EURASIP Journal on Advances in Signal Processing

**RESEARCH** **Open Access**

# Analysis of influencing factors on excellent teachers' professional growth based on DB-Kmeans method

Xu Gao[1,3], Xiaoming Ding[1,3*], Tingting Han[1,3] and Yueyuan Kang[2]

*Correspondence:
xmding@tjnu.edu.cn

[1] College of Artificial Intelligence, Tianjin Normal University, Tianjin 300387, China
[2] Faculty of Education, Tianjin Normal University, Tianjin 300387, China
[3] Tianjin Key Laboratory of Wireless Mobile Communications and Power Transmission, Tianjin Normal University, Tianjin 300387, China

## Abstract

The Kmeans clustering algorithm is widely used for the advantages of simplicity and efficient operation. However, the lack of clustering centers in the algorithm usually causes incorrect category of some discrete points. Therefore, in order to obtain more accurate clustering results when studying the factors affecting the professional growth of outstanding teachers, this paper proposes an improved algorithm of Kmeans combined with DBSCAN. Observing the clustering results of the influencing factors and calculating the evaluation standard values of the clustering results, it is found that the optimized DB-Kmeans algorithm has obvious improvements in the accuracy of the clustering results, and the clustering effect of the algorithm on edge points is more advantageous than the original algorithms according to the scatter diagram.

**Keywords:** Clustering, Kmeans, DBSCAN, Education, Teachers' professional growth

## 1 Introduction

Rejuvenating the country through science and education is an important policy of our country. The professional growth of teachers affects the development and future of national education. For an ordinary teacher to grow into an excellent teacher, in addition to his/her own efforts, he/she also needs to learn useful experience from other excellent teachers, which can effectively help his/her own growth. The interview records of excellent teachers are an effective summary of teachers' professional growth experience. Extracting key information from these interview texts and clustering the influencing factors of excellent teachers' professional growth can systematically provide valuable guidance for the professional development of teachers. It is of great and far-reaching significance to improve the professional quality of teachers and promote the development of education in our country.

How to mine and analyze the interview texts of these excellent teachers is of great significance and research value. Under the modern background, the information retrieval of texts puts forward higher requirements for the clustering algorithms. In order to obtain more accurate and effective information more efficiently, it is necessary to optimize the traditional clustering algorithm to deal with all kinds of texts to achieve more efficient in-depth analysis. In this paper, when researching the professional growth factors from the interview texts in the growth process of outstanding teachers, the Kmeans and Density-Based Spatial Clustering of Applications with Noise (DBSCAN) clustering algorithms are improved, and DB-Kmeans (DBSCAN—Kmeans) is proposed. The initial value of the clustering is optimized, and the selection method of the cluster centroid is improved, which effectively improves the accuracy of the clustering results.

## 2 Related works

Kmeans clustering is an unsupervised method and a common partition-based clustering algorithm, which is widely used in text analysis and cluster analysis. However, due to the lack of cluster centers in the Kmeans algorithm, it can still be seen that many discrete points are not classified into the correct category. DBSCAN clustering algorithm is a classical density-based spatial clustering algorithm. The algorithm starts with a randomly selected core point and recursively classifies points that satisfy the density requirement. Finally, the maximized area containing the core points and boundary points is obtained. The DBSCAN algorithm does not need to specify the number of clusters in advance, but only needs two parameters: Eps (Epsilon) and MinPts (Minimum Points). However, this algorithm is computationally inefficient, and the computation speed is slow for relatively large datasets. Because the effect of the DBSCAN clustering algorithm depends on the parameters Eps and MinPts, and improper selection of parameters will directly lead to the decline of the clustering quality, it is necessary to conduct multiple experiments to obtain a set of values with better effects. With the continuous development of technology, many scholars have made great improvements to the K-center clustering algorithm and DBSCAN, bringing great benefits to the mining of big data and information acquisition.

