

HYPER-GRAPH BASED VISUALIZED CLUSTERING APPROACH (HVCA) FOR EFFECTIVE DATA PARTITIONING

B.V RamNaresh Yadav¹, P Anjaiah²

¹Department of CSE, JNTUH, Hyderabad

²Department of CSE, Institute of Aeronautical Engineering, Hyderabad

¹bvramnaresh@gmail.com, ²anjaiah.pole@gmail.com

Abstract: Clustering is primarily used to detect the similar objects for determining of groups or clusters. Top clustering methods, such as k-means and hierarchical clustering are unable to find the initial number of clusters, i.e., it cannot access the value of clustering tendency. Many automatic cluster number detection methods are investigated for this purpose. It is studied that visual access tendency (VAT) is one of approach for effective cluster detection method. It takes the input as dissimilarity features for a set of objects and reorders the dissimilarity matrix according to order of most similar objects. It shows the number of clusters as square shaped dark blocks visually in the image of reordered dissimilarity matrix. Visualized clustering approach (VCA) captures the clusters labels of objects during clustering process. However, context-aware based dissimilarity information gives more informative cluster assessment in VCA. Thus, hyper graph is constructed to create context based aware for a set of objects in the proposed hyper-graph based visualized clustering approach (HVCA). Two performance measures are used for demonstrating the efficiency of proposed HVCA in the experimental study.

Keywords: Clustering, VAT, Clustering Tendency, HVCA, Clustering Accuracy, Normalized Mutual Information.

1. Introduction

The bunching is utilized as a part of an extensive variety of genuine applications, incorporates information grouping [8], picture division [17], design acknowledgment, video movement division and so on. The fundamental idea of grouping is unsupervised since it plays out the bunching undertaking of unlabeled information. Most grouping calculations are explored, for example, k-implies [8], progressive bunching idea [9], thickness based grouping idea [20], and chart based bunching [10]. The k-means is a classical algorithm and it creates the initial number of clusters with initially assigned objects. Initial number of clusters may influence the clustering results. The number of clusters

for k-means [8] is given by the user itself. However, sometimes the user may attempt the unsuitable number of clusters in k-means. We have observed that same issue is arisen in other clustering algorithms [7]. After the rigorous study, we address the best solution of clustering tendency (or number of clusters) from [1, 2, 3, 15, 16], they proposed a visual access tendency (VAT) [2] (with different versions such as SpecVAT, and iVAT). The VAT is an automatic cluster detection tool, which can determine the number of clusters in a visual view, hence, is known as the visualized procedure. The major contribution of the VAT is finding the dissimilarity features between data objects. These features are derived using metrics such as Euclidean, and Cosine. The exact discrimination between objects is helped for VAT to detect the number of clusters. The context-based dissimilarity features provide more assessment of discrimination between data objects. Thus, we use a context-aware [3] based dissimilarity features in our extended VAT for improving the access tendency of clusters. The present clustering algorithms use a clustering tendency in their algorithms for accessing of clustering results but it leads to very expensive steps since we merge the two procedures for achieving of clustering results, which are as follows: (1) VAT and (2) any clustering algorithm. VAT detects the number of clusters 'c' from unlabeled data, and clustering algorithms use a value of 'c' as input, and it generates the clustering results for 'c' clusters. Due to the concern of expensive steps of the hybrid procedure, we propose a straight clustering procedure, which is capable to extract the clustering results directly from context-aware-based-VAT [13]. The context-aware based dissimilarity between data objects is derived in [10], which use a k-nearest neighbor concept for a finding of hyperedges, and it finds the dissimilarity features of data objects in respect to hyperedges [11]. Our method extracts the efficient clustering results from context-aware based Visualized Clustering Procedure (CAVCA). The technical procedure of VCA is described in [13]. The summary of contributions of the paper is defined as follows:

1. The optimal automatic cluster detection method is introduced

2. The context-based dissimilarity matrix is derived.
3. The basic procedure of VAT is re-designed by context-aware based dissimilarity matrix; the respective algorithm is introduced in this paper
4. The newer version of VCA is presented for the purpose of extracting the efficient clustering results from CAVAT or (VAT)
5. The empirical analysis is performed for our results
6. The experimental evaluations are performed to proving the effectiveness of our method
7. The clustering validation index (CVI) [14, 15] measures are evaluated for the purpose of validating our clustering results.

