International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Comprehensive Study on Density Peak Clustering and its Variants

Sarvani Anandarao*[1], Sweetlin Hemalatha Chellasamy[2]

Submitted: 13/11/2022 **Accepted**: 15/02/2023

Abstract: Clustering is a technique used to group similar datapoints/samples. Similar group of datapoints can be formed by using distance measure or by density. Density peak clustering (DPC) groups datapoints based on the density. This paper shows variations and improvements of DPC and also the performance of DPC over other clustering algorithms. This paper also addresses the problem in DPC with random selection of cut-off distance parameter(d_c). Local density of the datapoint is calculated based on d_c. The improper selection of d_c leads to wrong clustering results. The issue in the random choice of dc is addressed by using gini index or Gaussian function to make a valid guess on d_c.. Here we have chosen homogeneity, completeness, silhouette coefficient as the three parameters to compare results of DPC, DPC with gini index, DPC with gaussian function.

Keywords: *Density peak clustering (DPC), cut-off distance parameter, homogeneity, completeness, silhouette coefficient*

1. Introduction

Nowadays we find data everywhere. Classification and clustering are two major approaches employed to analysis the data. Classification [1] is referred as supervised learning while clustering as unsupervised learning. Mapping input data to labelled classes is called classification, $x \in R$ to $y \in 1, \dots. C$, where x is the input data, y is the class label, C is the total number of class labels. Unlike classification, in clustering predefined class labels are not present, but need to group the data into discrete sets. The input points having similar behaviour are grouped into one cluster and dissimilar points are placed in another cluster. Consider $A = \{a_1, a_{2,} \dots \dots \dots \dots \dots a_N\}$ as the input data points and calculate the distance matrix d_{ij}. Partitional clustering strive to bring out K partitions of A, $C = \{C_1, C_{2,\dots\dots\dots\dots\dots} C_k\}$, where

- $C_i \neq \emptyset, i = 1 \dots \dots \dots .. K;$
- $\sum_{i=1}^{K} C_i = A\,;$
- $C_i \cap C_j = \emptyset, i, j = 1, \dots \dots . K \; and \; i \neq j$

Clustering technique has shown promising results in various fields like market survey, classification of customer behaviour, medical data, recommender system, pattern recognition, social network analysis, digital image processing.

2. Existing Clustering Techniques

One of the clustering techniques is hierarchical. This is carried out in bottom-up or top-down style called agglomerative approach and divisive approach respectively. Initially in the top-down style all the datapoints are grouped into one single cluster, later based on the similarity the larger cluster is divided into number of small clusters until the desired similarity is reached. Now in the bottom-up style, initially every single datapoints is considered as individual cluster, later the similar datapoints are merged to form larger cluster till the desired number of clusters are reached. This clustering technique has the high time complexity of $O(kd^2)$, where number of clusters is denoted by ‘k’ and total number of datapoints with ‘d’. Another technique is K-means [7] clustering in which distance measure is used to group the similar datapoints. Initially, need to choose the number of clusters and cluster centroids for every cluster. The cluster centroid value is calculated recursively by taking the mean value of all the datapoints present in the cluster. This calculation of centroid is done for every iteration till stable centroid values are obtained. K-means have three limitations, first one is wrong selection of number of clusters and cluster centroid leads to improper clustering results. Second one is, can bring out accurate results only when shape of the cluster is spherical. Third one is, completely based on the distance measure is not a meaningful assumption. The time complexity of k-means clustering is $O(n^2)$, where total number of datapoints denoted by "n". The time complexity of k-means is lesser than hierarchical clustering. These two clustering algorithms are based on distance measure and fail in creating arbitrary shape of cluster while DPC considers

1,2 School of Computer Science and Engineering, Vellore Institute of Technology (VIT), Chennai 600127, Tamil Nadu, India
** Corresponding Author Email: sarvani.anandarao@gmail.com*

the densities instead of the distances. The area with greater number of data points is considered as the denser area. This denser area is considered as a cluster. The DPC performs well even on the noisy data and also returns arbitrary shape of clusters [8]. To prove this, experimental study has been done which is shown in the result section. This DPC algorithm has wide range of application some of them are autonomous vehicle navigation [9], moving object detection [10], electricity customer segmentation [11],document summarization [12] and overlapping community detection [13].