Wu Ying proposed a Canopy-Kmeans clustering algorithm. First, the Canopy algorithm was used to cluster the samples, and then, the initial clustering was obtained. The clustering result was used as the initial center and number of clusters of the Kmeans algorithm to get the result [1]. Gao Xin proposed a DT-Kmeans clustering algorithm, which first randomly selected a cluster center point, and determined the remaining clusters according to the data object density information and the distance information between the data object and the existing cluster center point [2]. Yan Minghui et al. proposed the introduction of Gaussian kernel density estimation to obtain the maximum probability to improve the way of Kmeans cluster center acquisition and finally improved the research effect of the traditional method [3]. Hima Bindu et al. proposed an improved algorithm of Firefly Algorithm (FA) mixed with Kmeans to find the optimal cluster center [4]. Valarmathy proposed a clustering algorithm combining DBSCAN density clustering and K-Distance tree algorithm [5]. Manogaran et al. proposed a modeling method combining Hidden Markov Model (HMM) and DBSCAN with GMM [6, 7]. Zhong Jun et al. proposed a hybrid algorithm of convolutional auto-encoding and Gaussian mixture, which was applied to the feature extraction of ECG signals, and saved a lot of time and effort of manual labeling [8]. Shi Yongge et al. proposed a hybrid algorithm of Kmeans and Extreme Gradient Boosting (XGBoost) to mine designated telecom customers with special behaviors from the vast voice communication records of telecom companies [9].

In view of the shortcomings of the Kmeans and DBSCAN algorithms and the hybrid algorithm idea proposed by the above scholars, this paper proposes an improved algorithm of DBSCAN combined with the Kmeans algorithm. The accuracy of the results in this paper is improved, and it provides data support for our research on the influencing factors of the growth process of outstanding teachers.

## 3 Kmeans and DBSCAN algorithms

The Kmeans algorithm first randomly selects $K$ objects as cluster centers, then assigns the sample points to the class with the closest centroid according to the Euclidean distance, finally calculates the mean of the sample points in the class and updates the centroids until the results converge.

The specific steps of the algorithm are as follows:

- Randomly generate $K$ centroids;
- Calculate the distance between all points in the sample and a random centroid, and classify each data into the cluster corresponding to the centroid that is closest to it. The distance between the object and the random centroid is the Euclidean distance. The formula is as (1);

$$d(x_t, z_k) = \sqrt{\sum_{i=1}^{N} (x_t, z_k)^2} \tag{1}$$

  Among them, $N$ represents the sample set $\{x_1, x_2, \ldots, x_t\}$, $\{z_1, z_2, \ldots, z_k\}$ represents $K$ centroids, and the sample points in $N$ are divided into the class closest to the centroid [10].
- Recompute the $K$ centroids based on the average distance between all points of the class result;
- Repeat steps 2 and 3 until the sum of the distances of all sample points and their corresponding centroids of the class is minimized. The results tend to converge through multiple iterations.

The algorithm flowchart of Kmeans is shown in Fig. 1.

DBSCAN can assume that the clustering results can be determined by the tightness of the sample distribution. The data that are clustered into the same category are closely connected, that is, there must be data belonging to the same category around a certain

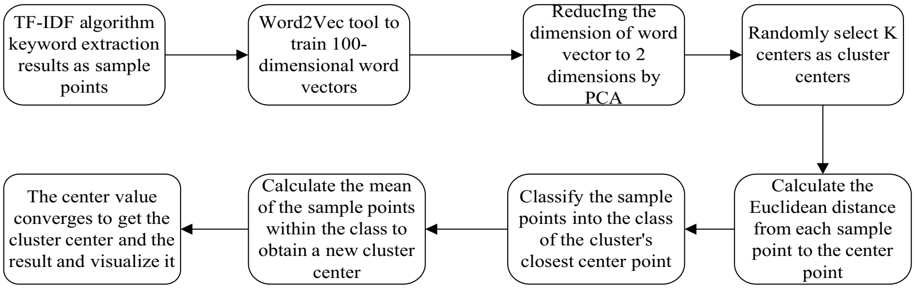

TF-IDF algorithm keyword extraction results as sample points
Word2Vec tool to train 100-dimensional word vectors
ReducIng the dimension of word vector to 2 dimensions by PCA
Randomly select K centers as cluster centers
Calculate the Euclidean distance from each sample point to the center point
Classify the sample points into the class of the cluster's closest center point
Calculate the mean of the sample points within the class to obtain a new cluster center
The center value converges to get the cluster center and the result and visualize it

**Fig. 1** Flowchart of Kmeans clustering algorithm

data in the sample. The final clustering result is obtained by dividing all the closely connected words in the sample set into categories and displaying the results in the form of scatter plots. Different categories are represented by different colors, which are presented with a more intuitive visual experience.