Section 2 displays the writing investigation of the work, Section 3 talks about the imagined grouping approach (VCA), Section 4 introduces the proposed technique, Section 5 reports the trial investigation of the paper, and Section 6 exhibits the conclusion and future extent of the work.

2. Literature Study

Automatic detection of a number of clusters is recognized as a major problem in the clustering of unlabeled data. Many clustering techniques [7] require the prior number of clusters 'c' for achieving the efficient clustering results because the quality of clusters heavily depends on the correct estimation of 'c'. Seek the correct number of clusters for unlabeled data is usually called as clustering tendency. We studied some methods for the purpose of finding clustering tendency. The statistically based technique for performing the clustering tendency assessment [11] is discussed in Yingkang, H.; Hathaway [4]. Mean-Shift (MS) [19] succeeds in the area of speech clustering, and it finds the modes (or clusters) for unlabeled data in a non-parametric form; it is known as a mode-seeking procedure, the motivation of MS is happening by its non-parametric nature. The clustering results of MS depend on bandwidth parameter. The selection of the value of this parameter in MS is a critical task due to the reason of its dynamic nature, so it is one of the limitations of the mean-shift procedure. Bayesian information criteria(BIC) is the popular approach for data segmentation [17]. We can estimate the number of clusters by BIC but it may be very sensitive to large data. During the investigation of automatic cluster detection, we borrowed the vital concept from Bezdek et al. [5,12], they developed VAT, which uses a pairwise dissimilarity information in $n \times n$ image form for a set of n objects as the input of VAT. It shows the highlights of potential clusters information using reordered dissimilarity image (RDI) [2] as output. VAT is a clustering tendency tool, this is used for the purpose of exploratory data analysis. It aims to generate the potential clusters as groups along the diagonal in the result of VAT Image. Recently, several algorithms extend the VAT for related

clustering assessments problems [1, 2, 3], the coVAT[19] is designed for detecting the clustering tendency in the co-clustering problem. For very large data cases, we use a bigVAT[12] for solving the issue of scalability. Most of the clustering applications involve the complex data structures; it demands the more reliable approach. In the concern or assumptions of compact and well clustering properties, Lian Wang et al.[16] have presented the SpecVAT, it uses a spectral (or Eigen concept) in the VAT for reliable estimation of clusters via visual inspection of SpecVAT image. An improved VAT (iVAT) [14, 15] is recently proposed by C. Havens et al. [14], they suggest that iVAT has made a significant visual improvement in the clarity of resulting image compared to VAT image. Thus, it helps to easily detect a number of clusters in iVAT. These present techniques have detected the prior clusters of unlabeled data, it also required to find the clustering labels of every data object i.e we need to generate the "good" partitioning results. From these VAT procedures, we have the knowledge of clusters but we require clustering algorithm for generating the clustering results. This hybrid approach(both VAT and any clustering algorithm) is capable to produce the clustering results along with a known number of clusters. But it is the very expensive approach. For solving this issue, we propose a directly visualized clustering procedure of VAT (and SpecVAT) for generating the clustering results as well as clustering tendency. In our paper, we can also refine the procedure for finding the dissimilarity matrix by context-aware concept [3]. The context-aware based dissimilarity matrix increases the performance of visualized clustering procedure. It is explained in the proposed work. Context-aware based dissimilarity measure lead to the robust clustering results even the data have noisy. The contextual dissimilarity information of a set of objects is in more stable form, the main reason is that it takes the consideration of local grouping problem and nearest neighbor information of each vertex. The main motivation of this paper is to find the prior clustering tendency, and then design a robust data clustering algorithm for VAT (and SpecVAT). We attempt the two issues in this paper, i.e to how to detect the correct clustering tendency in an unsupervised way using context-aware based VAT (and SpecVAT), and how to extract the clustering results from a visualized clustering approaches, namely, CAVAT, and CASpecVAT. The different versions of the VAT are discussed in the following section.

3. Proposed Methodology

Our work is initially built upon data partitioning using VAT, so we use the visualized clustering approach (VCA) algorithm. We can give the first glance is to detect the prior clustering tendency, later we find the respective clustering results by VCA for the

hypergraph, hence it is known as HVCA. We follow the three- vital steps in VCA and these are described here: 1) The VAT image is derived from the input of dissimilarity matrix of a set of data objects. The flexible metrics are used to compute the dissimilarity features in a matrix form. 2) The derived VAT image always suggests that members of clusters. The crisp

partition matrix defines the data partition results from VAT image. 3) The current labels of data members are compared with ground truth labels in order to find the clustering accuracy. The HVCA algorithm is outlined as follows.