DPC [23] has shown the promising results in detecting clusters of arbitrary shapes and outliers. DPC consumes more execution time than the partitioning-based clustering method. Many real-world problems use both DPC and partitioning-based clustering to produce accurate results. Rodriguez and Laio proposed DPC algorithm [24]. Since many years DPC [25-35] is able to bring out meaningful results for many challenging problems. Accuracy of the data plays vital role in thermal Engineering, heat diffusion [25], this can be achieved by DPC. Quick identification of cluster centroid [26] by using potential entropy in DPC, improves the performance. DPC along with K-means [27] reduces the computational cost. This combination has shown the outstanding performance on many real world datasets, mixed data, cancer data sets.[28,29,30,34]. Merging DPC with k-nearest neighbours [31,32] and its applicability on synthetic and real-world data has replaced many existing clustering with its best performance. Numerical and categorical data is very common in real world dataset, handling this type of data is a challenging task which can be effectively clustered by merging DPC with fuzzy logic [33]. DPC have shown promising results on tensor imaging (DTI) of paediatric brain development and many medical related data[34]

The main motive of this paper is to show the variations in DPC techniques along with feasibility study of applying DPC to hot topic detection. The performance of the traditional DPC is compared with other two variations of DPC.

3. Density Peak Clustering and Its Variants

DPC starts by identifying density peaks (datapoint surrounded by maximum number of other datapoints) followed by the formation of clusters. The datapoints with highest local density and highest separation distance are considered as the density peaks. For the dataset X, ρ(x_i) is the number of data points in the neighbourhood of x_i and is defined as the local density which is given by equation 1:

$$\rho(x_i)=|B(x_i)| \qquad (1)$$

where B(x_i) denotes the neighbourhood of x_i and whose distance is less than user-specified parameter d_c. This is represented using the equation 2.

$$B(x_i) = \{ x_i \in X | d(x_i, x_j) < d_c \} \qquad (2)$$

where $d(x_i, x_j)$ is the distance between x_i, x_j.

Separation distance is given by equation 3

$$\delta_i = \begin{cases} \min(d_{ij}) & \text{if there exists data point with a local density } \rho(\text{xi}) \\ \max(d_{ij}) & otherwise \end{cases}$$

(3)

Separation distance as shown in eq.[3] and local density as shown in eq.[1] is calculated for every datapoint x_i. Then for each datapoint x_i, decision value (γ_i) must be calculated using the equation [4]

$$\gamma_i = \rho_i \delta_i \qquad (4)$$

Assign the datapoints to the cluster's centers. Top γ values are selected as cluster centers. Traditional DPC chose some random value as d_c. An inappropriate value of d_c shows the negative impact on the cluster formation. This is an important parameter in the calculation of local density and the cluster formation is completely dependent on the value of local density. Later many variations of DPC came into existence to show the solutions for the shortcomings in traditional DPC. The traditional DPC even fails to produce accurate local density when the dataset is small. To address this issue another density metric [36] is used shown in the equation 5.

$$\rho_i = \sum_j \exp\left(-\frac{d_{ij}^2}{d_c^2}\right) \qquad (5)$$

The equation [5] is applicable only when the dataset is small but there is no particular method to decide whether the dataset is small or large. So, the equation 6, 7 show the uniform solution for the calculation of local density without worrying about the dataset size.

The local density proposed based on K nearest neighbours DPC-KNN[37] is given by Eq. (6)

$$\rho_i = \exp\left(-\frac{1}{K}\sum_{j\epsilon KNN_i} d_{ij}^2\right) \qquad (6)$$

where K is the input parameter and KNN_i is the set of K nearest neighbours of point i.

Fuzzy weighted *K*-Nearest Neighbors Density Peak Clustering (FKNN-DPC) [37] proposed local density

$$\rho_i = \sum_{j\epsilon KNN_i} \exp(-d_{ij}) \qquad (7)$$

Separation distance is calculated by scanning half of the dataset in the traditional DPC. This leads to reduction in computation speed. This issue can be solved by scanning only the nearby datapoints [38]. This accelerates the computation. Early detection of non-peak datapoints avoids the unnecessary calculation of separation distance to obtain the density peaks. This is shown in the equation 8

$$\rho(x_i) < \max \rho(x_j) \text{ then } x_i \text{ is not density peak} \quad (8)$$

In the high-density area, if many density peaks are identified, only the first density peak is considered as the cluster center. The equation 9,10 shows the calculation of cluster member and cluster center respectively.

$$C_m = \min(\delta_{ij}) \quad if\ \delta_{ij} < d_c \qquad (9)$$

where C_m is the cluster member.

