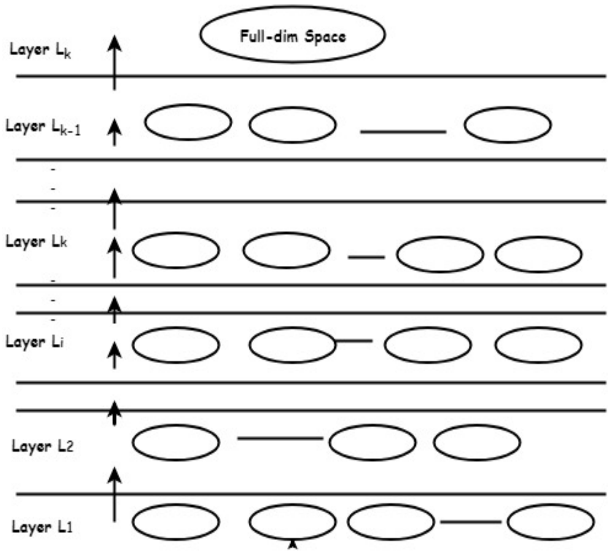


**Fig. 2:**AprioriApproach Subspace Lattice [9].

### 2.2. Top-down approach subspace clustering approach

The generation of subspace clusters is maintained towards lower-level dimensional spaces. PROCLUS(Projected Clustering)[19], ORCULUS(Arbitrarily Oriented projected Clustering)[20] and FINDIT(Fast and Intelligent Subspace Clustering )[21] techniques. The selection of these algorithms are suitable for dataset disjoint partitions, but not suitable for the applications of overlapping clusters.

### 2.3. Soft subspace clustering approach

FWKM (Fuzzy weighting K-means)[22], FSC (Fuzzy subspace clustering)[23], AWA (Attribute weighting algorithm)[24], A weight is selected to every dimension level to the formation of a specific subspace cluster and various clusters of subspaces can be determined by the value of weights. However these algorithms fail to assess the significance of cluster improvements in higher dimensions.

### 2.4. Enhanced subspace clustering approaches

K.Sim et.al [7] reviewed desired properties of enhanced subspace clustering algorithms. These are categorized into two important features such as and improving cluster results and handling complex data.

#### 2.4.1. Improving clustering results

The following algorithms involved in improving clustering results. One of them is Significance subspace clustering [25], which is Entropy based subspaces overlapping subspaces clusters and orthogonal Model subspace clusters [26]. And Constraint based subspace clustering [27] is to find the interesting subspaces. The model Overcoming Parameter sensitive subspace clustering is used to tune the efficiency of the process clusters are Rough set based subspace clustering approach ISC(Interesting Subspace Clusters)[28] and Efficient Density subspace clustering called EDENCOS(Efficient Density Conscious Subspace Clustering)[29] which is used top-down and bottom up approaches. To overcome the problems in conventional subspace algorithms, depth first search method is used in EDENCOS. It is an effective, incremental updating, FP mining method which keeps FPs efficiently .This technique follows FP-Tree with children table and trailer table in order to avoid the repetitions in the scanning, reconstructing and computing. But it takes more run time for increasing of attributes in the high dimensional data sets. Some design models developed in Enhanced subspace clustering model [30] and it supports to interpret the subspace clustering framework.

### 2.4.2. Handling complex data

Three dimensional data 3D, Categorical data, Noisy data such as sensor data in smart phone, financial data in finance, gene expression data in biology are handled by these subspace clustering algorithms.
All these enhanced subspace clustering algorithms are used to improve the potential results of high dimensional data. Research in the field of enhanced subspace clustering is an interesting genre to mine dimensional data sets effectively.
In summary, existing techniques are inefficient as it generates many redundant lower dimensional subspaces and high runtime in formation of clustering process. Apriori approaches did not generate the promising clusters and expected scalability. They have to extracted many non-candidate subspaces before reaching the promising subspaces. In particular some existing subspace clustering algorithms fails due to the inherent sparsity of the objects and unable to produce accurate clusters.
Next section, formulate a new Enhanced Subspace Clustering Model called "ENSUBCLU" for high dimensional data to fulfill the relevant subspace generations.

## 3. Proposed work

A New enhanced subspace clustering algorithm ENSUBCLU proposed to avoid sequential processing and also to overcome the drawbacks in existing subspace clustering algorithms. Some preliminaries are defined and overview of ENSUBCLU processing in the following sections.

