International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878, Volume-8 Issue-5, January 2020

# Visual Social Data Clustersfor Effective Topics Tendnecy with Hybrid Machine Learning Techniques

**Upendar Penmetcha, K. Rajendra Prasad**

***Abstract:*** ***The machine learning is an emerging field in social classification of data, which enable the learning of social data patterns and classify the data by unsupervised approaches. Majorly, k-means and graph-based machine learning algorithms are used for discovering of social data clusters based on similarity features of user views, opinions. This paper presents the sentimental analysis of social users for the topics using the cluster tendency of derived clusters. The experimental of social data clusters and the cluster tendency are visualized for effective sentiment of topics analysis.***

***Keywords:*** ***Machine Learning, Classification, Cluster tendency, Social Data Clusters, Sentiment Analysis.***

## I. INTRODCUTION

Machine learning algorithms are the most important data mining techniques, which are used for classification of data based on similarity features of data objects for real-life social data. Social data is extracted from various sources, include, twitter, Facebook, and LinkedIn, etc. and its data is voluminous. Clustering of big social data is an emerging problem, which needs the initial assessment of pre-clusters estimation. Pre-determination of initial clusters is known as the cluster tendency. Related cluster tendency methods are surveyed for effective assessment of clusters in major clustering methods, such as k-means and graph-based clustering methods.These machine learning methods discover the quality of clusters for big social data. Procedure for finding quality of clusters through assessment of pre-estimated clusters is shown in Fig. 1, its derived proposed procedures finds the clusters estimation by visual access tendency (VAT) [8] and quality of clustering results are discovered by either k-means or minimum-spanning tree based clustering (graph based clustering), shortly known as MST-based clustering. These hybrid machine learning algorithms are known as VAT-based-k-means (VKM) and VAT-based-MST clustering (VMC), which are effectively works for big social datasets.

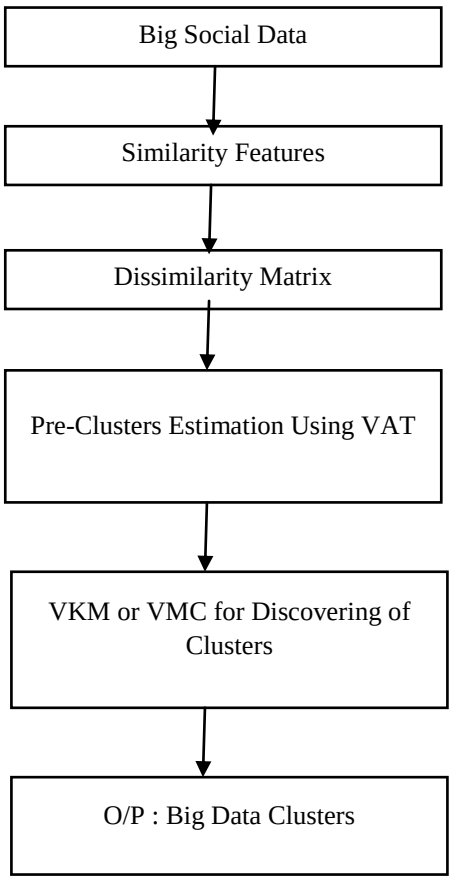



**Fig. 1. Big Data Clustering Using Machine Algorithms (VKM or VMC)**

Big social data is collected through the tweets and either in Euclidean space or cosine space for extraction of similarity features. Later, the dissimilarity matrix is derived and it is reordered using VAT [8] procedures for pre-clusters estimation compared. Proposed hybrid models, such as VKM or VMC are used for discovering of quality of big data clusters. Details of proposed hybrid machine learning algorithms are presented further in this paper.

Summary of the contributions of the work are described as follows:

Manuscript received on January 02, 2020.
Revised Manuscript received on January 15, 2020.
Manuscript published on January 30, 2020.

