Journal of Autonomous Intelligence (2024) Volume 7 Issue 1
doi: 10.32629/jai.v7i1.903

ORIGINAL RESEARCH ARTICLE

Construction of agricultural product consumer group portrait and analysis of precision marketing strategies based on *K*-means cluster analysis

Cheng Kong[1,*], **Haliyana Khalid**[1], **Zhihao Gao**[2]

[1] *Azman Hashim International Business School, Universiti Teknologi Malaysia, Jalan Sultan Yahya Petra, Kuala Lumpur 54100, Malaysia*

[2] *Claro M. Recto Academy of Advanced Studies, Lyceum of the Philippines University Muralla Cor. Real Sts., Intramuros, Manila 1002, Philippines*

*** Corresponding author:** Cheng Kong, kongcheng1116@163.com

ABSTRACT

In today's technologically advanced landscape, where data flows in unparalleled volumes, the power of data mining stands out as a transformative force. Its capabilities extend beyond mere analysis; data mining can be harnessed to create intricate and detailed profiles of various consumer groups. This is particularly pertinent to the agricultural sector, which has long grappled with challenges of reaching its consumers effectively. The quest to refine and elevate marketing strategies for agricultural products amidst this data deluge led us to a methodical approach. We initiated this by sourcing and cleansing customer analysis datasets from Baidu AI, a leading platform in the realm of artificial intelligence and data analytics. Such a foundational step ensured that the data underpinning our analysis was robust and free of inconsistencies. The subsequent analytical journey, comprising rigorous data exploration and the utilization of the *K*-means clustering method, allowed us to dissect and segment the vast data pool. Through this, we crafted a comprehensive consumer profile that is tailor-made for agricultural product consumption. Such segmentation offers invaluable insights, paving the way for marketers and producers to understand their audience's nuances. Our research findings highlight the remarkable prowess of the *K*-means clustering technique. When underpinned by sophisticated intelligent algorithms, it doesn't just cluster data; it offers a pathway to identify distinct customer segments, shed light on core product offerings that resonate with them, and sculpt effective marketing strategies. By integrating these insights with modern media channels, we can craft marketing narratives that echo the desires and needs of the target demographic. Such a precision-driven approach ensures a symbiotic relationship where products align seamlessly with customer preferences. Beyond meeting immediate needs, this alignment has broader implications. It allows enterprises to tap into the vast opportunities presented by the digital era, positioning them on a trajectory of sustained growth and relevance in an ever-evolving market landscape.

Keywords: *K*-means cluster; agricultural product; consumer group; marketing strategies

ARTICLE INFO

Received: 5 July 2023
Accepted: 22 August 2023
Available online: 13 October 2023

1. Introduction

Rural e-commerce services include online farmers' markets, digital farmhouses, characteristic tourism, characteristic economy, and investment attraction. Online farmers' markets. Quickly transmit supply and demand information for agriculture, forestry, fisheries, and animal husbandry, assist foreign enterprises in entering and exiting local markets, and assist local farmers in exploring domestic and foreign markets. Conduct agricultural product market trends and dynamic express delivery, match business opportunities, and publish product information. Green agricultural products mainly refer to pollution-free, safe, high-quality, nutritious, and healthy agricultural products that follow the sustainable development strategy advocated by the country, are produced according to specific production methods, are recognized by professional institutions at the national, provincial, and municipal levels, and are licensed to use green food labels. Such as green wheat, green rice, green meat, green eggs, green fruits, etc.[1]. The development of green agricultural products in China is still in its early stages, with the production area accounting for only one thousandth of the total production area. The total consumption only accounts for two thousandths of the total food consumption, with broad prospects. However, the market supply of green agricultural products is imbalanced, mainly composed of grains, beans, and vegetables, with relatively few processed products such as wine and cosmetics. Mainly concentrated in major cities such as Beijing, Shanghai, Guangzhou, and Nanjing. By delving deep into consumer profiles specific to agricultural products, it becomes feasible to cater to consumers in a manner that emphasizes convenience, quality, and affordability. Further, these insights facilitate the nurturing of contemporary consumption patterns, particularly in the burgeoning realm of e-commerce. With the meteoric rise of platforms that support interactive e-commerce and live-streamed sales, consumer behavior is shifting. There's a pronounced tilt towards these new e-commerce paradigms where consumers exhibit a preference for procuring agricultural products online, primarily through localized living platforms and notable e-commerce entities[2]. The digital space has undeniably emerged as the predominant avenue for disseminating information about agricultural products.

