Journal of Harbin Engineering University
ISSN: 1006-7043

Vol 44 No. 10
October 2023

E-commerce Personalization Revolutionized by Birch and FCM Clustering

Shyamala Devi J.[1], Anoop M.[2] *, Dr. Allan J. Wilson[3]

[1]Assistant professor, Department of computer science and applications, SRM institute of science and technology, Ramapuram, chennai-89

[2]Assistant professor, Department of computer science and applications, SRM institute of science and technology, Ramapuram, chennai-89

[3]Assistant Professor, Department of Electronics and Communication Engineering, Amrita College of Engineering and Technology

Abstract:

Understanding customer behaviour in the fast-paced e-commerce environment of today is essential for online success. E-commerce platforms collect enormous amounts of clickstream data, which may be used to reveal consumer preferences and improve the online shopping experience. In order to segment and categories clients based on their clickstream data, this study offers a novel approach (BFC-HT) that combines Birch clustering, Fuzzy C-Means (FCM) clustering, and dynamic hyperparameter tuning via Grid Search and Random Search. In a highly competitive online world, these analytics enable e-commerce businesses to personalise marketing initiatives, improve user interactions, and eventually promote increased consumer happiness and conversions. In doing so, our research significantly advances the rapidly developing fields of e-commerce analytics and customer segmentation by providing a solid framework for extracting useful information from the vast amount of available clickstream data.

***Keywords**:* E-commerce Clickstream data, Birch clustering, Fuzzy C-Means clustering, Hyperparameter tuning, Grid Search, Random Search

1. Introduction

Internet store Recent years have seen a significant amount of research into clickstream data, which provides important insights on user behaviour, preferences, and the nuances of online buying experiences. We can observe the development of our knowledge of clickstream data and its implications for e-commerce enterprises by drawing on earlier research papers. Early research in this field frequently concentrated on fundamental clickstream analysis with the goal of understanding user navigation patterns. These studies provided the groundwork for appreciating clickstream data's relevance as a valuable source of knowledge. In order to understand the paths users take, the sites they view, and the length of their trips, researchers looked into the sequential nature of user interactions.

As research progressed, focus switched to the use of machine learning and data mining tools to uncover more profound discoveries. Birch and Fuzzy C-Means, among other clustering algorithms, have become effective tools for segmenting users based on their clickstream behaviour. By grouping consumers into discrete groups with common traits, researchers were able to develop targeted marketing campaigns and individualised recommendations. As the efficiency of clustering algorithms heavily depended on fine-tuning parameters like the number of clusters and the fuzziness exponent, hyperparameter tuning emerged as a key component of this research. An important step was taken with the introduction of Grid Search and Random Search as hyperparameter optimisation techniques, which improved clustering accuracy and made it possible to identify complex user segments.

The practical ramifications of clickstream analysis for e-commerce enterprises were also covered in research articles. They looked at how user segmentation based on behaviour could increase customer happiness, boost sales, and make the most of advertising dollars. Additionally, they shed light on the difficulties associated with handling clickstream data while maintaining data privacy and security, particularly in light of changing rules. In hindsight, these earlier research publications served as the foundation for current e-commerce analytics techniques. They highlighted the potential of clickstream data to drive business growth and underscored the importance of robust clustering algorithms like Birch and Fuzzy C-Means. The integration of hyperparameter tuning techniques like Grid Search and Random Search showcased the adaptability of these approaches to evolving data analysis needs.

In sum, the historical research on e-commerce website clickstream data offers a valuable perspective on how the field has evolved from basic clickstream analysis to sophisticated clustering and hyperparameter optimization techniques. These insights continue to shape modern e-commerce analytics practices and have profound implications for businesses seeking to harness the power of user behavior data.

2. Related Work

Data samples are grouped during clustering according to some sort of similarity metric. The literature [1][3] lists numerous clustering techniques in great detail. The centroid-based clustering techniques K-Means [4], hierarchical clustering [1], spectral clustering [5], and Gaussian mixture model [6] are the most often used.Multiple weak clustering methods are used as the foundation for ensemble clustering techniques [4]. The authors of [6] suggested a hybrid system to predict timeseries problems combining neural trees, genetic programming, and particle swarm optimisation. To anticipate time-series problems, the authors of [7] suggested a similar hybrid system utilising RBF neural networks and differential evolution algorithm.