* Correspondence Author
**UpendarPenmetcha**,Dept. of CSE, Mahatma Gandhi Institute of Technology, Hyderabad, India.
**Dr. K Rajendra Prasad***, Professor and Head, Dept. of CSE, Institute of Aeronautical Engineering, Dundigal, Hyderabad, India. Email: krprgm@gmail.com

1. Similarity features are extracted from the social big data using either Euclidean or cosine space
2. Pre-clusters are estimated using the VAT procedure
3. Hybrid machine learning algorithms are developed for finding both cluster tendency as well as quality of clusters for big social data
4. Empirical analysis is performed for demonstrating the efficiency of proposed hybrid machine learning methods

Organization of the remaining part of the paper is described as follows: Section 2 presents the related work, Section 3 describes the proposed work, Section 4 overviews the experimental study, and Section 5 presents the conclusion and future scope of the work.

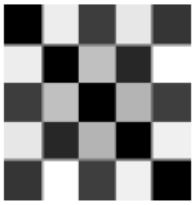
## II. RELATED WORK

Clustering methods are suitable for classification of unlabeled datasets using the approach of unsupervised machine learning. Top-clustering methods [1] are surveyed for the data partitioning of big data. These methods, include, k-means [3], MST-based clustering [5]. These are produce the quality of clusters when they have knowledge about pre-determined clusters. User interference is required for prediction of number of clusters (or cluster tendency). This predicted cluster tendency may be intractable and it is required to get exact pre-determine clusters. During the survey of related cluster assessment techniques, it is investigated that visual access tendency (VAT) [12] is one of the best approach for pre-determination of clusters. The basic steps of VAT procedure as follows:

1. Extraction of dissimilarity features for the set of data objects and place into matrix form, is known as dissimilarity matrix (DM) [17].
2. Re-order the DM using the logic of Prim's algorithm and placed the resulting values into another matrix, known as re-ordered dissimilarity matrix (RDM) [18].
3. Normalize the values of RDM between 0 and 1, and display the image of normalized RDM
4. Count the number of square shaped dark colored blocks along the diagonal, its count is referred as the number of clusters (or cluster tendency)
5. Save the output of cluster tendency.

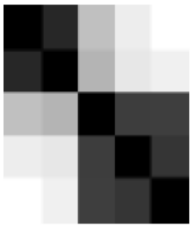
Following Fig. 2 shows the sample data of VAT procedure and respective VAT Images, which are shows the cluster tendency of given sample data. In this sample, pre-determined clusters were identified as '2'. i.e. cluster tendency of sample data is accessible with '2'. Its pre-estimation of cluster tendency would improve the quality of clustering results.

$$R = \begin{pmatrix} 0 & 0.73 & 0.19 & 0.71 & 0.16 \\ 0.73 & 0 & 0.59 & 0.12 & 0.78 \\ 0.19 & 0.59 & 0 & 0.55 & 0.19 \\ 0.71 & 0.12 & 0.55 & 0 & 0.74 \\ 0.16 & 0.78 & 0.19 & 0.74 & 0 \end{pmatrix} = I$$



**(a) Dissimilarity Image 'I' (before applying the VAT) [ ]**

$$\tilde{R} = \left(\begin{array}{cc|ccc} 0 & 0.12 & 0.59 & 0.73 & 0.78 \\ 0.12 & 0 & 0.55 & 0.71 & 0.74 \\ \hline 0.59 & 0.55 & 0 & 0.19 & 0.19 \\ 0.73 & 0.71 & 0.19 & 0 & 0.16 \\ 0.78 & 0.74 & 0.19 & 0.16 & 0 \end{array}\right) = \tilde{I}$$



**(b) After the VAT Procedure**

**Fig. 2 VAT Images Illustration**

VAT algorithms are developed using Euclidean space, later it is developed in cosine space for extraction of accurate similarity features, the cosine based VAT is known as cVAT [8]. Stare-of-the-art of VAT algorithms are studied and it is investigated that using these visual models, there is possible to estimate the correct number of clusters for big social data. Its pre-knowledge of number of clusters are accessible for big social for achieving of the best clustering results with the top-clustering machine learning algorithms [2].

k-means and MST-based clustering methods are need the cluster tendency information for discovering of quality of clusters. Visual models are access the cluster tendency information to overcome the problem of k-means and MST-based clustering methods [10].