### 3.1. Preliminary definitions

A unit of similar data in different subspaces of a given multi-dimensional data sets called subspace clustering. |DB| is considered as a set of data objects specified. P is considered as the specific unit of dimensions and forms a feature space called subspace as S ⊆ Pdefined. $2^{|P|}$-1 be the specified number of subspaces in a particular subset of dimensions P.

**Definition 1:** *The dense neighbors*

The ε-neighborhood of q in S, defined by $X_{\varepsilon}(\ q\ )$ and denoted as the following equation. (1).

$$\boldsymbol{X_{\varepsilon}}(q) = r \in |DB| \ / \ (dist\ (\boldsymbol{\pi_s}(q), \boldsymbol{\pi_s}(r) \leq \boldsymbol{\varepsilon} \quad (1)$$

Therefore,(dist ($\boldsymbol{\pi_s}$(q), $\boldsymbol{\pi_s}$(r) represents the distance between q and r in subset projection, 'S'. The data objects associated with $\boldsymbol{\varepsilon}$-neighborhood of 'q' are said to be $\boldsymbol{\varepsilon}$ neighbors of 'q'[9].
The minpoints$\mu$ is a positive integer and $c \in |DB|$. Data point $c$ is defined to core points and it has at least min points of $\varepsilon$- neighbors [9]
The data objects c, $d \in |DB|$.'c' is considered as direct density reachable from 'd' with respect to $\varepsilon$ and minpoints, that considered 'c' is $\varepsilon$- neighbor of 'd' [9],[28].

**Definition 2:** *Assigning Random integer sigcodes in number theory.*

Erdos et.al [31] suggested this Assigning unique- sigcode procedure is in number theory. If $P \geq 1$ is a positive integer then,$\{d_1 d_2, d_3 \dots d_{\partial}\}$ is denoted by specific partition, such that P=$\sum_{i=1}^{\partial} d_i$ for some $\partial \geq 1$ and summand points is defined $d_i >$ 0.The total number of partitions defined as $a_{\partial}(\rho)$and each partition has at most $\partial$ summand points. An asymptotic formula in these integer partitions defined in equation.(2).

$$\partial = o(P^{\frac{1}{3}}), \quad (2)$$

$$a_{\partial}(P) \sim \frac{\frac{P-1}{\partial-1}}{\partial!} \quad (2.1)$$

The number of unique fixed sized portions are ideally large. The probability of a specific region set of size $\partial$is. A very large integer P, if consider portion size$\partial$is

$$\frac{\frac{P-1}{\partial-1}}{\partial!} / \binom{P}{\partial} = \frac{(P-1)!\partial!(P-\partial)!}{(\partial-1)!(P-\partial)!\partial!P!} = \frac{1}{P(\partial-1)!} \quad (2.2)$$

Since P is the sum of labels of $\partial$
Therefore two dense region units U1 andU2 consists of same points with high probability, provided if summand($U_1$) =summand ($U_2$), thus sum is very large.

**Definition 3:***(discretization Point)$d_{\alpha}$.*

Discretization point is used to speed up subspace clustering process. Adjusting the discretization point with minimum points in the subspace to achieve the scalability. If discretization point $d_{\alpha}$ divide the range N into $d_{\alpha}$ for each hash computations maintained $\frac{N}{d_{\alpha}}$ entries for dense region units. Thus the value of $d_{\alpha}$ point to allocate hash tables holds the dataset points. Discretize the upper and lower limits of each hash-table computations.
Overview of ENSUBCLU processing
This Research paper designed a new ENSUBCLU frame work model. The following Figure 4 represents the proposed methodology view. Initially, ENSUBCLU construct the one- dimensional subspace dense units of a given dataset with epsilon$\varepsilon$,minpoints$\mu$ and discretization point$d_{\propto}$.
This research paper adopted signatures concept in the number theory to assign the unique random integer codes to each data points in the given input dataset. After that, we propose steering scheme to identify the promising subspaces. and their combinations of common points among 1-dimensional subspaces. This leads to the reduction of subspace processing and obtain high quality subspace clusters. Apply the steer jump approach to find all combinational dense core regions from all lower dimensions of dense units. The promising generic steering scheme is scalable to high quality relevant subspaces in high dimensional data. It finds the efficient information based on density paradigm. This steered approach overcomes the drawbacks of state-of-the-art processing methods and use these steered jumps directly to handle higher dimensional subspaces and prune many lower dimensional subspaces of search space.