If the shortest distance of point with its remaining points is less than d_c. Then that point is considered as the cluster member.

$$C_c = \min(\delta_{ij}) \quad if\ \delta_{ij} > d_c \qquad (10)$$

Where C_c is the cluster center.

If the shortest distance of point with its remaining points is greater than d_c. Then that point is considered as the cluster center.

The points are said to have the highest similarity, if both the points have many common neighbouring datapoints. The common neighbouring datapoints is calculated by K-nearest neighbour [39,40]. This is shown in the equation 11.

$$SNN(x_i, x_j) = \tau(x_i) \cap \tau(x_j) \qquad (11)$$

where $SNN(x_i, x_j)$ is the shared nearest neighbour between any two datapoints i,j , $\tau(x_i)$ number of common points with the datapoint j, $\tau(x_j)$ number of common points with the datapoint i.

The equation 12 shows the calculation of similarity between two datapoints.

$$Sim(x_i, x_j) = \begin{cases} \frac{|SNN(x_i,x_j)|^2}{\sum_{p \in SNN(x_i,x_j)} (d_{ip} + d_{jp})}, & if\ x_i, x_j \in SNN(x_i, x_j) \\ 0 & \end{cases} \qquad (12)$$

..

The equation 13 shows the calculation of local density

$$\rho_i = \sum_{x_j \epsilon L(x_i)} Sim(x_i, x_j) \qquad (13)$$

Two points i,j come under same cluster, when atleast half of their neighbourhood points are common for the two point i,j.

In this paper we are concentrating on calculation of cut-off distance The challenge of random cut-off distance d_c can be addressed by usage of gaussian function and gini index [39]. The equation 14 shows the gini index and equation 15 shows the calculation of cut-off distance by using gaussian function.

$$G = 1 - \sum_{i=1}^{n} (\frac{\delta_i}{Z})^2 \qquad (14)$$

Where $z = \sum_{i=1}^{n} \delta_i$

$$d_c' = \sum_{j \in X, j \neq i} \exp\left(-\frac{{d_{ij}}^2}{{d_c}^2}\right) \qquad (15)$$

where, d_{ij} is the distance between any two points

4. Experiments and Results

4.1 DPC Vs Hierarchical Clustering Vs K-means

To prove the efficiency of over hierarchical clustering and k-means. We have done a small experimental study. We have created a dataset with two features. The two features are radius and number of data points in each radius. Here we have taken 3 radii to form 3 circles for easy understanding. The created dataset is shown in the Table 1.

Table.1. Created dataset and noise

Circle/Noise	Radius	Number of datapoints
Circle 1	500	1000
Circle 2	300	700
Circle 3	100	300
Noise	-600 to 600	

The k-means clustering is applied on the created dataset with '4' predefined number of clusters and 300 iterations. The figure 1 shows the clustering results of k-means. In the figure 1 red(0), blue(1), orange(2) , brown(3) indicates 4 clusters.

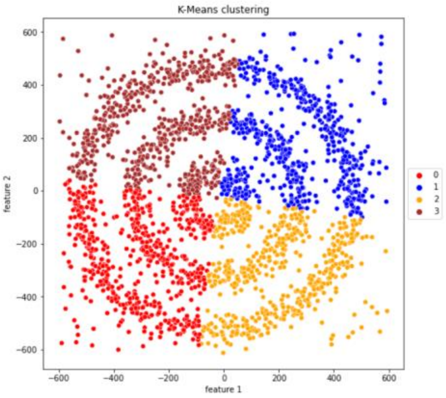


Fig..1. Clustering results of K-means

The hierarchical clustering is applied on the created dataset with '4' predefined number of clusters. The figure 2 shows the results of hierarchical clustering

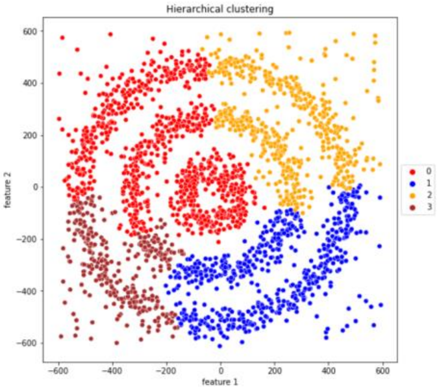


Fig..2. Results of hierarchical clustering

DPC is applied on the same dataset which is previously used for above two clustering techniques. The cut-off distance is calculated by drawing the k-distance graph by using NearestNeighbors technique. Figure 3 shows the k-graph where the maximum curvature is at '30'. So the cut-off distance is taken as '30'. Number of clusters are chosen automatically by the DPC algorithm i.e. 3 clusters. Figure 4 shows the results of DPC. In the figure 4 we can observe the 'red' colour dots which are outliers/noise which is not grouped into any of the three clusters. This process of eliminating outlier is not observed in hierarchical clustering and k-means clustering. This proves that DPC is able to identify the noisy data.