In VAT, cluster tendency is estimated by extracting the information of number of clusters through visual count of number of square-shaped dark colored blocks along the diagonal of VAT Image. Fig. 2b shows the two square-shaped dark colored blocks along the diagonal of VAT Image. By this information, the given data shown in Fig 2a has totally two different clusters [11][13]. The VAT Image finds the dissimilarity information using either metrics of cosine or Euclidean distance formulas. In most of the cases, VAT computes the cluster tendency accurately using cosine space than Euclidean space [14]. Thus, the proposed work is defined in terms of VKM and VMC with considering of cosine space and it is presented in next following sub-sections.

## III. PROPOSED WORK

Two machine learning algorithms are proposed, namely, VAT based k-means (VKM) and VAT based MST-clustering (VMC) for discovering of both cluster tendency and clustering results.

These two are initially determines the dissimilarity information for the set of data objects, further, they are estimated prior number of clusters for producing the quality of clusters. These are presented in Algorithm 1 and Algorithm 2 respectively.

***Algorithm 1:*** VAT based k-means (VKM)

*Input:* D- Set of Data Objects, n-total number of objects

*Output:* Cluster Tendency, Clusters and its objects

*Method:*

1. Compute the dissimilarity matrix (DM) for D using cosine metric [4]
2. Find re-ordered dissimilarity matrix of DM, namely, RDM using the Prim's logic according to sorted order of dissimilarity information of DM as per VAT procedure [9]
3. Display Image of RDM, saved as 'VAT Image'
4. Print the output of cluster tendency, whereas cluster tendency is determined by counting of square-shaped dark colored blocks along the diagonal
5. Assign the cluster tendency to variable 'k' and print the clustering results using k-means procedure

***Algorithm 2:*** VAT based MST clustering (VMC)

*Input:* D- Set of Data Objects, n-total number of objects

*Output:* Cluster Tendency, Clusters and its objects

*Method:*

1. Compute the dissimilarity matrix in a similar way earlier proposed algorithm
2. As per the procedure given step 2 of VKM, here also RDM is derived using VAT [7]
3. Find the image of RDM, known as VAT Image in VMC
4. Access the prior number of clusters information from 'VMC VAT Image'
5. Derive the clustering results in sub-trees of minimum spanning tree (MST) form, in which each distinguished sub-tree is considered as cluster for similar set of objects.

The above algorithms are the hybrid machine learning techniques, which are initially determines the similarity features for the set of data objects, later, subtract their values from '1' for computing the dissimilarity matrix. The values of dissimilarity matrix are normalized between 0 and 1. Prim's logic construct the tree by adding the edges according to order of edge weights. Step 2 of both algorithms uses the same concept, in which edge weight refers to distance (or dissimilarity) between nodes (here nodes are denoted as data objects). In this step, all the dissimilarity values of pair of nodes are re-ordered, then computes the re-ordered dissimilarity matrix, it is shortly called as 'RDM'. Step 3 derives the image of RDM, known as VAT Image, which shows the square shaped dark colored blocks with the information re-ordered dissimilarity information. Most similar objects are grouped as clusters and these clusters are represented as square-shaped dark colored blocks in VAT Image. Step 4 extract the information of estimated clusters and its known as cluster tendency (for example, its say 'k' value in k-means). Step 5 of both algorithms use the value of cluster tendency for tracking of correct number of clusters in k-means and MST-based clustering algorithms.

Proposed hybrid machine learning algorithms are the most beneficial for unlabeled datasets, which are may be synthetic or real-life datasets. These algorithms address the problem of cluster tendency using visual assessment of clusters.