Algorithm: HVCA

Input: Dissimilarity Matrix for hypergraph

N: Number of objects

Output: k: Number of Clusters

Method:

Step 1:

[RD] = VAT (D)

VAT_Im=Image (RD);

k=No_of_Square_Shaped_Dark_Blocks (VAT_Im);

Step2:

For i = 1: k

Find data objects at each partition i using crisp- partition matrix

End for

Step 3:

Map the data objects and find the ground truth labels using Khun-munkres function for finding of clustering accuracy

Recently, the spectral concept is very familiar because it is used in a graph analysis for various applications such as dimensionality reduction [22], video segmentation [17], and data clustering [18]. The spectral graphs use the Eigenvectors for analyzing of graph adjacency, it is used to define a geometrical graph, hence the spectral based clustering results reveals the hidden cluster structure even the data is in

complex-shaped. The respective VAT algorithm is introduced in [16], called the SpecVAT. After the thorough study of SpecVAT, we recommend the same VCA to retrieving the better data partition results from the SpecVAT image instead of VAT image in step 2. The proposed changes are done in the spectral based HVCA, called the SpecVCA. The HSpecVCA algorithm is described as follows.

Algorithm: HSpecVCA

Input: Dissimilarity Matrix

N: Number of objects

Output: k: Number of Clusters

Method:

Step 1:

[RD] = SpecVAT (D)

SpecVAT_Im=Image (RD);

k=No_of_Square_Shaped_Dark_Blocks (SpecVAT_Im);

Step2:

For i = 1: k

Find data objects at each partition i using crisp- partition matrix

End for

Step 3:

Map the data objects and find the ground truth labels using Khun-munkres function for finding of clustering accuracy

We have emerged to find the best optimal clustering results from either VAT or SpecVAT image. The nature of bunching comes about relies upon either best VAT picture or SpecVAT picture as far as clearness of piece structures. Fig. 1 shows the clearness of piece structures for both VAT and SpecVAT pictures for manufactured information.

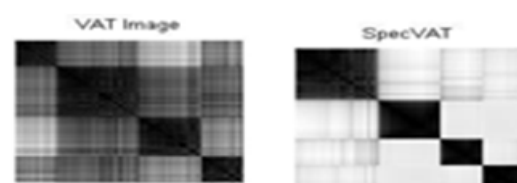


Figure 1. VAT image and SpecVAT image for synthetic data

The HSpecVAT finishes up the more lucidity for various groups than VAT picture utilizing piece structures of hyper-chart. In light of this perception of a ghostly idea, we emphatically wish to install the difference highlights of an arrangement of articles in a k-dimensional otherworldly insert space, where k alludes the number of Eigenvectors and each datum point is supplanted with another ghastry space. The various important steps are noted in the HSpecVAT as follows.

1. Compute the distance matrix (W) for a set of objects using the local statistics of k-nearest neighbors, which results maximize affinities within clusters, and minimize affinities across clusters.
2. We construct the normalized Laplacian matrix (L) as follows

$$L = M^{(-1/2)} (M - W) M^{(-1/2)}$$

3. Choose the k largest Eigenvectors of L to form the Eigen matrix $V = [v_1, v_2, \dots, v_k] \in \mathbb{R}^{n \times k}$
4. Normalized the V with Euclidean norm to define V^1
5. Find a new dissimilarity matrix(Dnew) of V^1 in respect to n objects (of k-Eigen dimensions)
6. Apply the VAT for Dnew to obtain the SpecVAT image

The cSpecVAT is another improvement of SpecVAT, in which the dissimilarity matrix is computed using cosine metric instead of Euclidean in step 5. The cSpecVAT is more robust than SpecVAT in most of the real datasets. The cSpecVAT is described in [16].

Therefore, we made an observation that finding the accurate dissimilarities is a crucial step in clustering. To enable the accurate dissimilarities from pairwise objects is addressed using the surrounding contexts of objects and this idea is borrowed originally from [21]. Required steps are and changes are proposed in the present systems for improving the quality of clustering results in our work.

4.Experimental Results

In this experimental study, proposed HVCA is evaluated using synthetic datasets, real data sets, and speech datasets. The clustering accuracy (CA) [5, 6], and normalized mutual information (NMI) [5, 6] are used in the experimental study for evaluating the proposed methods.