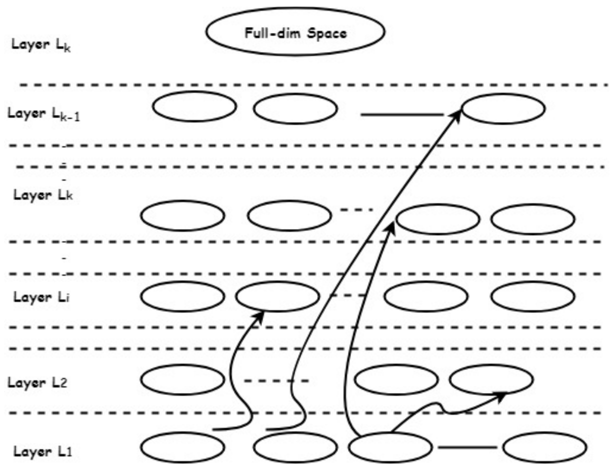

**Fig. 3:**Steer Subspace Exploration of Ensubclu.

The following figure 3 represents the steered approach to identify the common dense core points in lower level subspaces then to directly jump to higher dimensional subspaces. The promising common dense points of these all lower level attributes will direct to the dense core units in the next higher level attributes of sub spaces.

Steer the dense-core points in lower dimensional levels to make up the relevant subspaces in 2-dimensional, 3-dimensional, n-dimensional subspaces.

These high quality steered jumps with the steered indicators are used for removing redundant computations of combinatorial dense core regions.

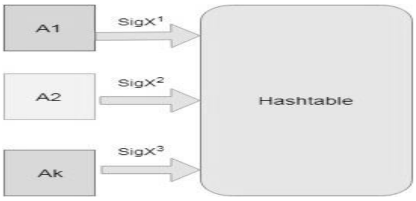


**Fig. 5:**H-Table Mapping with sigcode Units of Each Attributes.

Thus, it also handles an efficient exponential subspace search processing. Another important feature hash data structure is used in hash computations to maintain the improvement in the performance of ENSUBCLU algorithm for high- dimensional data sets. It also identifies the optimal sigcode units of dense region units sums. Each dimension of dense core units sums mapped with indexed dimensions of h-mapping Find collisions of high probability of same sigcode units of dense units and it merges the summands of dense core units. Thus obtain the maximal subspaces. Some times while dealing with high dimensional data it faces some memory bottleneck problems and it must be able in a position to tackle a more number of collisions of dense core units. Thus hash data structure should be able to derive all sigcodes of given datasets range N. The discretization point d$\alpha$ is given along with epsilon and minpts. The $d\alpha$ point granulize the P dimensions in the given datasets and various hash computations provided in each discretization slice. Thus ENSUBCLU technique has more benefits which are useful to discrete d$\alpha$ size to granularity of P dimensions in the given data. Discrete this range N into number of chunks and each chunk range can be processed in each hash table computations independently. And adjusting the discretization point d$\alpha$ with minpoints in the dense setsto scalable the higher subspaces. The discretization Point proportional to (maintain the) hash computations. Finally it merges the dense code units in order to eliminate the collisions of dense objects in each discretization ofhash computations.

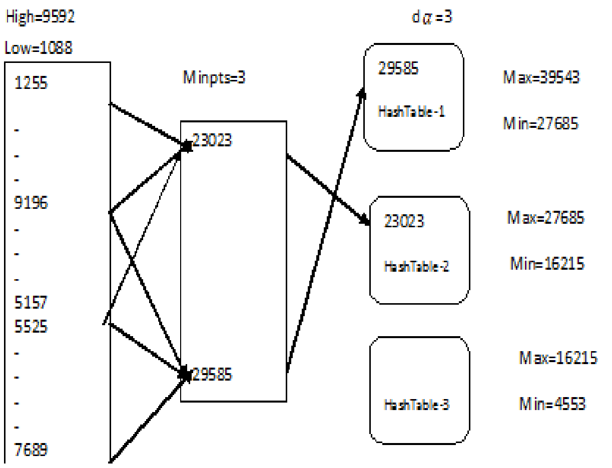


**Fig. 6:** Discretizing Hash-Table Computations.

Finally our model eliminates the redundant subspace clusters and it scales well with increasing attributes. It also refines the clusters from each found maximal subspaces and detect the accuracy of clusters using cluster metrics [32]. Our ENSUBCLU generic model experiments the results on various real world datasets [33] and synthetic datasets. It outperforms than traditional subspace clustering algorithms.