The clustering results are derived with the prior information of cluster tendency, i.e., here number of clusters 'k' value is known either in k-means and MST-based clustering methods, thus, they are efficient in producing of clustering results. In VMC, initially the cyclic graph is generated from the features of data objects, later, minimum spanning tree is constructed for the given graph using either Prim's [8] or Kruskal's [6] algorithms. In further steps of VMC, an inconsistency edge is identified and generated sub-trees of MST. Each sub-tree is considered as an individual cluster. Repeat the same procedure until all inconsistency edges are removed, in such cases, distinguished clusters are generated sequentially. The experimental results are demonstrated in next following sub-section. In VKM, the distance between mean of cluster and other objects are calculated in order to find tight clusters and relationships between associated similar objects. Time efficiency of proposed algorithms are demonstrated based on synthetic and benchmarked real-life datasets [20]. These two hybrid machine learning algorithms must derive the pre-estimated clusters information for accessing of complete clustering results for the given datasets. MST structures of given data are also presented in next section with visual representation [19].

## IV. EXPERIMENTAL STUDY

Proposed VKM and VMC are experimented on both synthetic and real-life datasets for demonstrating the performance for comparative analysis purpose. Description of datasets and run time of VKM and VMC are described in Table 1 and Table 2 for synthetic and real-life datasets respectively.

**TABLE 1. RUNTIME COMPARISION(SYNTHETIC DATASETS)**

| Data Sets | Measure Runtime in Seconds | |
|---|---|---|
| | VMC | VKM |
| S-1 | 0.80 | 0.70 |
| S-2 | 1.32 | 0.70 |
| S-3 | 1.57 | 1.06 |

**TABLE 2. RUNTIME COMPARISION (REAL DATASETS)**

| Data Sets | Measure Runtime in Seconds | |
|---|---|---|
| | VAT Runtime | PSVAT Runtime |
| Iris | 1.43 | 1.33 |
| Voting | 2.59 | 1.78 |
| Wine | 1.45 | 1.33 |
| Glass | 1.58 | 1.4 |

From this experimental results, it is observed that VAT-based-k-means runs as much faster than MST-based clustering.

Visual representation of synthetic datasets is shown in Fig.3 (in sequence from left to right is 3-clusters, 4-clusters, and 5-clusters)

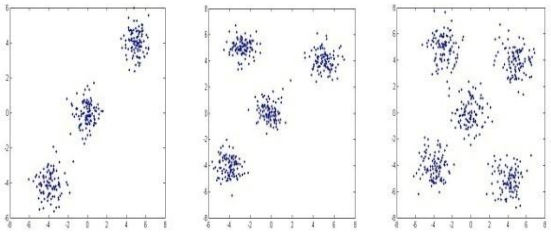


**Fig 3. Synthetic datasets**

**(S-1,S-2,andS-3 are from left to rignt)**

MST clusters are created after removing of inconsitent edges and recognition of inconsistent edges every incosistent edge removal divides a single tree into two clustered saperated trees. The figures(Fig. 4 ,Fig. 5,and Fig. 6) depicts the VAT based MST results of three synthetic datasets(S-1,S-2,andS-3 respectively).