4.1. Datasets Description

The accompanying benchmarked datasets are utilized as a part of the test ponder. The datasets are as per the following: four engineered datasets (S-1 to S-4 in Fig. 3), and eight genuine datasets(R-1 to R-8), incorporates iris, voting, wine, and seeds, and discourse datasets (R-

5 to R-8) from [21]. The subtle elements of these datasets are displayed in Table 2.

Table 2: Description of the Datasets

Real Dataset	Name of the Dataset	No of Clusters (or Classes)
R-1	Iris	3 (No of classes)
R-2	Voting	2(No of classes)
R-3	Wine	3
R-4	Seeds	3
R-5	Speech Data from TSP	2 Speakers data
R-6	Speech Data from TSP	3 Speakers data
R-7	Speech Data from TSP	4 Speakers data
R-8	Speech Data from TSP	5 Speakers data

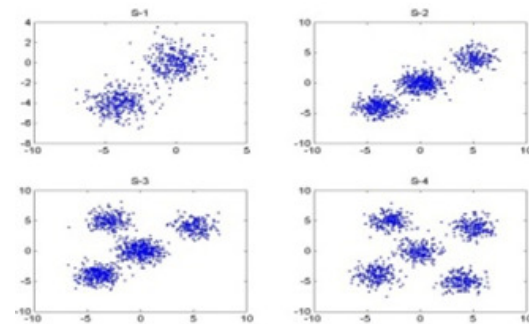


Figure 3. Synthetic Datasets (S-1 to S-4)

Existing methodologies VAT and SpecVAT determines the number of clusters from their VAT and SpecVAT Images; Our proposed methodologies i.e., HVCA and HSpecVCA access the clusters labels from the VAT and SpecVAT Images along with labels of objects by crisp partitioning concept. Cluster quality of HVCA and HSpecVCA is determined by two performance parameters, such as clustering accuracy and normalized mutual information in following tables Table 1 and Table 2

Table 1. Clustering Accuracy

Name of Dataset	HVCA	HSpecVCA
R-1	0.88	0.91
R-2	0.72	0.81
R-3	0.75	0.85
R-4	0.77	0.88
R-5	0.66	0.77
R-6	0.76	0.88
R-7	0.78	0.88
R-8	0.77	0.92
S-1	0.79	0.91
S-2	0.81	0.87
S-3	0.78	0.86
S-4	0.82	0.85

Table 2. Normalized Mutual Information

Name of Dataset	HVCA	HSpecVCA
R-1	0.81	0.89
R-2	0.77	0.88
R-3	0.72	0.83
R-4	0.73	0.84
R-5	0.64	0.75
R-6	0.75	0.86
R-7	0.76	0.84
R-8	0.73	0.95
S-1	0.75	0.89
S-2	0.78	0.89
S-3	0.88	0.91
S-4	0.79	0.92

From Tables 1 and 2, it is observed that CA and NMA values are high in HSpecVCA, thus it produces more quality of clusters than HVCA method.

5. Conclusion

Traditional methods are unable to access the prior knowledge about a number of clusters for comprehensive datasets. Therefore, this paper is focused on accessing of prior number of clusters along with the quality of clusters. Proposed methodologies, HVCA and HSpecVCA are addressed the problems of prior knowledge about a number of clusters from the VAT and SpecVAT Images and also discovers the quality of clusters by crisp partition matrix concepts. From the experimental results, it is observed that spectral based proposed methodology produces more quality of clusters.