According to [11], for the sample set $D=(x_1, x_2, \ldots, x_\mathrm{m})$, the DBSCAN algorithm includes 5 core definitions in the implementation process: Eps neighborhood, core object and boundary object, density direct, density reachable and density connected. There can be one or more core points in the cluster. If there is only one core point, other non-core point samples in the cluster are in the Eps neighborhood of this core point. If there are multiple core points, there must be one other core point in the Eps neighborhood of any core point in the cluster, otherwise the two core points cannot be density reachable. The collection of all samples in the Eps neighborhood of these core points forms a DBSCAN cluster.

The quality of the DBSCAN clustering algorithm depends on the parameters Eps and MinPts, so it is necessary to conduct multiple experiments to obtain a set of values with better quality. After many experiments, Eps = 1 and MinPts = 1 are selected. Kmeans and DBSCAN have their own advantages and disadvantages in the implementation process. The comparison results are as follows in Table 1.

It can be seen from the table that the Kmeans algorithm has simpler parameters than DBSCAN, is easy to implement and does not take too much time. On the contrary, there are many parameters in the DBSCAN algorithm, which have a great impact on the clustering results, but it does not need to specify the number of *K*, and the cluster centers all exist in the data samples, while the Kmeans algorithm is a randomly assigned centroid or a value calculated from the mean, not necessarily real in the sample.

## 4 The optimized DB-Kmeans algorithm

In view of the shortcomings of DBSCAN and Kmeans algorithms, this paper proposes a hybrid method of Kmeans and DBSCAN algorithms, referred to as DB-Kmeans. This algorithm can maximize the advantages of Kmeans and DBSCAN algorithms, and avoid the shortcomings to some extent to the clustering results.

Firstly, DBSCAN algorithm is used to perform rough clustering to obtain the number of cluster categories and to cluster center points, and then, Kmeans algorithm is performed for further clustering. This processing can benefit from the no need of *K*

**Table 1** Analysis of advantages and disadvantages of clustering algorithms

| Algorithm name | Advantages | Disadvantages |
|---|---|---|
| DBSCAN | (1) No need specify the number of clusters in advance<br>(2) Outliers and clusters of any shape can be found | (1) When the sample data is large, the clustering convergence time is long<br>(2) When the sample data is quite different, the clustering effect is poor<br>(3) The parameters are complex |
| Kmeans | (1) The clustering principle is simple and there are few parameters, so the clustering time is fast<br>(2) There are few parameters, so the process is simple<br>(3) The clustering effect is good and the interpretability is strong | (1) Specify the *K* value in advance, and the selection is not easy to grasp<br>(2) Generally, only applicable to convex datasets<br>(3) The initial value is completely random, so the results may belong to the local optimal solution |

value to obtain global optimal solution, and the results also can avoid the shortcoming of being sensitive to noise points and abnormal points of Kmeans algorithm. The clustering steps of DB-Kmeans are as follows:

- Through the initial clustering of the DBSCAN algorithm, all the data are divided according to the density, and the cluster center point is obtained;
- Use the cluster center and $K$ value in the above results as the initial centroid and the number of categories, respectively;
- Get the final clustering result and scatter plot based on Kmeans.

This algorithm solves the problem of slow clustering speed of DBSCAN algorithm and can greatly speed up the algorithm. The more accurate initial value is provided for Kmeans clustering algorithm, which is of great significance to the final segmentation result. The algorithm flowchart is shown in Fig. 2. The pseudocode is shown in Table 2 below.

## 5 Results of experiments

The experiment is performed on the platform with processing CPU Intel(R) CoreTMi5-1035G1, the Samsung 16G DDR4-3200 memory and Windows10 system. The development environment is Anaconda3 with the programming language python 3.6. The main data include the results of keyword extraction from 100 long texts, which are from the translation of the interview manuscripts of excellent teachers by laboratory personnel and the interview content of teachers from the open resources online. There are 550 keywords as a sample of subsequent keyword clusters.