Next section describes the sequence flow of ENSUBCLU and proposed methodology Algorithms. Prototype of ENSUBCLU and analysis the Experimental Results.

### 3.2. Proposed methodology algorithms

The proposed ENSUBCLU method finds the maximal subspace clusters in high dimensional data points. ENSUBCLU formulates the four Modules.

Module-1: Finds the dense core units in all one dimensional projections.

Module-2: Identifying the promising subspaces by using Steered scheme.

Module-3: Reduces the redundant subspaces using Hash data structure.(Merge dense neighbors)

Module-4: Refines the maximal subspaces and find clusters in each found subspaces.

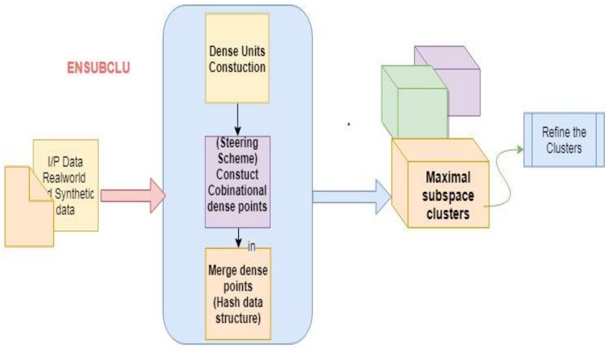


**Fig .4:** Overview of ENSUBCLU Framework.

The ENSUBCLU algorithm steps takes as follows.

**Algorithm-1**:Finds the dense core units in all one dimensional projections.

Input: |DB|, epsilon ε, minpoints$\boldsymbol{\mu}$ and discretization point $d_{\alpha}$.

Output: Found dense units in 1-dimensional subspace clusters.

Step 1: Assigning the Random integer codes to the each data point in the given data set.

Step 2: Initially, density connected subspace dense units are discovered in one-dimensional subspace of a given data with satisfying the ε value and minpoints$\boldsymbol{\mu}$.

Step 3: Based on discretization factor$\boldsymbol{d_{\alpha}}$ to scalable the high dimensional data.

Step 4: Allocate the h-tables based on the value of discretization factor $d_{\alpha}$ and max-min limits of random code integers are assigned to each h-table. Assign each dense units summands to hash table.

Step 5: Discretize the each h-table mapped with their summands of data points and identified the non- redundant dense regions due to the collisions.

Step 6: Finally obtain the 1-dimensional maximal dense code units in each attribute of dataset.

**Algorithm2**: Identifying the promising subspaces by using Subspace steered scheme and reduce the redundant subspaces:

Step 1: It presents how the core sets form lower dimensional subspaces to higher dimensional subspaces.

Step-2: Identifying the common subset dense points across these one-dimensional levels and direct to the dense-core units in next higher dimensional subspaces.

Step-3: Subspace steering technique for relevant mining of subspace clusters and steer indicators are used to reduce the combinational core set cluster computations.

Step 4: The overall efficient cluster results depending on the steer indicators selection.

Step 5: Hash their summands of random code integers then find the collision dense point units in all combinational sets of maximal subspaces.

//Refine the subspaces and find clusters in each found subspaces.

Step 6: Refine the merge dense points to apply any density parameters on found subspaces.

Step 7: Compute the accuracy clusters in the given sub sets.

Next section, interpret the results of an empirical study. The results that satisfy and fulfill the promising requirements.

## 4. Experimental results

The proposed algorithm ENSUBCLU organized to formulate the relevant subspace clusters on various real-world data sets. Compare ENSUBCLU model with the following EDENCOS [29], INSCY [18], and SUBCLU [9] subspace clustering methods which are represented in table 3. ENSUBCLU model implemented in Java language on 64 bit Windows 7 OS with Intel core i5-2600 and 12 GB RAM.
Use synthetic datasets and real world data sets from UCI repository system [33] to the assessment of quality maximal subspaces. Synthetic and real world datasets are maintained in table 1 and table 2. The performance of ENSUBCLU in terms of efficiency and accuracy on series of experiments. Compare the efficiency scalability and the accuracy of ENSUBCLU against existing well known density based subspace clustering algorithms are EDENCOS, INSCY, SUBCLU.