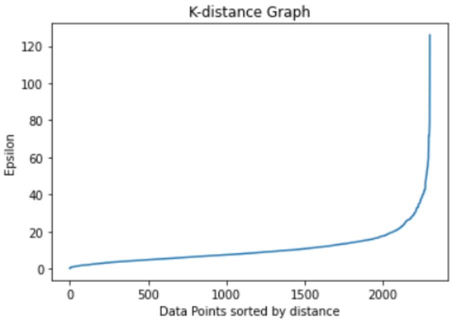


Fig.3. K-Graph

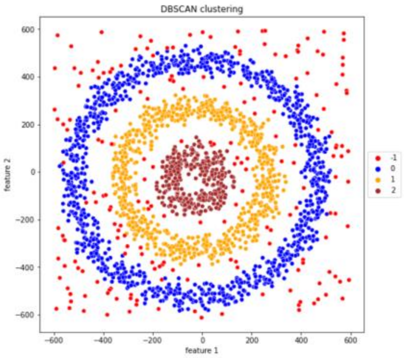


Fig..4. Results of density peak clustering

Table 2 shows the behaviour of DPC with increased size of the datapoints. The silhouette coefficient increased from 0.177 to 0.560 when the dataset size increased from 100 datapoints to 400 datapoints. This metric checks the correctness of the clustering result. The "1 value" indicates well-formed clusters, The "-1 value" indicates wrong formation of clusters, The "0 value" indicates insignificant distance between the clusters. Silhouette Score = (y-x)/max(x,y), where x= average intra-cluster distance, y= average inter-cluster distance. Homogeneity increased from 0.159 to 0.919, when all the datapoints in a cluster have the same label then its homogeneity is "1". Completeness increased from 0.431 to 0.786. When all the datapoints of same label present in dataset are assigned to same cluster, then its completeness is "1". V-measure increased from 0.232 to 0.847.

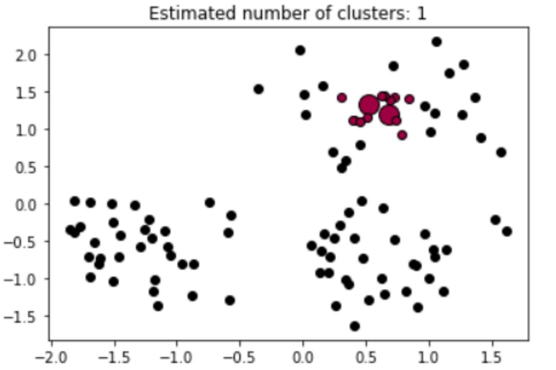
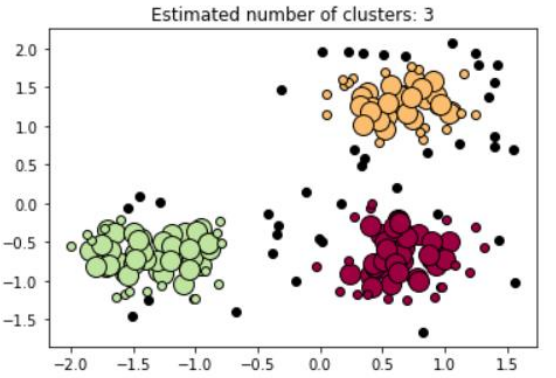
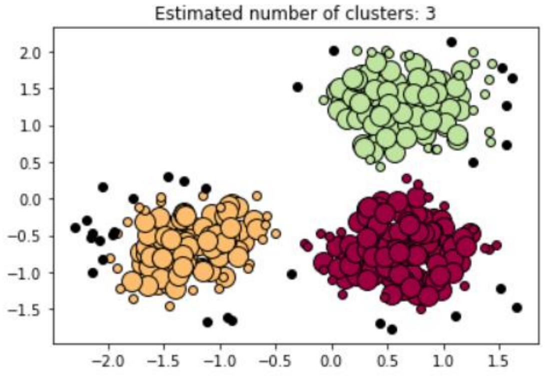
Homogeneity measures how many datapoints are similar in a cluster. It ranges from 0 to 1. This is ratio of number of datapoints labelled 'a' in a cluster 'b' by total number of datapoints in cluster 'b'. If all the datapoints in a single cluster have the same label then then the value of homogeneity equals to "1". Completeness calculates the number of similar datapoint put together. This is ratio of number of datapoints labelled 'a' in a cluster 'b' by total number of datapoints labelled 'a'. It ranges from 0 to 1. If all the datapoints of label 'a' are under one single cluster then the value of completeness equals to "1". Silhouette Coefficient measure the goodness of the cluster. This ranges between -1 to 1. The "1" indicates the cluster are far apart with clear boundary.