The manually extracted keywords are vectorized, and the keywords are clustered on the two-dimensional vector. The scatter plot is shown in Fig. 3 below. In these scattergrams, different colors represent different clusters. Through many experiments in this research, the clustering result is the best when $K=7$, so 7 clusters of different colors can be seen in the figure.

Looking at the scattergrams, a lot of data in Figs. 1 and 2 that are not clustered together correctly or are at the category boundary are divided into appropriate categories in Fig. 3. In order to analyze the results more accurately and objectively, the results of the 6 clustering algorithms used in this paper are labeled according to the category of the cluster to compare with the results of manual clustering. The confusion moment certificate is exported to show the corresponding results of the real and predicted labels of the classification model. Finally, the standard values corresponding to the 7 clusters are calculated according to the confusion matrix.

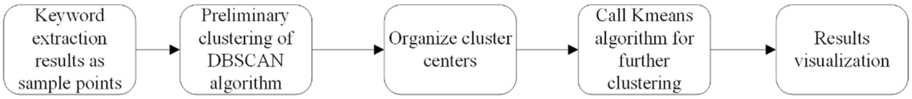
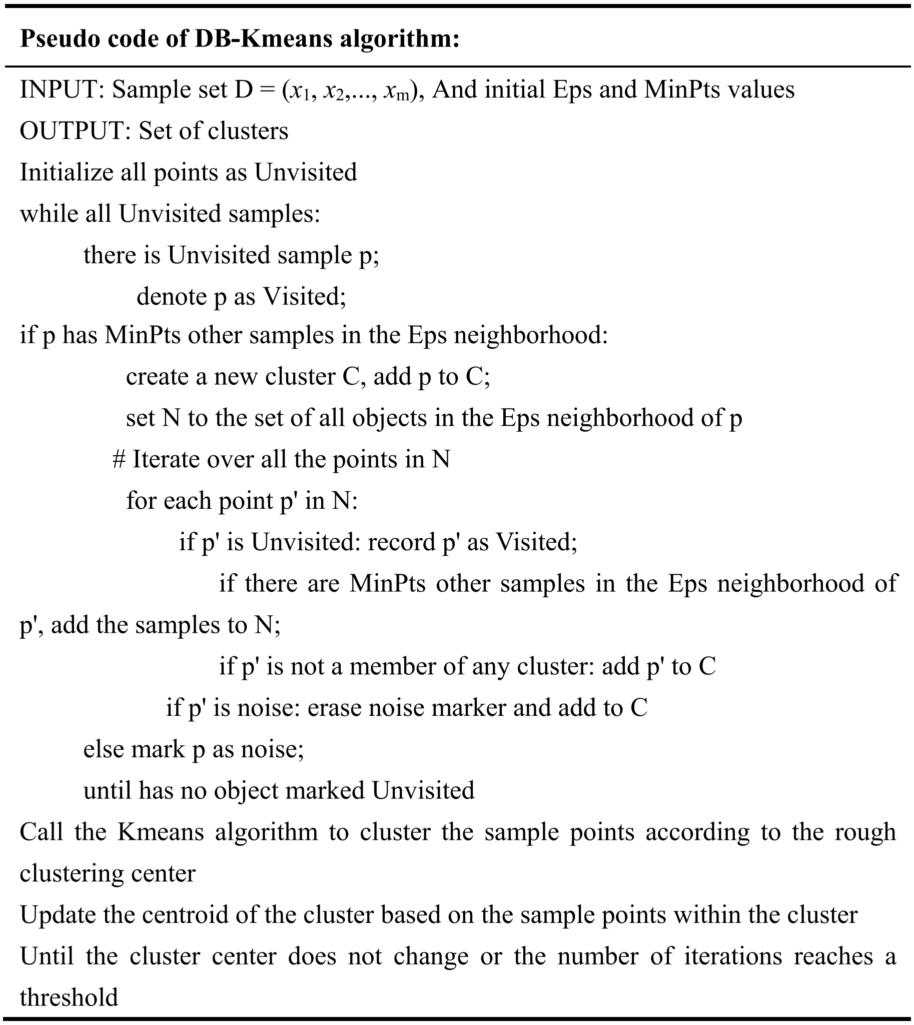

Keyword extraction results as sample points
Preliminary clustering of DBSCAN algorithm
Organize cluster centers
Call Kmeans algorithm for further clustering
Results visualization