**Table 1:**Synthetic Datasets

| Synthetic Datasets | Dimensions | Size |
|---|---|---|
| D20 | 20 | 5120 |
| D40 | 40 | 5120 |
| D80 | 80 | 5120 |
| D160 | 160 | 5120 |
| D200 | 200 | 2000 |

### 4.1 Result analysis

The efficiency of all mentioned algorithms on a series of real world data sets and synthetic data sets are compared. The time scalability with respect to the growing of data dimensionality on synthetic datasets represented in the following Figure 7.

**Table.2:**Real World Datasets

| Real world datasets | Dimensions | Size |
|---|---|---|
| Diabetes | 8 | 768 |
| Vowel | 9 | 991 |
| Wages | 11 | 534 |
| Varying Size | 29 | 5000 |
| Face50 | 50 | 2250 |
| Swedishleaf | 50 | 1125 |

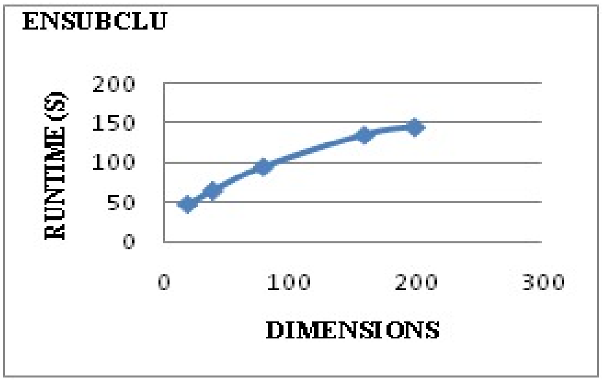


**Fig. 7:** Dimensions vs Runtime.

Figure.8 represents the time scalability with respect to the growing of real world data set size.
Existing algorithms are not scalable for increasing dimensionality, which also consumes extra quite time and compare with ENSUBCLU. Figure 9 represents run time of various traditional subspace clustering algorithms with the proposed model ENSUBCLU. Muller et.al [32] developed open subspace framework and it is used to evaluate various conventional subspace clustering algorithms and compute the F1 score, which is quality measure of obtained subspace results. In this, F1 measure which is computed as the harmonic mean of precision and recall values and defines better quality that is high F1 score.

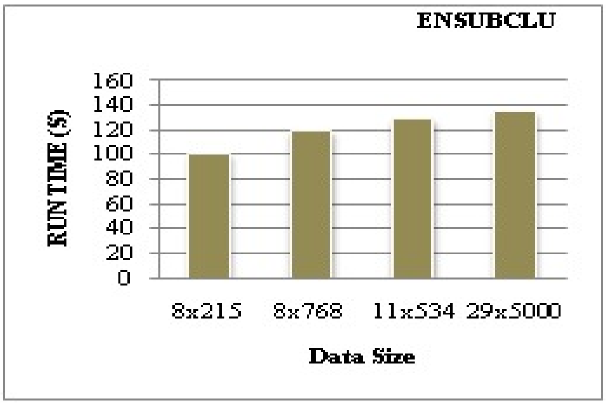


**Fig. 8:**Data Size vs Run Time.

The following figure.9 refers the runtime comparisons in various subspace clustering models with proposed ENSUBCLU model better results.