V-measure is the harmonic mean between homogeneity and completeness. The equation 16 shows the calculation of v-measure.

$$\mathrm{V-measure} = 2 * \frac{\mathrm{h*c}}{\mathrm{h+c}} \qquad (16)$$

where h *and* c are homogeneity and completeness respectively.

Table 2: Behaviour of DPC

Number of datapoints	Results	Clusters formed
100	Estimated number of clusters: 1 Estimated number of noise points: 86 Homogeneity: 0.159 Completeness: 0.431 V-measure: 0.232 Silhouette Coefficient: 0.177	
200	Estimated number of clusters: 3 Estimated number of noise points: 40 Homogeneity: 0.811 Completeness: 0.648 V-measure: 0.720 Silhouette Coefficient: 0.457	
400	Estimated number of clusters: 3 Estimated number of noise points: 31 Homogeneity: 0.919 Completeness: 0.786 V-measure: 0.847 Silhouette Coefficient: 0.560	

4.2 DPC Vs DPC with gini index Vs DPC with gaussian function

DPC provides promising results for non-spherical clusters and number of clusters is determined automatically based on the density and peaks. However, the cut-off distance d_c is selected manually, and hence wrong selection of d_c leads to improper local density and results. In the traditional DPC, d_c is selected manually in a random fashion. This issue can be resolved by adopting the variants of DPC such as DPC with gini index and DPC with Gaussian function. Here we have considered homogeneity, completeness, silhouette coefficient as parameters for comparing the results. The comparison results are shown in the Table 3. In the Table 3, the homogeneity is higher for DPC with gaussian than DPC with gini index and DPC. The completeness for both DPC and DPC with gaussian are almost equal. The Silhouette Coefficient is higher for DPC with gaussian than DPC and then DPC with gini index. This shows, DPC with gaussian has given better result than the other two.

Table 3: Comparison results

	Homogeneity	**Completeness**	**Silhouette Coefficient**
DPC	0.577	0.795	0.406
DPC with gini index	0.670	0.539	0.293
DPC with gaussian	**0.919**	**0.786**	**0.560**