For forecasting student grades using prior data and student behaviour, a hybrid method based on fuzzy C-Means clustering and multi-variable regression is proposed in [8]. In comparison to other prediction systems, the suggested approach produces better predictions. [9] describes a hybrid fuzzy picture segmentation method that relies on both the fuzzy C-Means algorithm and graph cut theory. The image is first divided into small sections, and from each of these regions, colour histogram features are retrieved and clustered using the fuzzy C-Means algorithm.

An application of Self-Organizing Feature Map Neural Network based on K-means clustering in Network Intrusion Detection was introduced by Tan et al. (2019) [14]. The NSL-KDD network intrusion detection database provided the data set for this study. The clustering model increased pattern recognition and greatly decreased training time, according to experimental results. Supervised Self-organizing Maps (SuSi) is a framework that Riese et al. (2020) [12] presented. On high-dimensional data, supervised and semi-supervised classification are used. Evaluation based on information on soil moisture data demonstrated that our Deep Learning model produced positive outcomes and outperformed the random forest in the moisture in the soil declining. A Deep Learning model that may be utilised for Predicting Movies Rates Category was proposed by Nasser et al. (2019) [10].

Based on a power metre, Zhang et al. (2020) [13] proposed a method to determine the kind and condition of electric equipment. time elapsed. To determine the kind of appliance, a convolutional neural network was trained. A k-means method was used to determine how many states the appliance had. The findings revealed that this model had a considerable impact. in increasing the precision with which electrical appliances' types and states of operation are recognised. A Deep Learning model called HKM-ANN was recently suggested by Nguyen Nasser et al. (2020) [11] for predicting Exploiting a Hybrid Model Based on Clustering and Artificial Neural Networks to Reduce Blast-Induced Ground Vibration in an Open-Pit Mine Network neural. Prior to instructing the prediction model, a Hierarchical K-Means clustering technique (HKM) models.

3. Problem Definition

Modern digital marketing and analytics are grappling with the issue of user segmentation and personalised suggestions using clickstream data from e-commerce websites. Starting with the crucial stage of data gathering and preprocessing, it comprises a variety of subproblems. To derive actionable insights, clickstream data must be of a high standard and consistency. The next step is feature engineering, where scientists and researchers must imaginatively pull pertinent elements from the data to enable efficient clustering. It is equally important to assess how well the clustering model performs and the resulting user segments, as this affects how successful the strategy is. Using these user segments for personalised suggestions and targeted marketing directly affects a business's capacity to engage customers and encourage sales.The size and complexity of clickstream data might make data processing and storage difficult. Additionally, the delicate nature of user behaviour data prompts worries regarding data security and privacy. Complying with changing legislation and making adjustments for altering user behaviour over time complicate the process even more. However, businesses and researchers may fully utilise the potential of clickstream data to adapt their marketing efforts and improve the online purchasing experiences of their clients by solving these issues and utilising the learnings from prior research. The knowledge and use of clickstream data are constantly evolving, which emphasises their importance in the digital age and how they are affecting the strategies of e-commerce companies all over the world.

Proposal

- To segment e-commerce customers into distinct groups based on their clickstream behavior using Birch and Fuzzy C-Means clustering.
- To uncover hidden patterns and behaviors within the clickstream data using Fuzzy C-Means clustering and fine-tuning hyperparameters.
- To use the clustering results and hyperparameter tuning to optimize marketing campaigns, including targeting specific customer segments with relevant offers and promotions.

4. Process Flow

The process begins with Birch clustering, which serves as both a dimensionality reduction technique and a means to identify initial subclusters. Subsequently, FCM clustering takes center stage, operating on these subclusters to reveal latent patterns in customer behavior. However, achieving the finest clustering outcomes requires meticulous fine-tuning of pivotal hyperparameters, specifically the number of subclusters within Birch and the count of clusters (K) within FCM.To accomplish this fine-tuning, hyperparameter optimization unfolds through a harmonious blend of Grid Search and Random Search. Grid Search methodically explores predefined hyperparameter values within established boundaries, providing structure to the search process. Conversely, Random Search injects an element of randomness, further refining the parameters. This iterative amalgamation dynamically adapts the search strategy, pinpointing the most suitable hyperparameters and elevating the precision of clustering.