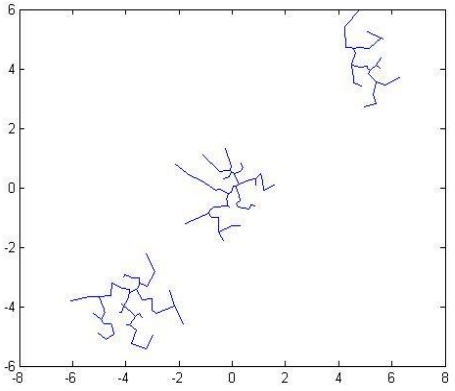


**Fig. 4. VAT-based MST Clustering results (S-1 dataset)**

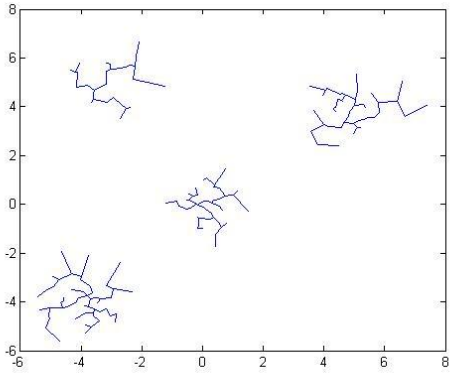


**Fig. 5. VAT-based MST Clustering results (S-2 dataset)**

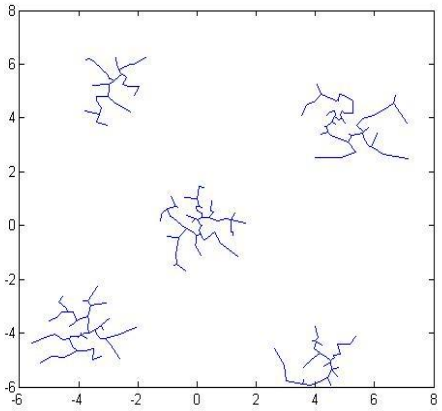


**Fig. 6. VAT-based MST Clustering results (S-3 dataset)**

Table 3 shows the evaluation of clustring performance using post-clusterng validity index measure. This table presents the clusters performance informative assessment as per the cluster tenedency.

**Table 3. DUNN'S INDEX: 'VAT' BASED MST CLUSTERING (VMC)**

| Datasets | Dunn's Index for Number of clusters (C) | | | | | Cluster Tendency |
|---|---|---|---|---|---|---|
| | C=2 | C=3 | C=4 | C=5 | C=6 | |
| S-1 | 0.22 | **0.45** | 0.18 | 0.22 | 0.23 | 3 |
| S-2 | 0.21 | 0.36 | **0.49** | 0.29 | 0.21 | 4 |
| S-3 | 0.21 | 0.16 | 0.32 | **0.43** | 0.16 | 5 |
| Iris | **0.45** | 0.16 | 0.26 | 0.24 | 0.25 | 2 |
| Voting | **0.48** | 0.28 | 0.26 | 0.24 | 0.26 | 2 |
| Wine | **0.22** | 0.01 | 0.02 | 0.03 | 0.07 | 2 |
| Glass | 0.01 | **0.04** | 0.02 | 0.03 | 0.02 | 3 |

It is proved that, it achives the best Dunn's cluster validity index value for respective cluster tendency value.

## V. CONCLUSION AND FUTURE SCOPE

Hybrid machine algorithms, namely, VKM and VMC are proposed in this paper for addressing the problem of prior assessment of cluster tendnecy. MST clusters genrations are visually depicted in this paper and respective performance is computed using the post cluster validity index measure , Dunn's index. Experimental study demonstrated that VAT-based k-means runs as faster than VAT-based MST clustering method for both synthetic and real-life datasets. Futhure scope of the work is to extend the procedures of VKM and VMC for big data with scalable approaches of proposed methods.