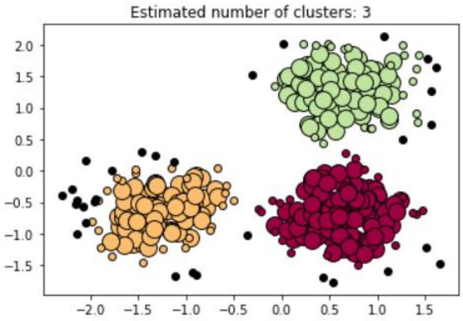


Fig. 5. Density peak clustering with gaussian

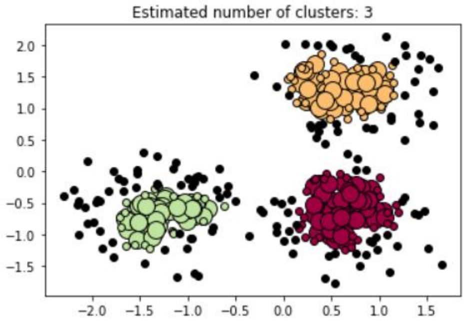


Fig. 6. Density peak clustering with gini index

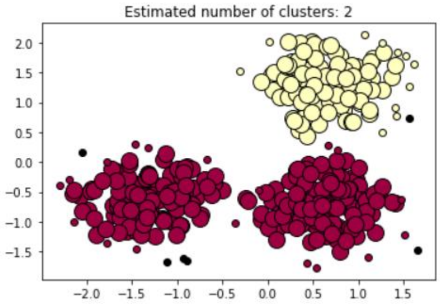


Fig. 7. Density peak clustering

5. Feasibility Study of DPC for Hot Topic Detection

5.1 Text Clustering

Document/Text Clustering is one among the various applications of clustering. Text mining is used in many trending fields [2] like information retrieval (IR), computational linguistics and data mining. Text clustering [3] is useful in predicting the topic or grouping the text related data. The three major steps in Text Clustering are text pre-processing, feature extraction and clustering. Cleaning the unstructured data is carried out in text pre-processing [4]. This step is a combination of stop words removal, stemming and TF/scoring technique. The conventional methods of stopward removal are Zipf's Law, the mutual information method, term based random sampling (TBRS). Text cleaning can be achieved by regular expression(re), NLTK (Natural language toolkit) [5] and spacy libraries for removing stop words, named entity recognition, part of speech tagging, phrase matching, expand shortened, lower casing, remove unwanted words/digits. To process any textual data, representation of text to vector format is mandate this can be done by bag of words (BOW) or TF-IDF technique. BOW look for unique words in documents referred to as vocabulary. Sentences/documents can be represented with vector format by placing 1 for the presence of words and 0 otherwise. Alternatively, the feature vector can be the frequency of the words in the document or it can be the TF/IDF score. Feature vector is fed as an input to clustering algorithm. This provides group of similar documents/similar sentences referred as clusters.

Effective text analysis can be performed by using text clustering [6]. Three basic stages of text clustering are

5.2 Hot topic detection

The papers [14-22] discussed the conventional hot topic detection. This section discusses the various applications of DPC.

Large amount of information is available in professional blogs. Analysis of candidate topics results in detection of hot topic in professional blogs [14]. Keywords with high frequency and its co-occurrences keywords are picked to construct word network then by analysing the word network candidate topic is identified. Opinion analysis is used to construct opinion network by considering topics in different time interval. Merging user participation degree, opinion degree and candidate topics results in hot topic.

Hokama et al. [15] detected hot topics by hierarchical clustering in relation to the timestamp. He et al. [16] proposed an advanced TF-IDF model and incremental clustering algorithm to identify the hot events. In Chen et al. [17], hot topic is identified by extracting the sentences with high frequency words distributed over time. Zhou et al. [18] extricate trend topics by clustering the words using density-based clustering.

Yamanaka et al. [19] showed the event detection using extraction method in targeted area. This uses support vector machine (SVM) model to label the messages by considering GPS information. Message labels helps to group the message. Finally, burst-detection method is applied on every cluster. Dense time interval between the messages is detected using burst-detection method, but this needs a predefined query set. Tweets along with their corresponding words form the weighted matrix. Words are clustered based on similarity measure to form the topics. Sakaki et al. [20] showed the event-detection method particular for twitter. This method is limited only to certain words related to the event and also, adaptability is lacking in this method. Relationship between every pair of words in tweets are built by giving weightage to the link. The weightage of the link is directly proportional to the frequency of word combination present in other tweets. Clustering method is used to group the words into topics and later burst-detection method detects the frequency of the topic. Hot topics related to health from the online communities/sites [21] can be extracted by automatic topic detection. Keyword-based features and medical domain-specific features are integrated to extract

document level, sentence level or word level. Document level clustering rearranges texts on the same topic. News stories, emails, search engines utilize document clustering to group the similar content whereas in "sentence level clustering" sentences gathered from numerous documents are clustered. Tweet analysis comes under sentence level clustering and in word level clustering words having similar meaning are grouped together.

the messages from online health communities/sites. Similarity between news stories and topics based on temporal distance factor [22] produced better result than the traditional measures and followed by short-term topic detection by using weighted term.

6. Feasibility of DPC for Hot topic Detection

DPC considers denser datapoints as the cluster centroids. Similarly, to assign a label to the tweet/sentence highest frequency word acts like a centroid. The highest frequency is one of the important parameter to assign accurate label to the text. This similarity bring out feasible solution by adopting DPC and its variants for text classification

7. Conclusion

This paper showed a comprehensive review of hot topic detection done by using various methods. In addition, this also gave an overview of various versions and improvements of the existing DPC and also able to highlight the issue with the existing DPC. This existing issue is resolved up to some extent by two solutions discussed in this paper i.e., gini index and gaussian function along with their corresponding results. The DPC is able to bring out favourable results. Homogeneity, Completeness, Silhouette Coefficient are the three parameters considered to compare with the existing algorithm. This study is limited to only certain DPC and its variations. Our future study focuses on applicability of improved density peak clustering on hot topic detection.