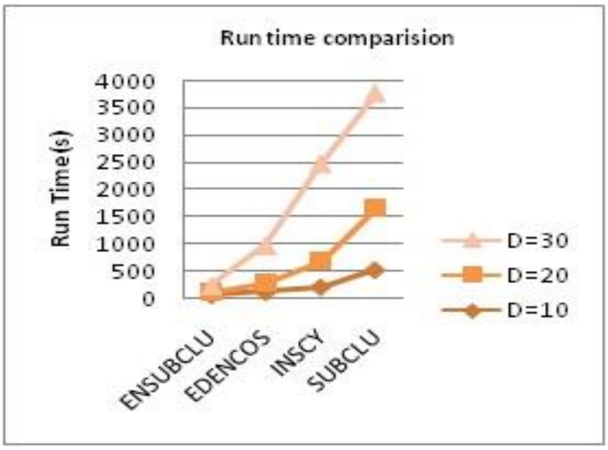


**Fig. 9:** Comparison of Subspace Clustering Techniques Runtime.

Figure10 represents the F1 score of accuracy clusters quality with various epsilon ε deterministic values. Note that, a larger epsilon ε value is not to get the better quality clusters. Scalability: The choice of epsilon ε and dataset size derives the number of dense units. The larger produces more number of dense core sets.
So, the ENSUBCLU algorithm tackles the more number of collisions of dense core units by using Hash data structure. Deploy the Hash-tables computations to handle the memory bottleneck problems of high-dimensional data.

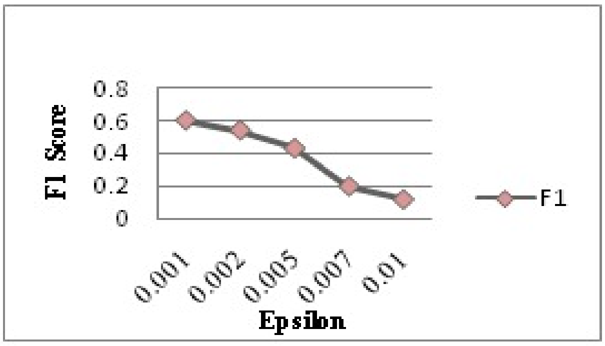


**Fig.10:** Epsilon ε vs F1 Score.

Use the discretization point $d_\alpha$ to granulize the more number of dimensions in given datasets. Discretize the dataset size N into $d_\alpha$ partitions and each hash table maintains N/$d_\alpha$entries in order to reduce processing of subspace computations. Always the runtime is proportional to $d_\alpha$point. Figure.11 represents the time scalability of ENSUBCLU algorithm with respect to the various discretization points on various synthetic datasets referred from table 1.

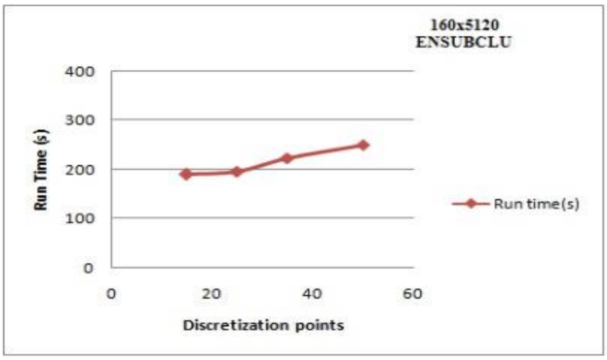


**Fig.11:** Discretization point $d_{\alpha}$vs Time Scalability.

The below mentioned table [3] refers to the accuracy and F1 score of the various subspace algorithms. ENSUBCLU, the generic enhanced subspace-clustering algorithm maintains high F1 score than some existing subspace clustering algorithms EDENCOS, INSCY, SUBCLU and thus, better accuracy.

**Table 3:**F1 Score and Accuracy of Subspace Clustering Algorithms with ENSUBCLU

| Datasets(mxn) | ENSUBCLU | | EDENCOS | | INSCY | | SUBCLU | |
|---|---|---|---|---|---|---|---|---|
| | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy | F1 | Accuracy |
| Diabetes(8x768) | 0.81 | 0.91 | 0.56 | 0.67 | 0.36 | 0.65 | 0.25 | 0.56 |
| Vowel(9x991) | 0.45 | 0.63 | 0.37 | 0.29 | 0.07 | 0.16 | 0.17 | 0.22 |
| Wages(11x534) | 0.58 | 0.6 | 0.45 | 0.57 | 0.34 | 0.26 | 0.14 | 0.16 |
| Swedish leaf(50x1125) | 0.35 | 0.41 | 0.23 | 0.34 | 0.19 | 0.29 | 0.09 | 0.2 |
| Face(50x2250) | 0.6 | 0.71 | 0.46 | 0.54 | 0.17 | 0.24 | 0.1 | 0.14 |
| Varying size (29x5000) | 0.53 | 0.64 | 0.36 | 0.41 | 0.11 | 0.34 | 0.01 | 0.34 |