Given this shift in consumer behavior and the challenges it presents to businesses; a pertinent question arises: How can agricultural product stakeholders elevate their sales strategies to not just sell more effectively but also cultivate a loyal consumer base? Solutions tailored for the agricultural sector should ideally cater to diverse stakeholders, including origin brands and operational entities in agriculture. Such solutions ought to provide a holistic approach to O2O (online-to-offline) business strategies. This would empower merchants with multi-faceted promotional avenues, foster the establishment of dedicated membership ecosystems, and leverage an assortment of marketing and distribution instruments. These strategies, in tandem, would assist businesses in product outreach, amplifying activity, and making purchase decisions anchored in tangible sales data. By aligning online and offline strategies, it's possible to secure customer loyalty and drive consistent growth in performance metrics. Delving into the sales model, the concept of order-based green agricultural products finds its roots in order-based agriculture. This paradigm emphasizes production driven by specific orders, harnessing the market's power to allocate resources optimally[3]. Such a strategy ensures production is both precise and efficient, with order specifics encapsulating product quantity, quality, pricing, and even predetermined sales channels. This proactive approach, intertwining production and sales, offers distinct advantages. It curtails the pitfalls of production in the absence of a clear market demand, guarantees foundational interests of producers, and lends stability to market prices. Championing order-based agriculture is instrumental in catalyzing structural reforms in agriculture. This transformative model, moving away from the conventional "produce per locality's inclination", is geared towards aligning production with tangible market needs.

2. Related work

Consumer profiling is a perfect abstraction of a user's business profile by a company by collecting and analyzing data on their social attributes, lifestyle habits, and consumption behavior. It can be seen as the basic way for enterprises to apply big data technology. User profiles provide a sufficient information foundation for enterprises, which can help them quickly find accurate user groups and broader feedback information such as user needs. Consumer profiling is a perfect abstraction of a user's business profile by a company by collecting and analyzing data on their social attributes, lifestyle habits, and consumption behavior. It can be seen as the basic way for enterprises to apply big data technology. User profiles provide a sufficient information foundation for enterprises, which can help them quickly find accurate user groups and broader feedback information such as user needs[4].

The intermediary-based resale approach entails farmers channeling their produce to major supermarkets, wholesale markets, or directly supplying to restaurants via intermediary acquisition. This model currently dominates China's agricultural sales framework. Its prevalence is intricately linked to the country's logistics and transportation development level, especially given the geographically dispersed agricultural production across China. One of the pitfalls of this system is that it often diminishes the product's value by the time it reaches intermediaries or agents, leading to a market predominantly driven by sellers. For consumers, particularly when assessing green products, time plays a pivotal role. The elongation of the supply chain through additional intermediary steps can dampen consumer confidence and trust in the authenticity and quality of green agricultural products[5]. Regrettably, both the growers and intermediaries end up capturing significant profit margins. Given the inherently high costs associated with green agricultural products, this distribution of profit makes it challenging to foster enthusiasm and commitment to sustainable farming practices[6,7].

3. Methodology

3.1. *K*-means clustering algorithm

Clustering, as an important technique in data mining technology, is the process of discovering potentially regular information and knowledge in massive data letters by dividing the same feature data into a cluster[8–10]. The clustering result is based on the similarity of the characteristics among the data objects, the similarity of the data objects in the same cluster is as high as possible, and the difference between the different clusters is as large as possible, which is also an important index of the effectiveness of the clustering algorithm, i.e., compact density and separation degree. *K*-means clustering algorithm, also known as *K*-means clustering algorithm, is one of the partition-based clustering algorithms. It adopts a heuristic iterative process to repartition the data objects and renew the cluster center[11]. With the m-dimensional data set as *X*, its clustering is divided into *k* clusters, and their cluster centers are $c_1, c_2, \ldots, c_k$, respectively. The number of data points in the cluster w_i is denoted by *i*, and the number of objects in the data set *X* is denoted by *n*. Including:

$$c_i = \frac{1}{n}\sum_{x \in w_i} x \tag{1}$$

The basic idea of the algorithm is to assume a set of *N* element objects and the quantity value *K* of the cluster to be generated. In the first round, *K*-sampled elements were randomly selected as the initial cluster centers, and the length of the *K*-cluster centers was calculated from the distance of the other sample elements, which were divided into *K* clusters according to the distance. In each of the following rounds, the iterative operation of the above steps is carried out, and the average value of the element object is taken as the center of clustering for the next round, and the clustering success is represented until the condition that the cluster center point is no longer changed in the iterative process is satisfied[12,13].