References

- [1] Bezdek, J.C.; Hathaway, R.; VAT: A tool for visual assessment (cluster) tendency, Proc. IJCNN, Honolulu, 2002, 25-30
- [2] Bezdek, J.C.; Hathaway, R.J.; Huband, J.; Visual Assessment of Clustering Tendency for Rectangular Dissimilarity Matrices, IEEE Transaction Fuzzy Systems, 15(5), 2007, 890-903.
- [3] K. Rajendra Prasad, C. Raghavendra, K Sai Saranya, "A Review on Classification of Breast Cancer Detection using Combination of The Feature Extraction Models", International Journal of Pure and Applied Mathematics, Volume 116 No. 21 2017, 203-208.
- [4] C. Raghavendra, A. Kumaravel and S. Sivasuramanyan, "Features Subset Selection using Improved Teaching Learning based Optimisation (ITLBO) Algorithms for Iris Recognition", Indian Journal of Science and Technology, Vol 10(34), DOI: 10.17485/ijst/2017/v10i34/118307, September 2017.
- [5] Bezdek, J.C.; Pal, N.R.; Some New Indexes of Cluster Validity, IEEE Trans on Systems ,Man, and Cybernetics, 28(3), 1998,301-315
- [6] Dehak, N.; Dehak R.; Glass, J.; Reynolds, D.; Kenny, P.; Cosine similarity scoring without score normalization techniques, Proceedings IEEE Odyssey workshop, Brno, Czech, Republic, 2010
- [7] B. Eswara Reddy and K. Rajendra Prasad, "Improving the Performance of Visualized Clustering Method", International Journal System Assurance Engineering, Springer, Vol. 7, Supplement 1, pp. 102–111 Feb. 2015.
- [8] Jain, A.K.; Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31(8), 2010, 651-666.
- [9] K. Rajendra Prasad, C. Raghavendra, Effective Mammogram Classification Using Various Texture Features, Journal of Advanced Research in Dynamical and Control Systems, Vol. 9. Sp– 12 / 2017.
- [10] Ding, C.H.Q.; Xiaofeng, H.; Hongyuan, Z.; Ming, G.; A Min-Max Cut Algorithm for Graph Partitioning and Data Clustering, ICDM 2001, 2001,107-114
- [11] Jain, A.K.; Dubes, R.C.; Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, NJ, U.S.A., 1988.
- [12] Huband, J.; Bezdek, J.C.; Hathaway, R.; Bigvat: Visual Assessment of Cluster Tendency for Large Data Sets, Pattern Recognition, 38(11), 2005, 1875-1886.
- [13] Huband, J.; Bezdek, J.C.; Hathaway, R.; Revised Visual Assessment of (Cluster) Tendency (reVAT), Fuzzy Information, (1), 2004,101-104
- [14] Havens, T.C.; Bezdek, J.C.; An Efficient Formulation of the Improved Visual Assessment of Clustering Tendency (iVAT) Algorithm, IEEE Transactions on Knowledge and Data Engineering, 24(5),2012, 813-822
- [15] Wang, L.; Uyen, T.V.Nguyen.;Bezdek, J.C.;Leckie, C.;Ramamohanarao, K.; iVAT and aVAT: Enhanced Visual Analysis for Cluster Tendency Assessment, PAKDD, (1), 2010, 16-27

- [16] C. Raghavendra, A. Kumaravel, S. Sivasubramanian, "Iris Technology: A Review on Iris Based Biometric Systems for Unique Human Identification", International Conference on Algorithms, Methodology, Models and Applications in Emerging Technologies – 2017, association With IEEE.
- [17] Bernd, F.; Thomas, Z.; Joachim, M.B.; Path Based Pairwise Data Clustering with Application to Texture Segmentation, EMMCVPR 2001, LNCS 2134, Springer-verlag Berlin Heidelberg, 2001, 235-250
- [18] Xiaofei, H.; Deng, C.; Yuanlong, S.; Hujun, B.; Jiawei, H.; Laplacian Regularized Gaussian Mixture Model for Data Clustering, IEEE Transactions on Knowledge and Data Engineering. 23(9), 2011, 1406-1418
- [19] Senoussaoui, M.; Kenny, P.; Stafylakis, T.; A study of the cosine distance-based mean shift for telephone speech diarization, IEEE Transactions on Audio, Speech, and Language Processing, 22(1), 2014, 217-227
- [20] Ester, M.; Kriegel, H.P.; Sander, J.; Xu, X; A Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, 1996, 67-82.
- [21] [http:// archive.ics.uci.edu/ml/datasets.html](http://archive.ics.uci.edu/ml/datasets.html)
- [22] Yanchang Z.; Zhang, S.; Generalized dimension-reduction framework for recent biased time series analysis, IEEE Transactions on data and knowledge engineering, 18(2), 2006, 231-244.
- [23] T. Padmapriya and V. Saminadan, "Inter-cell Load Balancing Technique for Multi- class Traffic in MIMO - LTE - A Networks", International Conference on Advanced Computer Science and Information Technology, Singapore, vol.3, no.8, July 2015.
- [24] S.V.Manikanthan and V.Rama "Optimal Performance of Key Predistribution Protocol In Wireless Sensor Networks" International Innovative Research Journal of Engineering and Technology, ISSN NO: 2456-1983, Vol-2, Issue –Special –March 2017.
- [25] S.V. Manikanthan, T. Padmapriya "An enhanced distributed evolved node-b architecture in 5G telecommunications network" International Journal of Engineering & Technology (UAE), Vol 7 Issues No (2.8) (2018) 248-254. March 2018