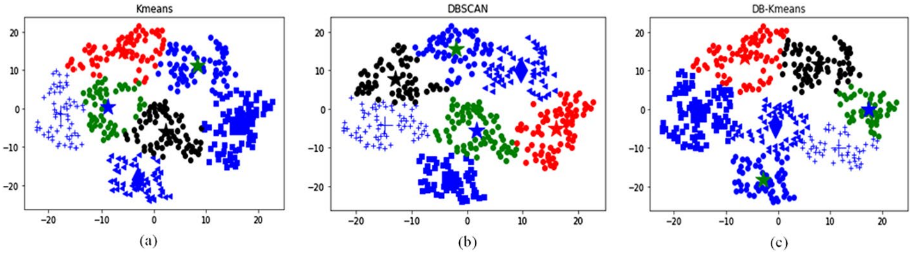
**Fig. 2** Flowchart of DB-Kmeans algorithm

**Table 2** Pseudocode of DB-Kmeans algorithm



**Pseudo code of DB-Kmeans algorithm:**

```
INPUT: Sample set D = ($x_1$, $x_2$,..., $x_m$), And initial Eps and MinPts values
OUTPUT: Set of clusters
Initialize all points as Unvisited
while all Unvisited samples:
        there is Unvisited sample p;
               denote p as Visited;
if p has MinPts other samples in the Eps neighborhood:
              create a new cluster C, add p to C;
              set N to the set of all objects in the Eps neighborhood of p
            # Iterate over all the points in N
              for each point p' in N:
                     if p' is Unvisited: record p' as Visited;
                          if there are MinPts other samples in the Eps neighborhood of p', add the samples to N;
                          if p' is not a member of any cluster: add p' to C
                   if p' is noise: erase noise marker and add to C
        else mark p as noise;
        until has no object marked Unvisited
Call the Kmeans algorithm to cluster the sample points according to the rough clustering center
Update the centroid of the cluster based on the sample points within the cluster
Until the cluster center does not change or the number of iterations reaches a threshold
```



(a) (b) (c)

**Fig. 3** Scatter plot

## 6 Evaluations and discussion

Clustering is an unsupervised learning process, but for the evaluation of clustering effect, we mark the effect of clustering manually and then use the evaluation indicators commonly used in machine learning classification models: Accuracy, Recall and $F1$ value ($H$-mean value) to evaluate the quality of the clustering effect.

**Table 3** Accuracy of each cluster

| | # 1 | # 2 | # 3 | # 4 | # 5 | # 6 | # 7 |
|---|---|---|---|---|---|---|---|
| AP | 89.56 | 91.78 | 89.56 | 93.56 | 88.44 | 90.67 | 92.89 |
| Meanshift | 68.33 | 68.75 | 62.75 | 50 | 73.33 | 70.89 | 81.48 |
| GMM | 92.67 | 94.44 | 91.78 | 94.89 | 90.22 | 92.89 | 94.67 |
| DBSCAN | 92.22 | 94.22 | 92.44 | 95.78 | 92.00 | 92.89 | 94.67 |
| Kmeans | 90.22 | 92.00 | 88.22 | 92.00 | 85.78 | 88.22 | 91.11 |
| DB-kmeans | **93.78** | **96.00** | **93.33** | **96.00** | **94.00** | **92.89** | **95.78** |

Bold values are the results of the method proposed in this paper

**Table 4** Recall of each cluster

| | # 1 | # 2 | # 3 | # 4 | # 5 | # 6 | # 7 |
|---|---|---|---|---|---|---|---|
| AP | 65 | 66.67 | 58.82 | 46.15 | 71.43 | 68.35 | 80.25 |
| Meanshift | 62.4 | 63.37 | 56.07 | 45.28 | 74.26 | 72 | 80.25 |
| GMM | 75 | 75 | 68.63 | 57.69 | 78.1 | 74.68 | 85.19 |
| DBSCAN | 73.33 | 72.92 | 72.55 | **73.08** | 80.00 | 75.95 | 83.95 |
| Kmeans | 55.00 | 58.33 | 56.86 | 46.15 | 67.62 | 67.09 | 75.31 |
| DB-kmeans | **78.33** | **83.33** | **76.47** | 61.54 | **82.86** | **79.75** | **88.89** |

Bold values are the results of the method proposed in this paper

- *Accuracy* Proportion of all predicted correct values to the total. The formula is (2).

$$\text{Accuracy} = \frac{\text{TP} + \text{FN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{2}$$

- *Recall rate* Recall rate, that is, the proportion of correct predictions that are positive to all actual positives. The formula is (3).