A data set *N* is known to contain *n* data objects from which arbitrarily selected or designated *k* data points are used as initial cluster centers $Z_j(I), j = 1,2, \ldots, k$. Calculate the distance from each data point to the *k* cluster centers in turn $D(x_i, Z_j(I)), i = 1,2, \ldots, n; j = 1,2, \ldots, k$, and redistribute the objects of each cluster based on the distance. If satisfied:

$$D(x_i, Z_j(I)) = min\{D(x_i, Z_i(I)), j = 1,2,3, \ldots, n\}, x_i \in w_k \tag{2}$$

In the domain of data mining, the *K*-means algorithm stands out as a frequently employed method due to its effectiveness and simplicity. This paper intends to provide a concise overview of the conventional *K*-means algorithm and subsequently, introduce an enhancement by integrating it with the neighbor propagation algorithm, resulting in the improved AP + *K*-means algorithm. The primary objective behind this amalgamation is to optimize customer segmentation systems specifically tailored for e-commerce websites.

The operational crux of this method involves first determining the data points that are closest in distance to the cluster's centroid. These data points are then amalgamated into their respective clusters. Post this aggregation, the mean of each cluster is recalculated to identify new centroid locations, setting the stage for subsequent iterative refinements. A critical metric, the error sum of squared criterion function *J*, is computed to measure the effectiveness of the clustering in each iteration.

$$J_c(I) = \sum_{j=1}^{k} \sum_{k=1}^{n_j} \left\| x_k^{(j)} - Z_j(I) \right\|^2 \tag{3}$$

For inexperienced users, in the use of *K*-means algorithm, the selection of *K*-value of clustering is not enough, only after many times of the cluster process and its results, to determine the appropriate *K*-value[14–16]. For the improvement method, some scholars have proposed the ISODATA algorithm, using the automatic merging and splitting of the class to obtain the more suitable *K*-value, and some will obtain the best *K* value by using the mixed *F* statistic according to the theory of variance analysis. The basic flowchart of *K*-means clustering algorithm and its process example **Figure 1**.

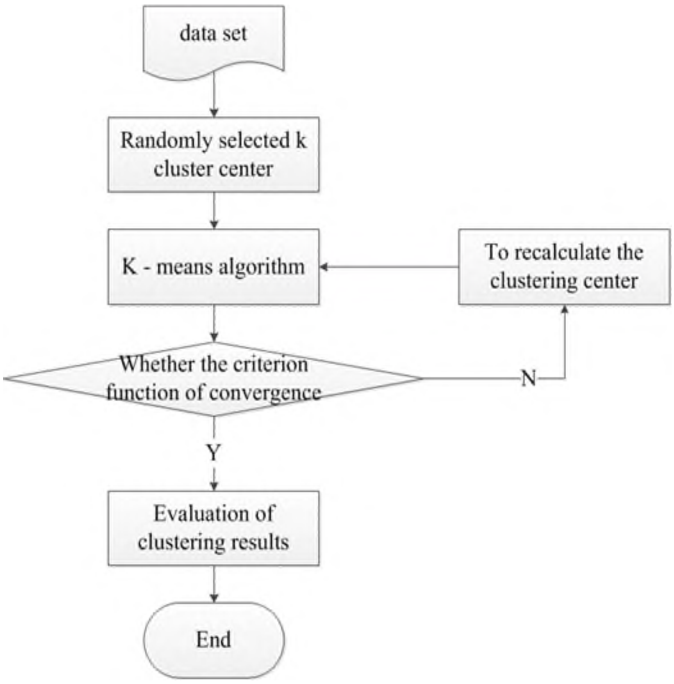

data set
Randomly selected k cluster center
K - means algorithm
To recalculate the clustering center
Whether the criterion function of convergence
N
Y
Evaluation of clustering results
End

Figure 1. The basic flowchart of the *K*-means clustering algorithm.

The use of the *K*-means algorithm must take into account the time complexity of its data sample calculation in order to achieve the clustering effect. Because the algorithm is an iterative adjustment algorithm, the selection of cluster centers must undergo repeated updated adjustments in order to achieve the best value, but the calculation process will result in tedious data and a significant time commitment. Some researchers suggested that in order to clear up this ambiguity and increase the effectiveness of the clustering process, the candidate set of the cluster center should be deleted if the features of the data items are comparable.