4.1. Birch Clustering:

In the realm of e-commerce websites, the analysis of clickstream data using BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies) clustering emerges as a valuable tool for extracting meaningful insights. Clickstream data, which records user interactions on a website in chronological order, contains a wealth of knowledge on user behaviour, preferences, and engagement patterns. It is feasible to hierarchically group user sessions or behaviours into clusters by using BIRCH clustering on this data, exposing hidden patterns and different user segments. This method helps in comprehending visitor navigational patterns, spotting trends, and improving the user experience. BIRCH is an effective option for e-commerce companies looking to improve customization, streamline marketing methods, and boost overall website performance because of its capacity to handle massive volumes of data rapidly while offering a hierarchical representation.Birch clustering, which uses preprocessing to identify subclusters within the clickstream data, first gives an overview of various user behavior patterns. The following formula results in this:

$$S = \sqrt{N/2}$$

Where 'S' represents the number of subclusters, and 'N' signifies the number of data points. Subsequently, Fuzzy C-Means (FCM) clustering is employed within each subcluster. FCM assigns data points to clusters based on degrees of membership using the following formula:

$$\mu(i, j) = 1 / \sum(k=1 \text{ to } c) [(||x(i) - c(j)|| / ||x(i) - c(k)||)^{(2/(m-1))}]$$

Where 'μ(i, j)' represents the degree of membership of data point 'i' to cluster 'j', 'x(i)' is the data point, 'c(j)' is the cluster center, 'c(k)' represents other cluster centers, and 'm' is a fuzziness parameter. Hyperparameter tuning is essential to optimize the number of subclusters in Birch and the number of clusters 'K' in FCM, ensuring that clustering outcomes are accurate and insightful.

4.2. Hyperparameter Tuning:

Fine-tuning the hyperparameters is crucial for achieving the best clustering outcomes. The key hyperparameters that need to be optimized are:

4.3. Number of subclusters within Birch.

BIRCH is often used as a preprocessing step before applying more traditional clustering methods. One of its main hyperparameters is the number of subclusters to create during the initial clustering phase. This hyperparameter determines how finely or coarsely the data is divided.Optimizing the number of subclusters in BIRCH is essential because it influences the granularity of clustering. If you choose too few subclusters, you might oversimplify the data, potentially missing valuable patterns. On the other hand, if you choose too many subclusters, you may introduce noise and overcomplicate the analysis.Techniques like Silhouette Score, of clustering results can help identify the optimal number of subclusters by evaluating the quality and meaningfulness of the clusters created.

4.4. Count of clusters (K) within FCM.

Fuzzy C-Means (FCM) is a clustering algorithm that assigns data points to clusters with degrees of membership. The key hyperparameter to fine-tune in FCM is the count of clusters (often denoted as 'K'). This parameter determines how many clusters the algorithm should aim to create from the data. Optimizing the value of 'K' in FCM is crucial because it directly impacts the granularity of the clustering result. A higher 'K' leads to more clusters, potentially capturing fine-grained patterns in the data, while a lower 'K' results in fewer clusters, grouping data points more broadly.Various methods can be employed to find the optimal 'K' in FCM, including the Fuzzy Partition Coefficient (FPC), the Fuzzy C-Means Objective Function (J_m), or even visual inspection of the clustering outcome. Cross-validation techniques can also be useful for this purpose.

5.Hyperparameter Optimization Methods:

To accomplish hyperparameter tuning, a combination of Grid Search and Random Search methods is employed.

5.1. Grid Search:

Grid Search is a hyperparameter optimization technique used to systematically search for the best combination of hyperparameter values for a machine learning algorithm. But it's important to be clear that Grid Search is frequently applied to supervised learning tasks, like classification or regression, where model performance is evaluated using measures like accuracy or mean squared error. Grid Search might not be as frequently employed for hyperparameter optimisation in the context of clustering algorithms like BIRCH and FCM applied to clickstream data from e-commerce websites.

However, you would need to establish suitable evaluation measures for clustering quality if you intended to use Grid Search in this situation to fine-tune hyperparameters. The Davies-Bouldin index and the Silhouette Score are two popular clustering metrics. A general description of possible uses for Grid Search is given below:

5.1.1. Define Hyperparameters: Identify the hyperparameters you want to optimize. In the case of BIRCH and FCM, these could include the number of subclusters within BIRCH and the count of clusters (K) within FCM.