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

- *F1 value* the arithmetic mean divided by the geometric mean, the larger the better. The formula is (4).

$$F1 = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{4}$$

Among them, TP (True Positive) represents the prediction of the true positive class as a positive class; FP (False Positive) represents the prediction of the true negative class as a positive class; TN (True Negative) represents the true negative class is predicted as a negative class; FN (False Negative) represents the prediction of the true positive class as a negative class.

In order to highlight the superiority of the DB-Kmeans algorithm more clearly, this paper introduces several other commonly used algorithms based on the two comparison algorithms to analyze the same data samples. According to the calculation formulas of the three standards, the evaluation result tables are obtained as Tables 3, 4 and 5, respectively.

**Table 5** *F*1 of each cluster

| | # 1 | # 2 | # 3 | # 4 | # 5 | # 6 | # 7 |
|---|---|---|---|---|---|---|---|
| AP | 62.4 | 63.37 | 56.07 | 45.28 | 74.26 | 72 | 80.25 |
| Meanshift | 67.21 | 66 | 59.26 | 50 | 76.62 | 73.2 | 80.49 |
| GMM | 73.17 | 74.23 | 65.42 | 56.6 | 78.85 | 78.67 | 85.19 |
| DBSCAN | 71.54 | 72.92 | 68.52 | **66.67** | 82.35 | 78.95 | 85.00 |
| Kmeans | 60.00 | 60.87 | 52.25 | 40.00 | 68.93 | 66.67 | 75.31 |
| DB-kmeans | **77.05** | **81.63** | **72.22** | 64.00 | **86.57** | **79.75** | **88.34** |

Bold values are the results of the method proposed in this paper

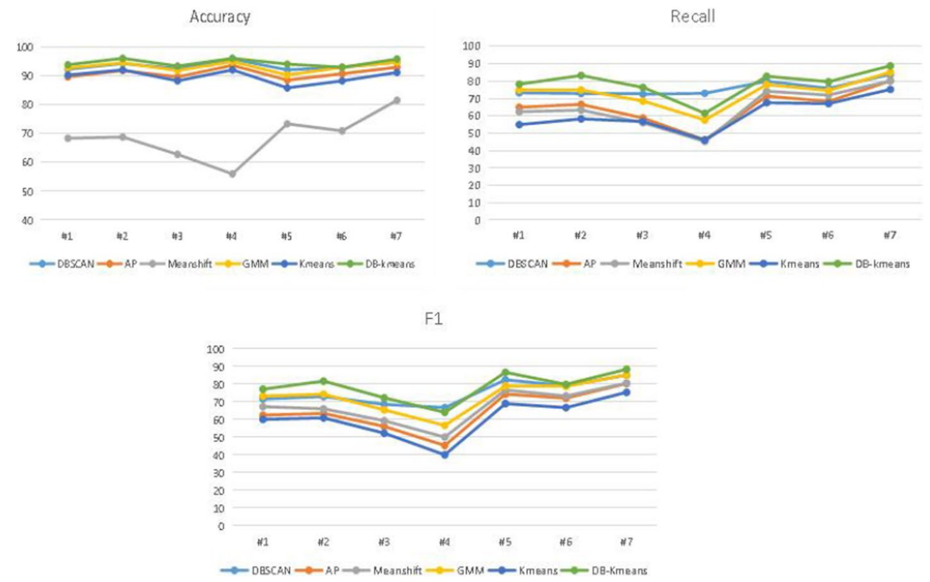


**Fig. 4** Line chart of results

By analyzing the table, we can find the Accuracy, Recall and the *F*1 values of DB-Kmeans are mostly the highest, which means that the proportion of words with correct clusters, the proportion of words predicted to be positive true values and the value of *F*1 are mostly the largest. The experiment result means that DB-Kmeans algorithm is the best in the traditional methods. In order to clarify the evaluation results, line charts are shown in Fig. 4.