The *K*-means algorithm's clustering criteria function, which is derived by simply summing the squared error values of each cluster in the data set, is unable to handle clusters of data sets with unequal densities and significant size disparities. A mixed *K*-means method clustering criterion function is established in order to achieve this. The function changes the way of reassigning the data object to the cluster by the conventional *K*-means algorithm and utilizes a mixed measure of data object to the center to replace the original *K*-means algorithm's weighting of the Euclidean distance according to a particular percentage. The data object is assigned to the cluster with the center point that minimizes the weighted distance using the Euclidean distance in the classical *K*-means method. The formula is as follows:

$$J = \alpha \sum_{i=1}^{K} \frac{X_i}{N} \sigma_i + (1-\alpha) J_c(I) \tag{4}$$

where N represents the total number of data sets, K represents the number of clustering, σ_i represents the standard deviation of the cluster i, x_i represents the number of cluster i, α is relaxation parameters. When the density is relatively loose, α set between [0.6, 1]. At this time the standard deviation in the cluster dominates. To make the sum of the standard deviation in the weighted cluster minimum, the weight is the proportion of the number of data objects in the cluster. The benchmark function ε contains the standard deviation σ_i in the cluster, its function is to make the data objects in the cluster as close as possible to its center point, the weight of σ_i is x_i/N , the function is to make the standard deviation of the cluster with more data objects contribute to the criterion function value, so that in each iteration of the K-means algorithm, the data object will be more inclined to be assigned to the cluster with less number of data objects.

3.2. Customer grouping

Distinct customer segments possess varied preferences, underscoring the essence of precision marketing. The core objective of this approach is to meticulously segment customers, subsequently deploying tailored marketing tactics, thus enhancing both the likelihood of marketing success and sales efficacy. For establishments like chain retail department stores, the demographics of the surrounding locale can exhibit unique characteristics and purchasing tendencies. A deep dive into these demographic distributions offers invaluable insights, shaping localized store-specific marketing strategies, and amplifying their precision. For instance, in regions densely populated by the working class, there's a notable inclination towards promotions and discounts, especially concerning everyday agricultural products. Here, time-sensitive deals on agricultural goods often see an uptick in sales. Conversely, in more upscale neighborhoods, there's a discernible emphasis on product quality—attributes like eco-friendliness, health benefits, freshness take precedence. The price elasticity in such areas is considerably lower, suggesting a robust market for premium agricultural commodities. Through comprehensive data collection and meticulous demographic segmentation, it's feasible to craft and execute marketing strategies that resonate profoundly with each specific group, ensuring a more targeted and impactful reach.

3.3. Customer value analysis

Customers stand as the bedrock upon which enterprises are built. As commercial competition intensifies, the cost associated with acquiring new customers has witnessed a surge. In this context, retaining the existing customer base isn't just a strategy; it's an imperative that conserves resources.

Drawing from the Pareto principle, it's observed that a significant 80% of an enterprise's profits are attributed to a mere 20% of its customer base. It's paramount that this 20% is nurtured and retained. Conversely, the residual 80% of the customers, while not as profit-centric, are indispensable in terms of generating traffic and engagement for the enterprise, making their retention essential as well.

Historical customer management methodologies heavily leaned on the instincts of marketing professionals and the volume of customer transactions as primary metrics. However, the advent of clustering algorithms has ushered in a more nuanced and data-driven approach to discern customer value.

At its most basic, customer categorization can segregate them into four distinct cohorts:

- Critical retention.
- Key development.
- Essential maintenance.
- Foundational.

For a more granular understanding, this segmentation can further be refined into eight distinct customer types:

- High value.
- Strategic development.

- Vital retention.
- Essential maintenance.
- Average value.
- Development potential.
- Standard maintenance.
- Basic retention.

Such detailed customer stratification enables businesses to tailor their strategies, ensuring optimized engagement and retention for each segment.

4. Construction of agricultural product consumer customer profile based on *K*-means cluster analysis

The analysis data for this experiment comes from the customer personality analysis project data on Baidu AI Studio, with a total of 2240 instance data. This data is the natural and social attributes of 2240 customers recorded in a certain shopping mall, as well as the consumption behavior data within one time cycle. The data has been desensitized, but has basic attributes such as year of birth, education level, and annual household income. At the same time, there is also the customer's consumption of vegetables, fruits, meat, fish, candies, and appliances in the past 2 years, which is suitable for the needs of this study.

