Research Article

International Journal of Advanced Computer Research, Vol 13(64)
ISSN (Print): 2249-7277 ISSN (Online): 2277-7970
http://dx.doi.org/10.19101/IJACR.2023.1362018

Enhancing data analysis through k-means with foggy centroid selection

Arun Sharma*, Surendra Vishwakarma and Animesh Kumar Dubey
Department of Computer Science, Patel College and Science and Technology, Bhopal, India

Received: 14-March-2023; Revised: 11-August-2023; Accepted: 16-August-2023

Abstract
An innovative approach, k-means with foggy centroid selection (KFCS) was proposed, for enhancing data clustering performance. This study focuses on the application of this method to the Pima Indians diabetes database, serving as a comprehensive evaluation ground. The process begins with preprocessing and data arrangement, involving scaling and normalization to ensure accurate computation. KFCS, combines k-means clustering with foggy centroid selection, utilizing both random initialization and iterative centroid calculation. The approach hinges on four distance algorithms – Euclidean, Pearson Coefficient, Chebyshev, and Canberra – to gauge similarity. A detailed exploration of distance estimation enhances dataset understanding. Through rigorous evaluation, KFCS demonstrates superiority in terms of computation time and error analysis, with Canberra algorithm emerging as a standout performer. This work contributes a comprehensive methodology for improved data clustering and analysis.

Keywords
K-means, Euclidean, Pearson coefficient, Chebyshev and Canberra.

1.Introduction
Clustering algorithms have emerged as indispensable tools across diverse sectors such as engineering, e-commerce, and healthcare in the contemporary era. Among the prominently utilized methods, k-means, fuzzy c-means, and hierarchical clustering have garnered substantial attention [1, 2]. The effectiveness of these algorithms hinges on their ability to discern data clusters through inter-centroid calculations [2].

Within the context of specific needs, a range of grouping methodologies has been established, typically falling into two or three distinct approaches [3]. These methods encompass techniques like partitioning and progressive strategies [3]. The terminologies used to describe these methods can be tailored to suit contexts [4]. The quality of clustering outcomes is intimately linked to characteristics of centroids, initialization techniques, divergence metrics, and central focus points [1, 2].

In the contemporary landscape characterized by escalating database sizes, the notion of data pruning has surfaced as an optimal solution within the realm of data mining [5].

*Author for correspondence

Data pruning plays a pivotal role in data mining algorithms, greatly facilitating the identification of meaningful patterns. In the pursuit of knowledge exploration, both data mining and knowledge discovery in databases (KDD) hold pivotal roles [6-9]. The DM approach, which entails the autonomous discovery of recurrent patterns within data, serves as the nucleus of the learning and discovery process. This methodological framework encompasses a series of stages, including data selection, preprocessing, transformation, actual data mining, as well as the interpretation and evaluation of instances. Notably, experts have advocated for tailoring the data mining process according to the specifics of the dataset being studied [10-12]. One notable pursuit within the realm of knowledge discovery is high-utility data mining.

Traditional data mining techniques have gained significant traction in tasks that involve distance estimation [13]. While continuous pattern mining helps, it is essential to recognize that each item carries unique significance and often appears individually in transactions [13−17]. The challenge of high-utility pattern mining addresses this issue by acknowledging that each item may bear a weight that encapsulates valuable information, enhancing the search process [18−20]. A multitude of applications can leverage the extraction of high utility itemsets from databases that prioritize value-oriented information. These applications encompass a wide

array of use cases such as market basket analysis, clickstream analysis, and other mission-critical scenarios.

The selection and application of specific algorithms depend on the unique requirements of each domain. The ongoing growth of databases has necessitated solutions like data pruning to enable efficient data mining and pattern discovery. Meanwhile, the integration of data mining and knowledge discovery serves as a cornerstone for acquiring valuable insights from intricate datasets. High-utility data mining and innovative approaches to handling distinct patterns further enhance the applicability of data mining techniques in real-world scenarios.

The main objective of the study is to introduce and assess the effectiveness of a novel approach called "k-means with foggy centroid selection (KFCS)" for enhancing the performance of data clustering. The study specifically targets the Pima Indians diabetes database, aiming to comprehensively evaluate the proposed method's capabilities.