From the analysis of the line charts, the three evaluation standard values of the DB-Kmeans algorithm are mostly the highest among these algorithms; besides, the accuracy and recall rate can basically reach more than 70%. Therefore, combined with the analysis of the advantages and disadvantages of the algorithms and the comparison of experimental data, this paper finally selects the DB-Kmeans algorithm with the best clustering results to study the influencing factors of the growth process of excellent teachers. Part of the results of the DB-Kmeans algorithm are shown in Table 6.

From the table above, we find that the words are clustered into 7 categories:

The first category is inclined to the spiritual attitude of excellent teachers, such as lifelong learning, teaching, educating people, passion, knowledge ideal and teaching students in accordance with aptitude. The second category can be explained to be the environmental factors, including the school factors and family factors, such as

**Table 6** Clustering results

| Category | Clustering words |
|---|---|
| 1 | Lifelong learning, teaching, and educating people, concentration, passion, self-cultivation, heart, achievement; knowledge ideal, moral character, subject knowledge, sense of humor, teaching students in accordance with their aptitude |
| 2 | Classroom, interest, teachers and students, class, grades, overall, sense of responsibility, authority, teachers, external environment, leadership, parents, moral education, capital, educational institutes, children, atmosphere, courseware, heart, construction, space, discussion, infection, teaching research, lectures |
| 3 | Excellent teacher, excellence, classmates, caring, group, praise, training, creativity, professor, opportunity, emotion, student career, trivia, special education, general education, understanding students, reform, communication, self-awareness, innovative, class hours, at school, brilliant, hard work |
| 4 | Experiment, class teacher, professional knowledge, school-based, experience, implementation, sense of responsibility, morality, common sense, performance, teaching and research activities, advanced characters, campus, creation, intelligence, communication platform, teaching and research staff, concept, observation, family education, school hours, preaching, educational experience, teaching materials, honorary title, guidance, quality education, old teachers |
| 5 | Student, advanced, dedication, character, observation, exploration, certificate, serious study, concept, credibility, reading, educational work, collaboration, employment guidance, active participation, competition, style, dedication, creation, computer, occasional event, nurturing, monitor, policy, key teachers, pioneer |
| 6 | Noble, reflection, ideas, strength, telling, cooperation, love, thinking, gaining, confidence, principal, patience, proficiency, situation, incumbency, sense of achievement, continuous learning, work hard dry, treatment, communication, perseverance, classroom atmosphere, scientific Research, teaching profession, competition, mathematical knowledge, curriculum standards |
| 7 | Profession, awareness, competition, career planning, inquiry, study, attitude, role model, subject, college, charm, classroom teaching, happiness, group, education, comment, friend, enthusiasm, superiority, language ability, morality, responsibility, correction weaknesses, competition, moral work, values, innovation, superior leadership, organizational management ability, management, tolerance |

classroom, external environment, leadership, parents and children. The third, fourth and fifth categories can be roughly distinguished as figures and events impact, professional ethics and work contents. For examples, excellent teacher, professor, special education can be regarded as the figure and events impact. Similarly, morality and sense of responsibility are in the scope of professional ethics. The words in the sixth category is mainly about self-awareness and introspection, such as reflection, ideas, thinking and sense of achievement. The seventh category can be identified as the professional knowledge and ability, such as the profession, classroom teaching, language ability and organizational management ability.

Basically, compared with the manual clustering results, the seven categories shown in Table 6 include the main influence factors of the excellent teachers' professional growth, such as the factors of professional spirit, knowledge and ability, self-awareness and introspection. These factors can be summarized as the inner factors of a person. In contrast, the environment factors, figures and events impact can be summarized as the outer factors of a person. These clustering and analysis provide a significant reference for the research of influencing factors on excellent teachers' professional growth.

## 7 Conclusions

In view of the shortcomings of Kmeans and DBSCAN algorithms, this paper proposes an improved DB-Kmeans algorithm and evaluates the clustering results through three evaluation criteria. Experiments show that the optimized algorithm improves the accuracy of keyword clustering results in the analysis of influencing factors of excellent teachers' professional growth through interview records. However, due to the specified scope of the research field in education, the amount of data prepared is relatively limited. With the increase in the amount of data, it is necessary to further test the DB-Kmeans algorithm whether can still maintain high-speed and effective calculation. Therefore, the next step of research is to apply this improved algorithm to a wider research field, to calculate a larger amount of data and to further verify the superiority.