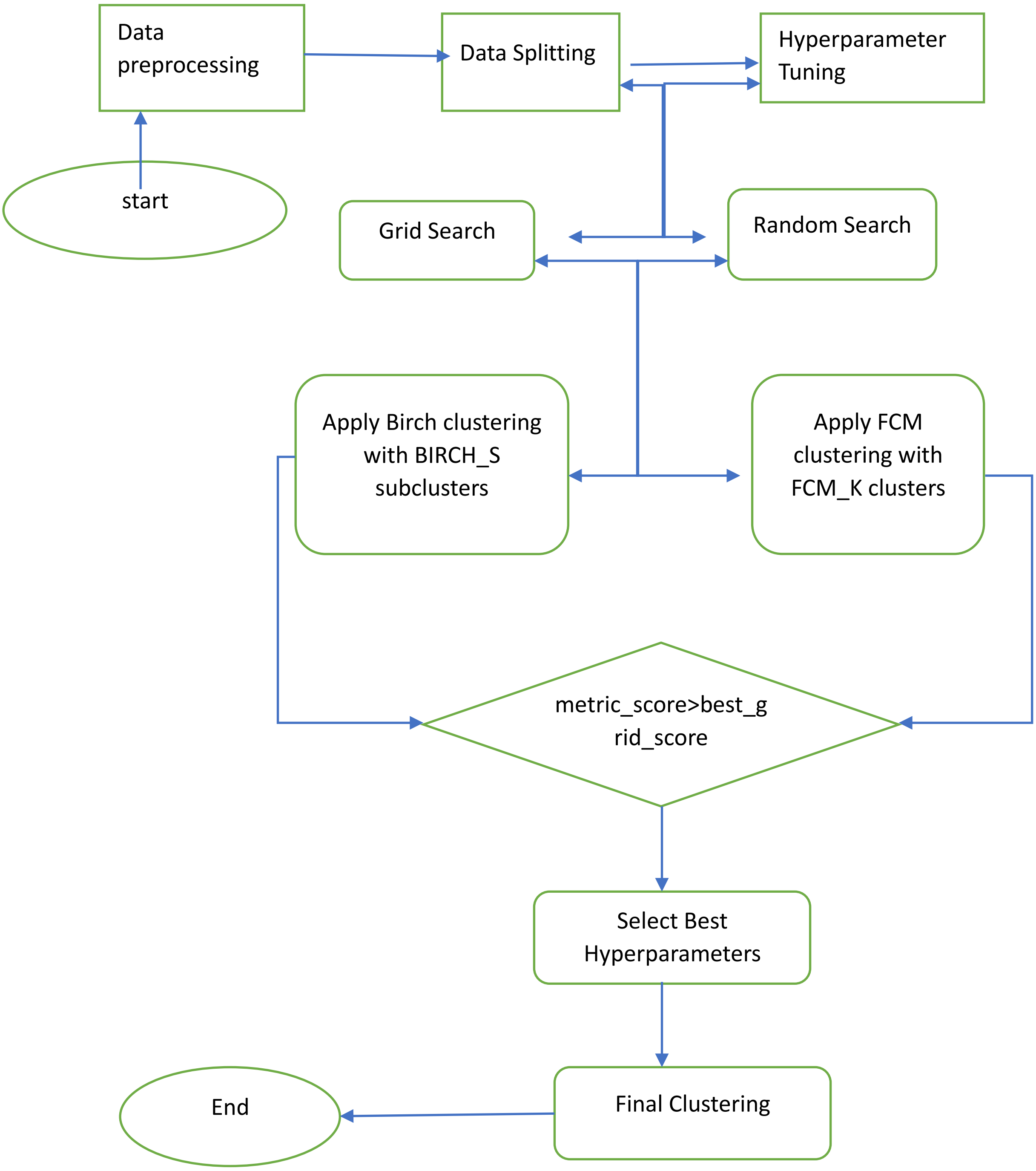

Fig. 1. Overall flow diagram

5.1.2. Define Parameter Grid: Create a grid of possible values for each hyperparameter. For example:

Number of Subclusters in BIRCH: [2, 3, 4, 5]

Count of Clusters (K) in FCM: [2, 3, 4, 5]

5.1.3. Evaluation Metric: Choose a clustering evaluation metric, such as the Silhouette Score or Davies-Bouldin index, to assess the quality of clustering for different hyperparameter combinations.