2.Literature survey

In 2019, Reddy et al. [21] addressed efficient large data handling, cost-effectiveness, and data loss challenges when data belongs to multiple clusters, particularly in healthcare big data. They proposed a fuzzy c-means algorithm utilizing midpoint to mitigate data loss.

In 2019, Vanitha et al. [22] explored tech's role in agriculture. They advocated Python as a front end for agri-data analysis, using Jupyter for crop prediction via k-means, k-nearest neighbors (KNN), support vector machine (SVM), and Bayesian network algorithms.

In 2019, Dai and Sheng [23] examined evolutionary algorithms in clustering. They proposed a multi-objective clustering ensemble algorithm with automatic k-determination, addressing issues of parameter requirement and diversity preservation.

In 2021, Anishfathima et al. [24] proposed a novel method for predicting type 2 diabetes using data mining techniques. Their model, incorporating enhanced SMO with random forest (RF) estimation, achieved 3.04% higher accuracy than previous studies. The approach holds promise for effective diabetes management and was tested on multiple datasets with favorable outcomes.

In 2022, Salsabila et al. [25] proposed an indoor positioning system (IPS) for tracking individuals using Wi-Fi signal clusterization via k-means algorithm. The approach achieved an average 77% accuracy in estimating positions, replacing global positioning system (GPS) in confined areas.

In 2022, DiAdamo et al. [26] employed quantum computers for unsupervised clustering, addressing practical predictive maintenance in energy operations. Quantum k-means clustering was proposed, improving accuracy by 67.8% compared to classical algorithms in real-world energy grid scenarios, overcoming hardware challenges.

In 2023, Jeyachidra et al. [27] introduced a novel k-means-artificial neural network (ANN) technique to enhance stability and sustainability of power systems amidst communication network congestion, attacks, and data issues. The approach utilizes k-means for outlier detection, ANN for missing data imputation, and has been validated through simulations and real-world data.

In 2023, Chusyairi et al. [28] addressed mysterious hepatitis cases in Indonesia. They proposed an intelligent hybrid technique integrating enhanced k-means clustering and ensemble learning for disease diagnosis. Notably, they emphasized the importance of features like age, bilirubin, alk_phosphate, sgot, albumin, and protime in their analysis, utilizing primary data sourced from kaggle.com.

In 2023, Siridhara et al. [29] responded to the growing demand for high-quality food by automating fruit and vegetable defect detection using image processing techniques like k-means clustering and Otsu's thresholding. This approach aims to reduce human errors and accelerate the detection process for imperfections, addressing the agricultural industry's need for efficiency.

In 2023, Hou et al. [30] introduced the dynamic weighted density clustering algorithm combined with k-means (DWDC-k-means) algorithm for enhancing user behavior analysis in substation areas. The method combines dynamic weighted density clustering and k-means. It optimizes cluster initialization and fusion-based k-means clustering, outperforming traditional methods in accuracy and efficiency, as demonstrated with smart grid data.

3.Proposed work

In this paper KFCS was proposed for the efficient cluster selection. The approach component has been discussed below:

Dataset discussion

The analysis of the Pima Indians diabetes database was delved into for experimentation. This dataset was the target of our approach, facilitating a comprehensive assessment of its effectiveness.

Preprocessing and data arrangement

The first phase involved preprocessing and arranging the data. To achieve meaningful data clustering and arrangement for refined clusters, the data was scaled according to an algometric approach. The data arrangement centered on patient attribute values, with normalization being employed to ensure meaningful computational processes.

KFCS

KFCS stands at the core of our algorithm. The K-means algorithm was employed for clustering and centroid calculation, making use of both random initialization and iterative centroids. This procedure was conducted on the preprocessed dataset values, effectively harnessing their potential.

The calculation of similarity scores was tackled in our paper through the utilization of four distinct distance algorithms: Euclidean, Pearson Coefficient, Chebyshev, and Canberra. These algorithms played a crucial role in achieving comprehensive similarity matching, thereby enhancing the accuracy of our approach.

The exploration of distance estimation was undertaken as a critical component of our data analysis efforts. Within this section, the calculation of distances between data points was encompassed, contributing to a deeper understanding of the structure inherent within the dataset.

The foundation for our analysis and experimentation was provided by the Pima Indians diabetes database. The preprocessing and data arrangement phase played a crucial role in this process, ensuring that the data was optimally prepared for the subsequent stages. The arrangement of data based on an algometric scale was undertaken to bolster clustering outcomes and streamline the identification of refined clusters. Additionally, the application of normalization further facilitated the effective harnessing of the data's potential.