Conflicts of interest

The authors declare no conflicts of interest.

Reference

[1] Xu, Rui, and Donald Wunsch. "Survey of clustering algorithms." IEEE Transactions on neural networks 16.3 (2005): 645-678.

[2]Salloum, S.A., Al-Emran, M., Monem, A.A., Shaalan, K.: A Survey of text mining in social media: facebook and twitter perspectives. Adv. Sci. Technol. Eng. Syst. J. (2017)

[3] https://devopedia.org/text-clustering

[4] Vijayarani, S., Ms J. Ilamathi, and Ms Nithya. "Preprocessing techniques for text mining-an overview." International Journal of Computer

Science & Communication Networks 5.1 (2015): 7-16

[5] Salloum, Said A., et al. "Using text mining techniques for extracting information from research articles." Intelligent natural language processing: Trends and Applications. Springer, Cham, 2018. 373-397

[6] Clifton, C., Cooley, R.: TopCat: Data mining for topic identification in a text corpus. In: European Conference on Principles of Data Mining and Knowledge Discovery, pp. 174–183. Springer, Heidelberg (1999)

[7] K. P. Sinaga and M. Yang, "Unsupervised K-Means Clustering Algorithm," in IEEE Access, vol. 8, pp. 80716-80727, 2020, doi: 10.1109/ACCESS.2020.2988796

[8] Patra, Bidyut Kr, Sukumar Nandi, and P. Viswanath. "A distance based clustering method for arbitrary shaped clusters in large datasets." Pattern Recognition 44.12 (2011): 2862-2870

[9] Lu, Kaiyue, Siyu Xia, and Chao Xia. "Clustering based road detection method." 2015 34th Chinese Control Conference (CCC). IEEE, 2015

[10] Zhang, Yuchi, et al. "A new algorithm for fast and accurate moving object detection based on motion segmentation by clustering." 2017 Fifteenth IAPR International Conference on Machine Vision Applications (MVA). IEEE, 2017

[11] Wang, Yi, et al. "Clustering of electricity consumption behavior dynamics toward big data applications." IEEE transactions on smart grid 7.5 (2016): 2437-2447

[12] Wang, Baoyan, et al. "Density peaks clustering based integrate framework for multi-document summarization." CAAI Transactions on Intelligence Technology 2.1 (2017): 26-30.

[13] Bai, Xueying, Peilin Yang, and Xiaohu Shi. "An overlapping community detection algorithm based on density peaks." Neurocomputing 226 (2017): 7-15

[14] Zhou, Erzhong, Ning Zhong, and Yuefeng Li. "Hot topic detection in professional blogs." International Conference on Active Media Technology. Springer, Berlin, Heidelberg, 2011

[15] Hokama, T., Kitagawa, H.: Detecting Hot Topics about a Person from Blogspace. In: Proc. of the Sixteenth European-Japaness Conference on Information Modeling and Knowledge Bases, pp. 290–294 (2006).

[16] He, T.T., Qu, G.Z., Li, S.W., Tu, X.H., Zhong, Y., Ren, H.: Semi-automatic Hot Event Detection. In: Proc. of the Second International Conference on Advanced Data Mining and Applications, pp. 1008–1016 (2006).

[17] Chen, K.Y., Luesukprasert, L., Chou, S.C.T.: Hot Topic Extraction Based on Timeline Analysis and Multidimensional Sentence Modeling. IEEE Transactions on Knowledge and Data Engineering 19(8), 1016–1025 (2007).

[18] Zhou, Y.D., Sun, Q.D., Guan, X.H., Li, W., Tao, J.: Internet Popular Topics Extraction of Traffic Content Words Correlation. Journal of Xian Jiao Tong University 41(10), 1142–1145 (2007)

[19] T. Yamanaka, Y. Tanaka, Y. Hijikata, and S. Nishida, “A Supporting System for Situation Assessment using Text Data with Spatio-temporal Information,” Journal of Japan Society for Fuzzy Theory and Intelligent Informatics, Vol. 22, No. 6. pp. 691–706, 2010. (in Japanese).