### Abbreviations

| | |
|---|---|
| DBSCAN | Density-based spatial clustering of applications with noise |
| Eps | Epsilon |
| MinPts | Minimum points |
| DT-Kmeans | Decision Tree-Kmeans |
| FA | Firefly algorithm |
| HMM | Hidden Markov model |
| GMM | Gaussian mixture mode |
| ECG | Electrocardiogram |
| XGBoost | Extreme gradient boosting |
| TF-IDF | Term frequency-inverse document frequency |
| TP | True positive |
| FP | False positive |
| TN | True negative |
| FN | False negative |

### Acknowledgements

Not applicable.

### Author contributions

XG and XD proposed the framework of the whole ideal, structure of the model and the algorithm; TH helped to perform the simulations and conduct the analysis of the results. YK provided the interview records, participated in the conception and helped to revise the manuscript. All authors read and approved the final manuscript.

### Funding

This work was supported by the Scientific Research Project of Tianjin Educational Committee (2021KJ182).

### Availability of data and materials

Please contact the authors for data requests.

## Declarations

### Ethics approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Competing interests

The authors declare that they have no competing interests.

Received: 31 July 2022 Accepted: 14 November 2022
Published online: 05 December 2022

### References

1. Y. Wu, *Research on Passenger Car Passenger Order Scheduling Based on Canopy-Kmeans Algorithm* (Shanxi University, 2020)
2. X. Gao, *An Improved K-means Clustering Algorithm and a New Clustering Effectiveness Index Research* (Anhui University, 2020)
3. M. Yan, X. Xie, W. Li, D. Wu, X. Cui, S. Pan, Morphological clustering algorithm of typical load curve based on Gaussian kernel density estimation. Electr. Meas. Instrum. 1–8 (2022). https://kns.cnki.net/kcms/detail/detail.aspx?dbcode=CAPJ%26dbname=CAPJLAST%26filename=DCYQ20210316003%26uniplatform=NZKPT%26v=fzYhayA03xb9KgYHcfE22jsZ7B3eNVIqtoqr0ToI60YAoWnfgwuDsWQj7-MOLMZ
4. G. HimaBindu, Ch. Raghu Kumar, C. H. Hemanand, N. Rama Krishna, Hybrid clustering algorithm to process big data using firefly optimization mechanism. Mater. Today Proc. (2020). https://doi.org/10.1016/j.matpr.2020.10.273
5. N. Valarmathy, S. Krishnaveni, A novel method to enhance the performance evaluation of DBSCAN clustering algorithm using different distinguished metrics. Mater. Today Proc. (2020). https://doi.org/10.1016/j.matpr.2020.09.623

6. G. Manogaran, V. Vijayakumar, R. Varatharajan, P.M. Kumar, R. Sundarasekar, C.-H. Hsu, Machine learning based big data processing framework for cancer diagnosis using hidden Markov model and GM clustering. Wirel. Pers. Commun. **102**(3), 2099–2116 (2018)
7. W. Jia, Y. Tan, L. Liu, J. Li, H. Zhang, K. Zhao, Hierarchical prediction based on two-level Gaussian mixture model clustering for bike-sharing system. Knowl.-Based Syst. **178**, 84–97 (2019)
8. J. Zhong, D. Hai, J. Cheng, C. Jiao, S. Gou, Y. Liu, H. Zhou, W. Zhu, Convolutional autoencoding and Gaussian mixture clustering for unsupervised beat-to-beat heart rate estimation of electrocardiograms from wearable sensors. Sensors **21**(21), 7163 (2021)
9. Y. Shi, S. Yan, M. He, X. Li, Hybrid data mining method of telecom customer based on improved Kmeans and XGBoost. J. Phys. Conf. Ser. **2010**(1), 120 (2021)
10. T. Li, Research on patent text clustering based on improved k-means algorithm. Hebei University of Engineering (2020)
11. J. Xiaoyun, Ru. Zheng, C. Jingxia, An EEG emotion recognition method based on multi-feature extraction. J. Shaanxi Univ. Sci. Technol. **36**(05), 152–158 (2018)

## Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.