Our algorithm selection centered on K-means, which featured both random initialization and iterative centroids. This choice ensured the attainment of robust clustering and the accurate determination of centroids. By operating on preprocessed data values, a higher degree of accuracy in results was achieved.

In the evaluation of our approach's efficacy, a range of similarity matching measures were employed, including Euclidean, Pearson Coefficient, Chebyshev, and Canberra. This multifaceted approach played a pivotal role in augmenting the precision of our results, enabling a more comprehensive understanding of the intricate relationships within the dataset.

Figure 1 presents the complete representation of our methodology. The flowchart encapsulates the entire procedure.

Algorithm: K-means with Foggy Centroid Selection (KFCS)
Input:
Dataset D with n data points and m features
Number of clusters k
Maximum number of iterations max_iterations
Threshold epsilon for convergence
Output:
Cluster assignments for each data point
Centroid values for each cluster

Steps:

1. Initialization:
   Initialize k centroids randomly from the dataset.
2. Iteration:
   For i in 1 to max_iterations:
      Initialize k empty clusters.
      For each data point x in the dataset D:
         Calculate the distance between x and each centroid.
         Assign x to the nearest centroid's cluster.
3. Centroid Update:
   For each cluster c:
      Calculate the new centroid by taking the mean of all data points in the cluster.
4. Foggy Centroid Selection:
   For each cluster c:
      Calculate the foggy centroid by selecting a point randomly from the cluster with a probability inversely proportional to its distance from the actual centroid. This step introduces randomness and exploration.
5. Check Convergence:

Calculate the difference between the old and new centroids.

If the difference is less than epsilon for all clusters, terminate the algorithm.

Output:

Return the cluster assignments for each data point and the updated centroid values.

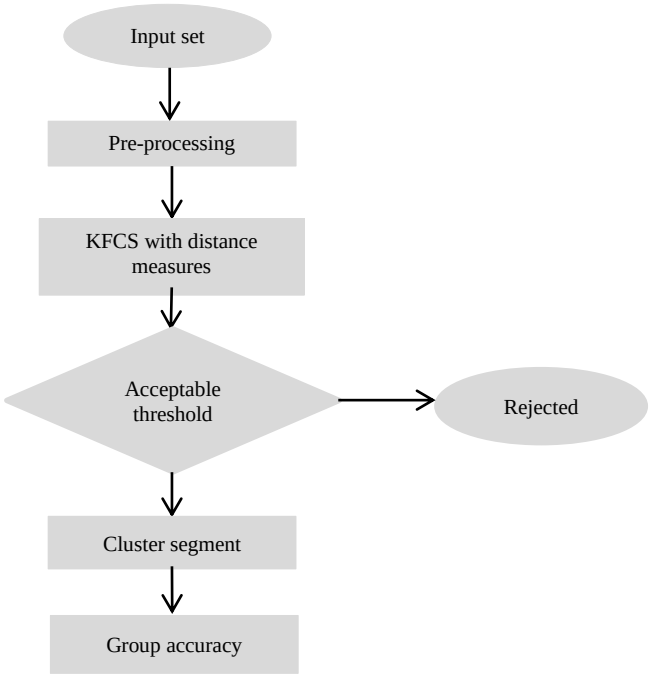

Figure 1 Complete steps of the proposed work

4.Results and discussion

Three samples have been chosen to compare and evaluate the results. In *Figure 2*, a time-based analysis is presented, showcasing various distance algorithms. It is observed that the Pearson (P) algorithm requires more computation time. Conversely, the Chebyshev (Ch) and Canberra (Ca) algorithms exhibit superior performance in terms of time computation when compared to the other methods. Additionally, the Euclidean (E) and Pearson algorithms exhibit similar computation times. In *Figure 3*, an error analysis is depicted, encompassing diverse distance algorithms. Notably, the Pearson algorithm exhibits higher clustering similarity errors compared to all others. Among the algorithms, Canberra emerges as the optimal choice, displaying the lowest error rates and thus standing out as the most effective option. A complete list of abbreviations is shown in *Appendix I.*

5.Conclusion

In this paper, we presented the KFCS approach for efficient data clustering, applied to the Pima Indians diabetes database. Through a systematic exploration, we showcased the efficacy of our method. Preprocessing and data arrangement set the foundation for optimal clustering by scaling and normalizing data. KFCS, featuring foggy centroid selection, exhibited robustness in calculating centroids, leading to accurate clusters. The use of four distance algorithms enhanced similarity measurement, deepening dataset understanding. Our results indicated that the Pearson algorithm consumed more computation time, while Chebyshev and Canberra algorithms excelled. In terms of error analysis, Canberra proved to be the most effective choice. Overall, our approach demonstrates promise in refining data clustering accuracy and offers a valuable contribution to the field of data analysis.