[20] T. Sakaki, M. Okazaki, and Y. Matsuo, “Earthquake Shakes Twitter Users: Real-time Event Detection by Social Sensors,” Proc. the 19th International Conference on World Wide Web (WWW), pp. 851–860, 2010

[21] Lu Y, Zhang P, Liu J, Li J, Deng S (2013) Health-Related Hot Topic Detection in Online Communities Using Text Clustering.

[22] Yu, RuiGuo, et al. "Online hot topic detection from web news archive in short terms." 2014 11th International Conference on Fuzzy Systems And Knowledge Discovery (Fskd). IEEE, 2014

[23] Liu, Peiyu, et al. "A text clustering algorithm based on find of density peaks." 2015 7th International Conference on Information Technology in Medicine and Education (ITME). IEEE, 2015

[24] Rodriguez, A.; Laio, A. Clustering by fast search and find of density peaks. Science 2014, 344, 1492

[25] Mehmood, R.; Zhang, G.; Bie, R.; Dawood, H.; Ahmad, H. Clustering by fast search and find of density peaks via heat diffusion. Neurocomput. 2016, 208, 210–217.

[26] Wang, S.; Wang, D.; Li, C.; Li, Y.; Ding, G. Clustering by fast search and find of density peaks with data field. Chinese J. Electron. 2016, 25, 397–402.

[27] Bai, L.; Cheng, X.; Liang, J.; Shen, H.; Guo, Y. Fast density clustering strategies based on the k-means algorithm. Pattern Recognit. 2017, 71, 375–386.

[28] Mehmood, R.; El-Ashram, S.; Bie, R.; Dawood, H.; Kos, A. Clustering by fast search and merge of local density peaks for gene expression microarray data. Sci. Reports 2017, 7, 45602.

[29] Liu, S.; Zhou, B.; Huang, D.; Shen, L. Clustering mixed data by fast search and find of density peaks. Math. Problems Eng. 2017, 2017, 7.

[30] Li, Z.; Tang, Y. Comparative density peaks clustering. Expert Syst. Appl. 2018, 95, 236–247.

[31] Du, M.; Ding, S.; Jia, H. Study on density peaks clustering based on k-nearest neighbors and principal component analysis. Knowledge-Based Syst. 2016, 99, 135–145.

[32] Yaohui, L.; Zhengming, M.; Fang, Y. Adaptive density peak clustering based on k-nearest neighbors with aggregating strategy. Knowledge-Based Syst. 2017, 133, 208–220.

[33] Ding, S.; Du, M.; Sun, T.; Xu, X.; Xue, Y. An entropy-based density peaks clustering algorithm for mixed type data employing fuzzy neighborhood. Knowledge-Based Syst. 2017, 133, 294–313.

[34] Yang, X.-H.; Zhu, Q.-P.; Huang, Y.-J.; Xiao, J.; Wang, L.; Tong, F.-C. Parameter-free laplacian centrality peaks clustering. Pattern Recognit. Letters 2017, 100, 167–173.

[35] Cheng, S.; Duan, Y.; Fan, X.; Zhang, D.; Cheng, H. Review of Fast Density-Peaks Clustering and Its Application to Pediatric White Matter Tracts. Annual Conference on Medical Image Understanding and Analysis. Springer International Publishing: Cham, Switzerland, 2017; pp 436–447.

[36] Yaohui, Liu, Ma Zhengming, and Yu Fang. "Adaptive density peak clustering based on K-nearest neighbors with aggregating strategy." Knowledge-Based Systems 133 (2017): 208-220.

[37] Xie, J. Y., Gao, H. C., Xie, W. X., Liu, X. H., Grant, P. W., Aug. 2016. Robust clustering by detecting density peaks and assigning points based on fuzzy weighted k-nearest neighbors. Information Sciences 354, 19–40.

[38] Lin, Jun-Lin. "Accelerating Density Peak Clustering Algorithm." Symmetry 11.7 (2019): 859.

[39] Wang, Zhechuan, and Yuping Wang. "A new density peak clustering algorithm for automatically determining clustering centers." 2020 International Workshop on Electronic Communication and Artificial Intelligence (IWECAI). IEEE, 2020.

[40] Lv, Yi, Mandan Liu, and Yue Xiang. "Fast Searching Density Peak Clustering Algorithm Based on Shared Nearest Neighbor and Adaptive Clustering Center." Symmetry 12.12 (2020): 2014