## 5. Conclusion

This research paper proposed a new generic enhanced subspace clustering model which is denoted named as ENSUBCLU. This algorithm based on density clustering paradigm and subspace steering approach is used to find the promising subspaces and different discretization points will be adjusted to detect the clusters in relevant subspace cardinalities. This model reduces the number of database scans and also to eliminate unnecessary subspaces used in hash mapping collision procedure. Next merged the dense point units and refined the clusters. These experimental results shown, ENSUBCLU can enhance the quality in subspace cluster results. ENSUBCLU proved to be efficient and it definitely replaces all other conventional subspace clustering algorithms with its features. This paper concludes that this model measures the obtained maximal subspace clusters and its accuracy by using cluster metrics. Further development of this work is to implement internal subspace cluster measures on unsupervised data clusters to validate the effectiveness of this methodology.

## References


[1] Zhang T, Ramakrishnan R, LivnyM, "BIRCH: an efficient data clustering method for very large databases",International conference on management of data, vol. 1.ACM Press, USA,1996,103–114 https://doi.org/10.1145/233269.233324.

[2] Ester M, Kriegel H, Sander J, Xu X, A density-based algorithm for discovering clusters in large spatial databases with noise, International Conference Knowledge Discovery Data Mining 96(34):1996, 226–231.

[3] Bellman RE, "Adaptive control processes: a guided tour", Princeton University Press, New Jersey, 1961.https://doi.org/10.1515/9781400874668.

[4] Beyer K, Goldstein J,"When is nearest neighbor meaningful? Proc 7th IntConf Database Theory. In: Database Theory–ICDT'99. Lecture Notes in Computer Science. Springer, Berlin Heidelberg Vol. 15, no. 40, 1999, pp 217-235.

[5] Joliffe IT,"Principle component analysis", 2nd edn. Springer, New York, 2002.

[6] Tan, Pa-Ning, Steinbach, Michael, Kumar, Vipi, Introduction to Data Mining. Addison-Wesley Longman Publishing Co., Inc, Boston, MA, 2005.

[7] Kelvin, Sim, Gopalkrishnan, Vivekanand, Zimek, Arthur, Cong, GAO, A survey onenhanced subspace clustering, Data Mining Knowl. Discovery,Vol. 26, no. 2, 2013,332–397.https://doi.org/10.1007/s10618-012-0258-x.

[8] Rama Devi.J, Venkateswara Rao M, An era of enhanced subspace clustering in high-dimensional data, i-manager's Journal on Computer Science, Vol. 4, no. 3,September –November 2016,29-36.

[9] Kailing K, Kriegel HP, Kroger P," Density-connected subspace clustering for high-dimensional data",In: SIAM international conference on data mining.2004,246–256,.

[10] Parsons L, Haque E, Liu H "Subspace clustering for high dimensional data: a review", ACM SIGKDD ExplorNewsl Vol.6, no.1, 2004, 90–105.https://doi.org/10.1145/1007730.1007731.

[11] Agrawal R, Gehrke J, Gunopulos D, Raghavan P, Automatic subspace clustering of high-dimensional data for data mining applications. In: Proceedings of the ACM international conference on management of data *(SIGMOD)*, 1998, 94–105.https://doi.org/10.1145/276304.276314.

[12] Jyoti Yadav, Dharmender Kumar Subspace Clustering using CLIQUE: An Exploratory Study, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Vol. 3, no. 2, February 2014.

[13] Cheng, C.H., Fu, A.C., Zhang, Y.:"Entropy-Based Subspace Clustering for Mining Numerical Data". In: Proc. ACM SIGKDD Int. Conf. on Knowledge Discovery in Databases, San Diego, FL. (1999)https://doi.org/10.1145/312129.312199.

[14] Goil, S., Nagesh, H., Choudhary, A.:"MAFIA: Efficiant and Scalable Subspace Clustering for Very Large Data Sets". Tech. Report No. CPDC-TR-9906-010, Center for Parallel and Distributed Computing, Dept. of Electrical and Computer Engineering, Northwestern University 1999.

[15] Sequeira, K., Zaki, M., SCHISM: A New Approach for Interesting Subspace Mining. In: The Proceedings of the Fourth IEEE Conference on Data Mining, .2004, 186–193.Kriegel H, Kroger P, Renz M, and Wurst S,

[16] A generic framework for efficient subspace clustering of high–dimensional data, In Proceedings of fifth IEEE International Conference on Data Mining,vol. 46, no.4, Mar 2004,255-271,.