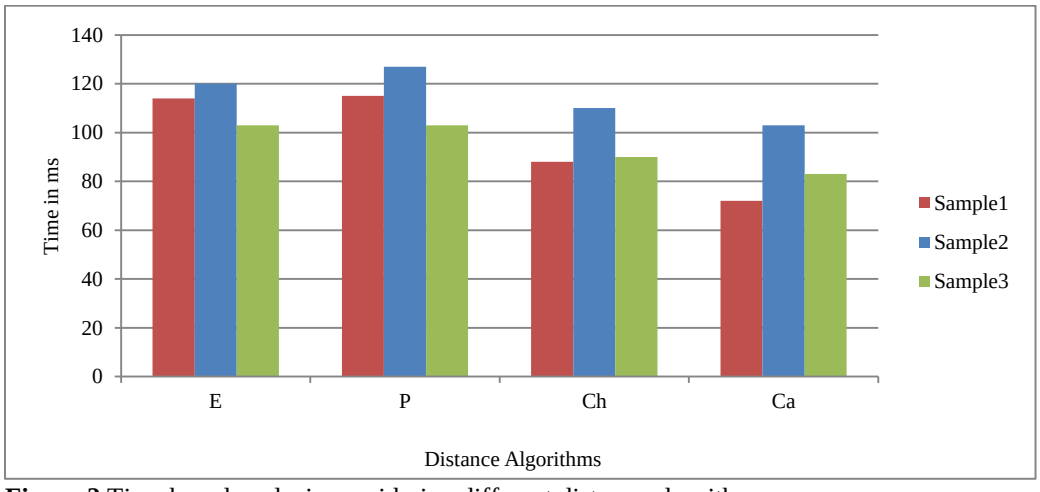

Figure 2 Time based analysis considering different distance algorithms

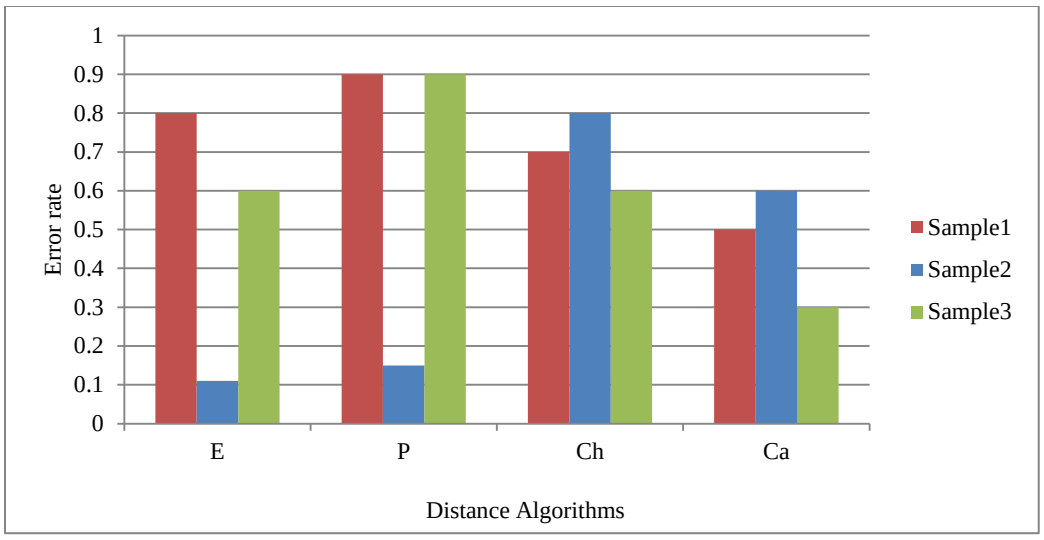

Figure 3 Error analysis considering different distance algorithms

Acknowledgment
None.