## REFERENCES


1. T.C. Havens, J.C. Bezdek, " An Efficient Formulation of the Improved Visual Assessment of Clustering Tendnecy (iVAT) Algorithm", IEEE Transactions on Knowledge and Data Engineering, Vol: 24,Issue:5, pp: 813-822,2012
2. Liang Wang, J.C. Bezdek, C.Leckie, and K. Ramamohanarao," Enhanced Visual Analysis for Cluster Tendency Assessment and Data Partitioning", IEEE Transactions on Knowledge and Data Engineering , Vol:22, Issue:10, Oct 2010
3. Liang Wang, C. Leckie, K. Ramamohanarao and J.C. Bezdek, " Automatically Determining the Number of Clusters in Unlabeled Data Sets", IEEE Transactions on Knowledge and Data Engineering, Vol:21, Issue: 3, Mar 2009
4. Jiawei Han and Micheline Kamber, Data Mining Concepts and Techniques, Elsevier, 2006
5. A.K. Jain and R.C. Dubes, Algorithms for Clustering data. Prentice Hall, 1988
6. Arun Pujari, Data Mining Techniques, University Press
7. M.Ester .P.Kriegel, J. Sander, and X.Xu, " A density based algorithms for discovering clusters in large databases with noise" , Int. Conf. on Knowledge Discovery and Data Mining,pp 226-231,1996
8. K Rajendra Prasad, C Raghavendra et.al, "Parallel Approach of Visualized Clustering Approach (VCA) for Effective Big Data Partitioning", Jour of Adv Research in Dynamical & Control Systems, Vol. 10, 04-Special Issue, 2018.
9. W.Wang, J.Yang, and R. Muntz, "STING: A Statistical Information Grid Approach to Spatial Data Mining", Int.Conf. on Very Large Databases,pp 186-195
10. Spec Vat: Enhanced Visual Cluster Analysis 2008 Eighth IEEE Conf on Data Mining
11. Wang D. Mitchell, " A Divide and Conquer Approach for Minimum Spanning Tree-Based Clustering", Vol21,July 2009,pp:945-958
12. C.T. Zahn. "Graph theoretical methods for detecting and describing gestalt clusters", IEEE Trans. Computers,vol. 20,no.1, pp. 68-86, Jan. 1971
13. Ying Xu, Victor Olman, Dong Xu, "Minimum spanning trees for gene expression data clustering", Genome Informatics 12:pg 24-33,2001
14. J.Bezdek and R.Hathaway, "VAT: A tool for visual assessment (cluster) tendency", in Proc. IJCNN, Honolulu, Hi,2002, pp.25-30
15. T.C. Havens, J.C.Bezdek, J.M.Keller,M. Popescu, " Dunn's Cluster Validity Index as Contrast Measure of VAT Images" Int Conf IEEE 2008
16. I.J.Sledge, J.M Huband, J.C. Bezdek, "(Automatic) Cluster Count Extraction from Unlabeled Data Sets", Fifth International Conf. On Fuzzy Systems and Knowledge Discovery, IEEE computer society, pg3-13
17. O.Grygorash, Yan Xhpu, Zach Jorgenson, " Minimum spanning tree based clustering algorithms"
18. A. Vathy-Fogarassy ,A.Kiss, and J.Abnoyi,"Hybrid Minimal Spanning tree based clustering and mixture of Gaussians based clustering algorithm", Foundations of Information and Knowledge systems, pp 313-330, Springer, 2006.
19. Liang Wang, Christopher Leckie, RamamohanaraoKotagiri, and James Bezdek, "Approximate pairwise clustering for largedatasets via sampling plus extension", Pattern Recognition,pp.225-235,Issue 44,2011
20. Liang Wang, Christopher Leckie, RamamohanaraoKotagiri, and James Bezdek, "Clustering on large relational datasets using active sampling and spectral extension" Int Proc 2007.


## AUTHORS PROFILE

**Upendar Penmetcha**, is studying his B.Tech in CSE Department at Mahatma Gandhi Institute of Technology, Hyderabad. His interested areas areMachine Learning, Data Science and Artificial Intelligence.

**Dr. K Rajendra,** Prasad is professor and head in the department of CSE. He received B.Tech degree in CSE from JNT University, Hyderabad; M.Tech degree in CSE from VTU, Belgaum; Ph.D. in CSE from JNT University, Anantapur. He has 14 years of indispensable experience in teaching. **Previously,** he served as a member in Board of Studies where he was actively involved in TEQIP work. He has published 15 papers in various international conferences and 20 papers in reputed journals. He is a member of CSI, IEEE professional societies. His major research interests are Data Mining, Pattern Recognition, Speech Processing, and Machine Learning.