5.1.4. Grid Search Algorithm: Use an implementation of Grid Search, which exhaustively evaluates all possible combinations of hyperparameters based on the chosen evaluation metric.

5.1.5. Evaluation and Selection: For each combination of hyperparameters, apply BIRCH and FCM clustering to your e-commerce clickstream data. Calculate the chosen clustering metric for each combination.

5.1.6. Best Hyperparameters: Select the combination of hyperparameters that yield the highest or most desirable value of the clustering metric.

Here are the formulas for the two common clustering evaluation metrics mentioned:

5.2. Random Search:

Random Search is another hyperparameter optimization technique that can be applied to fine-tune hyperparameters in machine learning algorithms, including clustering algorithms like BIRCH and FCM when used for e-commerce website clickstream data analysis. Unlike Grid Search, Random Search does not systematically explore all possible hyperparameter combinations but instead randomly samples hyperparameters from predefined ranges. Here's how Random Search can be applied, along with relevant formulas for evaluating clustering quality:

$$s'(i) = \frac{b'(i) - a'(i)}{\max\{a'(i), b'(i)\}}$$

5.2.1. Define Hyperparameters and Ranges:

Identify the hyperparameters you want to optimize, such as the number of subclusters in BIRCH and the count of clusters (K) in FCM.

Specify the ranges or distributions from which these hyperparameters will be randomly sampled.

5.2.2. Evaluation Metric:

Choose a clustering evaluation metric to assess the quality of clustering for different hyperparameter combinations. As mentioned earlier, common metrics include the Silhouette Score and Davies-Bouldin index.

6. Random Sampling:

Randomly sample hyperparameter combinations from the specified ranges. For example:

Number of Subclusters in BIRCH: Randomly select an integer between 2 and 10.

Count of Clusters (K) in FCM: Randomly select an integer between 2 and 10.

6.1. Clustering and Evaluation:

For each sampled hyperparameter combination, apply BIRCH and FCM clustering to your e-commerce clickstream data.

Calculate the chosen clustering metric for each combination using the formulas provided in the previous response.

6.2. Selection of Best Hyperparameters:

Identify the hyperparameter combination that yields the highest or most desirable value of the clustering metric.

6.3. Iterate and Refine:

Random Search can be repeated for a predefined number of iterations or until a satisfactory result is achieved. The randomness in sampling helps explore a wide range of hyperparameters efficiently.

Algorithm

```
# Step 1: Data Preprocessing
Load and preprocess e-commerce clickstream data
# Step 2: Define Hyperparameters and Ranges
Define BIRCH_S_RANGE as a list of possible values for the number of subclusters
Define FCM_K_RANGE as a list of possible values for the count of clusters (K)
# Step 3: Split Data (Optional)
Split the data into training and testing sets if needed
# Step 4: Grid Search (Optional)
best_grid_score = -inf
for BIRCH_S in BIRCH_S_RANGE:
  for FCM_K in FCM_K_RANGE:
    # Apply Birch clustering with BIRCH_S subclusters
    # Apply FCM clustering with FCM_K clusters
    # Calculate clustering evaluation metric (e.g., Silhouette Score)
metric_score = evaluate_clustering(BIRCH_S, FCM_K)
    # Record the metric value and hyperparameter combination
    if metric_score>best_grid_score:
best_grid_score = metric_score
best_BIRCH_S = BIRCH_S
best_FCM_K = FCM_K
# Step 5: Random Search
best_random_score = -inf
for _ in range(num_random_iterations):
```