Conflicts of interest
The authors have no conflicts of interest to declare.

References
[1] Dubey AK, Gupta U, Jain S. Analysis of k-means clustering approach on the breast cancer Wisconsin dataset. International Journal of Computer Assisted Radiology and Surgery. 2016; 11:2033-47.
[2] Dubey AK, Dubey AK, Agarwal V, Khandagre Y. Knowledge discovery with a subset-superset approach for mining heterogeneous data with dynamic support. In CSI sixth international conference on software engineering (CONSEG) 2012 (pp. 1-6). IEEE.
[3] Jiang R, Han S, Yu Y, Ding W. An access control model for medical big data based on clustering and risk. Information Sciences. 2023; 621:691-707.
[4] Alizadehsani R, Roshanzamir M, Izadi NH, Gravina R, Kabir HD, Nahavandi D, et al. Swarm intelligence in internet of medical things: a review. Sensors. 2023; 23(3):1466.
[5] Dubey AK, Gupta U, Jain S. Comparative study of K-means and fuzzy C-means algorithms on the breast cancer data. International Journal on Advanced Science, Engineering and Information Technology. 2018; 8(1):18-29.
[6] Fernández-de-Las-Peñas C, Martín-Guerrero JD, Florencio LL, Navarro-Pardo E, Rodríguez-Jiménez J, Torres-Macho J, et al. Clustering analysis reveals different profiles associating long-term post-COVID symptoms, COVID-19 symptoms at hospital admission and previous medical co-morbidities in previously hospitalized COVID-19 survivors. Infection. 2023; 51(1):61-9.
[7] Sangaiah AK, Rezaei S, Javadpour A, Zhang W. Explainable AI in big data intelligence of community detection for digitalization e-healthcare services. Applied Soft Computing. 2023; 136:110119.
[8] Liu B, Li X, Wang H, Zhao S, Li J, Qu G, Wang F. Retyping of triple-negative breast cancer based on clustering method. Expert Systems. 2023; 40(2):e12583.
[9] Setiawan KE, Kurniawan A, Chowanda A, Suhartono D. Clustering models for hospitals in Jakarta using fuzzy c-means and k-means. Procedia Computer Science. 2023; 216:356-63.
[10] Das RK, Shandilya M. Clustering based ACO and ABC algorithms for the shadow detection and removal. International Journal of Advanced Technology and Engineering Exploration. 2022; 9(91):839-53.
[11] Rao NT, Satyanarayana KV, Satyanarayana M, Joshua ES, Bhattacharyya D. Breast cancer classification using improved fuzzy C-Means algorithm. In smart technologies in data science and communication: proceedings of SMART-DSC 2022 2023 (pp. 197-204). Singapore: Springer Nature Singapore.
[12] Dubey AK, Sinhal AK, Sharma R. An improved auto categorical PSO with ML for heart disease prediction. Engineering, Technology & Applied Science Research. 2022; 12(3):8567-73.
[13] Kumar M, Dubey AK. An analysis and literature review of algorithms for frequent itemset mining. International Journal of Advanced Computer Research. 2023; 13(62):1-7.
[14] Ilango SS, Vimal S, Kaliappan M, Subbulakshmi P. Optimization using artificial bee colony based clustering approach for big data. Cluster Computing. 2019; 22:12169-77.
[15] Ladha GG, Pippal RK. An efficient distance estimation and centroid selection based on k-means clustering for small and large dataset. International Journal of Advanced Technology and Engineering Exploration. 2020; 7(73):234-40.