[17] Yi-Hong Chu, Jen-Wei Huang, Kun-Ta Chuang, De-Nian Yang, "Density Conscious Subspace Clustering for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering, vol. 22, no.1, January 2010. 16-30, https://doi.org/10.1109/TKDE.2008.224.

[18] Assent I, Emmanuel M, Seidl T, "Inscy: Indexing subspace clusters with in-process-removal of redundancy", in: Eighth IEEE international conference on data mining. IEEE.2008, 719–724.https://doi.org/10.1109/ICDM.2008.46.

[19] Aggarwal, C.C., Wolf, J.L., Yu, P.S., Cecilia, Procopiuc, Jong Soo, Park, 1999. Fast algorithms for projected clustering. In: Proc. ACM SIGMOD international conference on Management of data, New York, U.S.A., pp. 61–72.https://doi.org/10.1145/304182.304188.

[20] Aggarwal, C.C., Yu. P.S, Finding generalized projected clusters in high dimensional spaces. In: Proc. 1st International Conference on Management of Data, New york, U.S.A, 2000, 70-81

[21] Woo Kyou-Gu, Lee, Jeo-Hoon, Kim, Myou-Ho, Lee, Yo-Joon, and FINDIT: a fast and intelligent subspace-clustering algorithm. Inf. Softw. Technol. Vol. 46, no. 4, 2004, 255–271.

[22] Jing, L., Ng, M.K., Xu, J., Joshua Zhexue, Huang. Subspace clustering of text documents with feature weighting k-means algorithm. In: Proc. 9th Pacific-Asia conference on Advances in Knowledge Discovery and Data Mining, Springer, Berlin Heidelberg, 2005, 802–812.https://doi.org/10.1007/11430919_94.

[23] Gan, G., Wu, J, A convergence theorem for the fuzzy subspace clustering (FSC) algorithm,PatternRecogn. Vol. 41, no. 6, 2008, 1939–1947.https://doi.org/10.1016/j.patcog.2007.11.011.

[24] Chan, E.Y., Ching, W.K., Ng, M.K., Huang, Joshua Z, An optimization algorithm for clustering using weighted dissimilarity measures. Pattern recognition Vol. 37, no.5, 2004, 943–952.https://doi.org/10.1016/j.patcog.2003.11.003.

[25] Dash M, Choi K, Scheuermann P, and Liu H,."Feature selection for clustering-a filter solution. In: Proceedings of the two IEEE International Conference on Data Mining (ICDM), 2002, 115-122.

[26] Günnemann S, Müller E, Färber I, and Seidl T "Detection of orthogonal concepts in subspaces of high dimensional data". In: Proceedings of the 18 ACM Conference on Information and Knowledge Management (CIKM), 2009,1317-1326,https://doi.org/10.1145/1645953.1646120.

[27] Fromont É, Prado A, and Robardet C, "Constraint-based subspace clustering". In the Proceedings th of the 9 SIAM International Conference on Data Mining (SDM), 2009, 26-37.https://doi.org/10.1137/1.9781611972795.3.

[28] Lakshmi, B.J., et al. A rough set based subspace-clustering technique for high dimensional data. Journal of King Saud University – Computer and Information Sciences 2017. https://doi.org/10.1016/j.jksuci.2017.09.003.

[29] Suguna. M &Palaniammal. S, An Efficient Density Conscious Subspace Clustering Method using Top-down and Bottom-up Strategies‟, International Journal of Computer Science and Information Technologies, vol. 5, no. 3, 2014,3839-3842.

[30] Rama Devi.J, Venkateswara Rao M, Design a New Model for Enhanced Subspace Clustering of High –Dimensional Data, International Journal of Control Theory and Applications ISSN: 0974–5572, Vol. 9, no. 42, 2016, 201-208.

[31] Erdös P, LehnerJ ," The distribution of the number of summands in the partitions of a positive integer", Duke Mathematical Journal,Vol. 8,no.2,1941,335–345. https://doi.org/10.1215/S0012-7094-41-00826-8.

[32] Muller E, Gunnemann S, Assent I, Seidl T and Farber I, 2009.

[33] Lichman,UCI Machine Learning Repository, available at: 2007.http://archive.ics.uci.edu/m/.