BIRCH_S = random.sample(BIRCH_S_RANGE, 1)[0]
FCM_K = random.sample(FCM_K_RANGE, 1)[0]
 # Apply Birch clustering with sampled BIRCH_S subclusters
 # Apply FCM clustering with sampled FCM_K clusters
 # Calculate clustering evaluation metric
metric_score = evaluate_clustering(BIRCH_S, FCM_K)
 # Record the metric value and hyperparameter combination
if metric_score>best_random_score:
best_random_score = metric_score
best_random_BIRCH_S = BIRCH_S
best_random_FCM_K = FCM_K
 # Step 6: Select Best Hyperparameters
if best_grid_score>best_random_score:
BIRCH_S_OPTIMAL = best_BIRCH_S
FCM_K_OPTIMAL = best_FCM_K
else:
BIRCH_S_OPTIMAL = best_random_BIRCH_S
FCM_K_OPTIMAL = best_random_FCM_K
 # Step 7: Final Clustering
Apply Birch clustering with BIRCH_S_OPTIMAL subclusters
Apply FCM clustering with FCM_K_OPTIMAL clusters
 # Step 8: Interpret Results
Analyze clustering results and extract insights from the clickstream data

Algorithm 1begins by preparing e-commerce clickstream data through data cleaning and organization. Then, it enters the realm of hyperparameter selection, where we determine the best settings for grouping data. This involves considering how many groups (subclusters or K clusters) our analysis should create. Optionally, we may split our data into training and testing sets for validation purposes.Next, we embark on two different journeys. In one, we systematically explore various settings through a grid search, carefully noting the best ones. In the other, we take a more random approach, trying out different configurations to see if we stumble upon something exceptional.After this exploration, we make an important decision: selecting the ultimate settings that gave us the best results from both approaches. Armed with these optimal settings, we finally apply the clustering process to group our data.The last step involves scrutinizing these groups to uncover insights and patterns within the clickstream data, helping us gain a deeper understanding of user behavior on the e-commerce website. In essence, this algorithm helps us make sense of user interactions, potentially leading to more informed business decisions.

Result and Discussions:

An approach for grouping e-commerce clickstream data is explained in the prompt as a hyperparameter optimisation algorithm. The algorithm's main objective is to determine the best BIRCH and FCM clustering hyperparameters and use those values to cluster the data. The programme first use both Grid Search and Random Search to methodically explore possible combinations of hyperparameters. With the use of this procedure, the optimal values for the number of subclusters (BIRCH_S_OPTIMAL) and the number of clusters (FCM_K_OPTIMAL) that produce the optimum clustering performance may be determined. The algorithm applies the BIRCH and FCM clustering algorithms on the complete clickstream dataset using the best settings after identifying the optimum hyperparameters. The final clusters are produced in this step. The ultimate clusters are anticipated to represent user groups with comparable patterns of behaviour. These clusters can be examined to glean useful information about how users engage and browse the e-commerce website. Businesses can use this data, for instance, to more effectively target their marketing campaigns, personalise their suggestions, and enhance the look of their websites. Overall, the prompt's algorithm is a potent tool for grouping e-commerce clickstream data and gleaning insightful information from it. Here is an illustration of how the algorithm might be put to use in real life: In order to understand how people engage with its website and pinpoint certain user categories, an e-commerce company wishes to leverage clickstream data. The method can be used by the company to cluster the clickstream data and then draw conclusions from the clusters. For instance, the company might discover clusterswhereas other clusters of users are more likely to buy things that they have already viewed. of users who are more inclined to browse new products. This data can be used by the company to better effectively target its marketing campaigns and personalize suggestions. The algorithm can be used to spot unusual user behaviour as well. The company might discover, for instance, that a particular group of customers are coming from out of the ordinary places or are spending a lot of time on a particular set of sites. These clusters can be looked into by the company to find any potential fraud or security problems. In general, the algorithm is a useful tool for e-commerce companies looking to better understand their customers.

7. Future Enhancement

Further enhancements can include exploring ensemble clustering techniques, where multiple clustering algorithms collaborate to provide more comprehensive and stable results. Lastly, the development of real-time adaptation mechanisms would enable clustering models to continuously evolve and adapt to changing user behavior patterns, facilitating dynamic personalization and an improved user experience on e-commerce websites. These future directions collectively aim to elevate the capabilities of clickstream data analysis, enabling businesses to derive deeper insights and provide more tailored and engaging online experiences to their customers.Additionally, focusing on advanced feature engineering techniques holds great potential for refining the quality of clustering outcomes. By extracting more informative and relevant features from the clickstream data, clustering models can become more resilient to variations in hyperparameter choices, ultimately leading to more accurate and robust results. Moreover, automated hyperparameter optimization libraries, such as Optuna, Hyperopt, or Scikit-Optimize, offer intelligent and efficient strategies for finding optimal hyperparameter configurations, relieving the burden of manual tuning and further streamlining the clustering process.