[16] Dubey A, Gupta U, Jain S. Medical data clustering and classification using TLBO and machine learning algorithms. Computers, Materials and Continua. 2021; 70(3):4523-43.
[17] Dubey AK, Shandilya SK. A comprehensive survey of grid computing mechanism in J2ME for effective mobile computing techniques. In 5th international conference on industrial and information systems 2010 (pp. 207-12). IEEE.
[18] Muqtadiroh FA, Usagawa T, Rachmayanti RD, Nugroho SM, Yuniarno EM, Purnomo MH. Rules determination based on time-series data to classify unsupervised cases based on fuzzy expert system. International Journal of Intelligent Engineering and Systems. 2023; 16(3):258-68.
[19] Liu J, Peng B, Yin Z. A hybrid machine learning method for diabetes detection based on unsupervised clustering. In proceedings of the 2023 7th international conference on machine learning and soft computing 2023 (pp. 144-9).
[20] Vatresia A, Johar A. Fuzzy subtractive C-means for teacher distribution analysis. In mathematics and science education international seminar 2021 (MASEIS 2021) 2023 (pp. 233-44). Atlantis Press.
[21] Reddy BR, Kumar YV, Prabhakar M. Clustering large amounts of healthcare datasets using fuzzy c-means algorithm. In 5th international conference on advanced computing & communication systems 2019 (pp. 93-97). IEEE.
[22] Vanitha CN, Archana N, Sowmiya R. Agriculture analysis using data mining and machine learning techniques. In 5th international conference on advanced computing & communication systems 2019 (pp. 984-90). IEEE.
[23] Dai H, Sheng W. A multi-objective clustering ensemble algorithm with automatic k-determination. In 4th international conference on cloud computing and big data analysis 2019 (pp. 333-7). IEEE.
[24] Anishfathima B, Gautham P, Mahalakshmi BG, Jamadar SJ. Smart architecture for diabetic patients using machine learning. In 7th international conference on advanced computing and communication systems 2021 (pp. 1544-8). IEEE.
[25] Salsabila SS, Kristalina P, Santoso T. The implementation of optimal k-means clustering for indoor moving object localization. In international electronics symposium 2022 (pp. 210-5). IEEE.
[26] DiAdamo S, O'Meara C, Cortiana G, Bernabé-Moreno J. Practical quantum K-Means clustering: performance analysis and applications in energy grid classification. IEEE Transactions on Quantum Engineering. 2022; 3:1-6.
[27] Jeyachidra J, Logesh T, Nandhini K, Krithiga R. Hybrid K-Means clustering for training special children using utility pattern mining. In international conference on artificial intelligence and knowledge discovery in concurrent engineering 2023 (pp. 1-7). IEEE.
[28] Chusyairi A, Nurdiawan O, Sambath K, Hayat RN, Wijaya YA. Hepatitis cluster model with K-means algorithm. In international conference on computer science, information technology and engineering 2023 (pp. 811-5). IEEE.
[29] Siridhara AL, Manikanta KV, Yadav D, Varun P, Saragada J. Defect detection in fruits and vegetables using K Means segmentation and Otsu's thresholding. In international conference on networking and communications 2023 (pp. 1-5). IEEE.
[30] Hou Y, Lu H, Cao N, Wei Z. User behavior analysis of substation area based on improved K-means quadratic clustering algorithm. In 5th international conference on intelligent control, measurement and signal processing 2023 (pp. 1223-7). IEEE.

Arun Sharma is currently pursuing an M.Tech in Computer Science at PSCT, RGPV in Bhopal, Madhya Pradesh. He has completed his MCA program at RGPV Technical University in Bhopal, MP. His areas of interest include Data Mining, Optimization, Machine Learning, and Artificial Intelligence.
Email: arunsharma31087@gmail.com

Surendra Vishwakarma is working as Associate Professor in the Department of Computer Science and Engineering, at Patel College of Science and Technology, Bhopal, India.
More than 12 Publications in Reputed, Peer-reviewed National and International Journals and Conferences in the Research areas - Data Mining, Machine Learning, Cloud Computing and Artificial Intelligence**.**
Email:s.vish83@gmail.com

Animesh Kumar Dubey works as an Assistant Professor in the Department of Computer Science and Engineering at Patel College of Science and Technology in Bhopal, India. He holds a Bachelor of Engineering (B.E.) and an M.Tech. degree in Computer Science Engineering from Rajeev Gandhi Technical University in Bhopal, Madhya Pradesh. He has authored more than 15 publications in reputable, peer-reviewed national and international journals and conferences. His research interests encompass Data Mining, Optimization, Machine Learning, Cloud Computing, and Artificial Intelligence.
Email: animeshdubey123@gmail.com

Appendix I

S. No.	Abbreviation	Description
1	ANN	artificial neural network
2	DWDC	dynamic weighted density clustering algorithm
3	GPS	Global Positioning System
4	IPS	Indoor Positioning System
5	KFCS	k-Means With Foggy Centroid Selection
6	KDD	Knowledge Discovery In Databases
7	KNN	K-Nearest Neighbors
8	RF	Random Forest
9	SVM	Support Vector Machine