8. Conclusion:

In conclusion, the combination of Birch clustering, Fuzzy C-Means (FCM), Grid Search, and Random Search for hyperparameter tuning offers a powerful approach to analyze e-commerce website clickstream data. This method enables the discovery of meaningful user behavior patterns, which can drive improvements in marketing strategies, user personalization, and website optimization.By systematically exploring hyperparameter space through Grid Search and efficiently sampling from it using Random Search, we can find the most suitable hyperparameters for the clustering algorithms. However, this process should be accompanied by continuous monitoring and adaptation to account for changes in user behavior and evolving requirements.

Reference

[1] J. Kogan, C. Nicholas and M. Teboulle, “A Survey of Clustering Data Mining Techniques,” Berkhin, P., pp. 25–71, Springer, Heidelberg, 2006.

[2] A.K. Jain, “Data clustering: 50 years beyond K-means,” Pattern Recognition Letters, vol. 31, no. 8, pp. 651–666, 2010.

[3] A. Ben Ayed, M. Ben Halima and A.M. Alimi, “Survey on clustering methods: Towards fuzzy clustering for big data,” 6th International Conference of Soft Computing and Pattern Recognition (SoCPaR), Tunis, Tunisia, Aug 11-14, 2014, IEEE, pp. 331–336.

[4] Z.H. Zhou, “Ensemble methods: foundations and algorithms,” CRC press, Jun 6, 2012.

[5] A.Y. Ng, M. Jordan and Y. Weiss., “On spectral clustering: Analysis and an algorithm,” Advances in Neural Information Processing Systems (NIPS), vol. 14, no. 2, pp. 849-856, Dec 2001.

[6] H. Dhahri and A.M. Alimi, "The modified differential evolution and the RBF (MDE-RBF) neural network for time series prediction", IEEE International Conference on Neural Networks 2006, Conference Proceedings, pp. 2938.

[7] S. Bouaziz, H. Dhahri, A.M. Alimi and A. Abraham, "A hybrid learning algorithm for evolving Flexible Beta Basis Function Neural Tree Model", Neurocomputing, vol. 117, pp. 107-117, 2013.

[8] Z. Li, C. Shang and Q. Shen, "Fuzzy-clustering embedded regression for predicting student academic performance," IEEE International

Conference on Fuzzy Systems (FUZZ-IEEE), Jul 24, 2016, IEEE, pp. 344-351.

[9] M. Xu, M. Guo, L. Shang and X. Jia, “Multi-value image segmentation based on FCM algorithm and Graph Cut Theory,” IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), Jul 24, 2016, IEEE, pp. 1333-1340.

[10] Nasser, I.M., Al-Shawwa, M.O., Abu-Naser, S.S., 2019. A proposed artificial neural network for predicting movies rates category .

[11] Nguyen, H., Drebenstedt, C., Bui, X.N., Bui, D.T., 2020. Prediction of blast-induced ground vibration in an open-pit mine by a novel hybridmodel based on clustering and artificial neural network. Natural Resources Research 29, 691–709.

[12] Riese, F.M., Keller, S., Hinz, S., 2020. Supervised and semi-supervised self-organizing maps for regression and classification focusing onhyperspectral data. Remote Sensing 12, 7.

[13] Zhang, Y., Yin, B., Cong, Y., Du, Z., 2020. Multi-state household appliance identification based on convolutional neural networks and clustering. Energies 13, 792.

[14] Tan, L., Li, C., Xia, J., Cao, J., et al., 2019. Application of self-organizing feature map neural network based on k-means clustering in networkintrusion detection. COMPUTERS MATERIALS & CONTINUA 61, 275–288.

[15] Yang, M.S., Nataliani, Y., 2017. Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters. Pattern Recognition71, 45–59.

[16] Tian Zhang, Raghu Ramakrishnan & Miron Livny , pages141–182 (1997), "BIRCH: A New Data Clustering Algorithm and Its Applications",Data Mining and Knowledge Discovery .

[17] Fanny Ramadhani, Muhammad Zarlis, Saib Suwilo, January 2020,"Improve BIRCH algorithm for big data clustering, IOP Conference Series Materials Science and Engineering Fanny Ramadhani.