4.1. Customer clustering based on RFM model

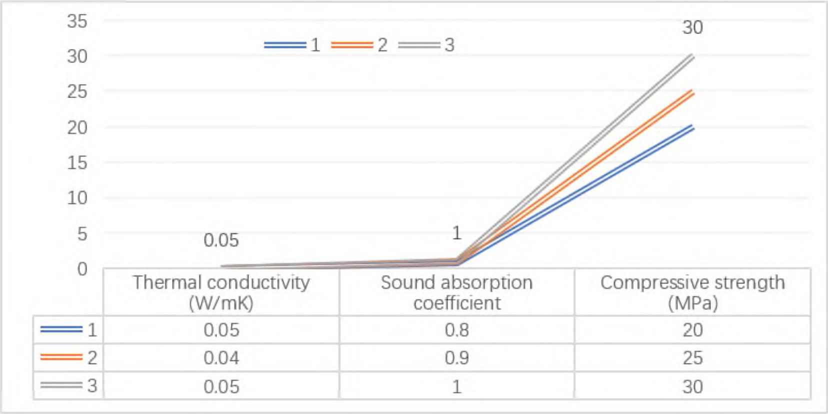
The primary role of the RFM model is to gauge the inherent value of a specific customer segment, subsequently assessing the potential benefits and value that can be derived from it. Within the dataset employed in this study, '*R*' (Recency) depicts the interval since the customer's last purchase. '*F*' (Frequency) quantifies the frequency of a customer's purchases within a specified timeframe. '*M*' (Monetary) signifies the cumulative expenditure on a range of products, including vegetables, fruits, meats, fish, candies, and appliances, spanning the past two years.

Upon data refinement, the sample was segmented into eight distinct customer categories based on varying metric values: lost customers, critical re-engagement customers, newcomers, high-value customers, essential maintenance customers, deep-engagement customers, and potential customers. A striking revelation was the significant attrition rate, with lost customers constituting 35.24% of the entire sample. Key re-engagement customers made up 22.11%, while high-value and potential customers represented around 10% and a mere 1.26% respectively.

In the realm of multi-feature customer clustering, this study leveraged 11 attributes, encompassing factors like age, consumption patterns, behaviors, and purchasing styles. These acted as vectors for the *K*-means clustering process. Given the non-static nature of the *K*-value, multiple iterations were tested to align the clustering output with business analytical requirements. The 'elbow method' served as the foundational approach for initial *K*-value determination, followed by a comparative analysis of the Sum of Squared Errors (SSE) for each potential *K*. The optimal *K*-*v*alue, marked by the minimal SSE, was identified as 7. To ensure dimensional consistency across all feature vectors utilized in clustering, each vector's spread was standardized, ensuring values ranged between 0 and 1. This normalization paved the way for the adoption of the Euclidean distance measure within the *K*-means clustering process. After judicious feature selection and outlier data filtration, the *K*-means algorithm was fully deployed for comprehensive customer analysis. This was further complemented by linear transformations of the raw data via the min-max standardization approach, with Euclidean distance serving as the foundational metric.

4.2. Precision marketing strategy selection

Precision marketing must target customers for marketing activities, and different customer groups need to adopt different marketing methods. So it is necessary to determine the target customers based on product characteristics before marketing in order to achieve precision and effectiveness. For example, for customer groups who prefer discounted products, for agricultural products with a relatively short shelf life, they can target these customers as marketing targets for some products that are on the market. You can even set up special zones, which are mainly used for customers with traffic. The target customers for agricultural products required for daily life are those with children and a high demand for agricultural products in daily life. Agricultural products that focus on high-end quality must target urban elites or small bourgeoisie groups who pursue delicacy, freshness, green, and health. Core products are usually the core values or benefits that a product can provide to customers, but the core values and benefits of different customer needs are different. There will not be a product that meets all customers, so it is necessary to define different core products for different target customers. Customers that purchase things at a discount do so on the justification that such products are inexpensive and need a specific discount. Their primary requirement is that the amount and diversity of agricultural products are sufficient and easy to acquire, which is particularly important for those with children who have a strong demand for agricultural products in their everyday life. Their hectic lifestyles are a source of frustration for this group, and they value attentive care above all else. **Figure 2** shows the sample results based on Rfm.



	Thermal conductivity (W/mK)	Sound absorption coefficient	Compressive strength (MPa)
1	0.05	0.8	20
2	0.04	0.9	25
3	0.05	1	30

Figure 2. Rfm based sample results.

5. Conclusions

The most effective aspect of marketing lies in addressing customer needs or pain points. The larger the needs or pain points, the better the marketing effect. In order to better achieve precision marketing of agricultural products in the era of big data., This study obtained customer personality analysis project data from Baidu AI Studio, conducted data cleaning, data analysis, *K*-means clustering analysis, customer clustering and information mining, and constructed a customer profile for agricultural product consumption, in order to provide reference for precision marketing of agricultural products.

Author contributions

Conceptualization, CK and HK; methodology, CK; validation, HK; formal analysis, HK; investigation, HK; resources, CK; data management, CK; writing—first draft preparation, CK; writing—review and editing, ZG; writing and execution, CK and HK; visualization, CK and HK; supervision, CK and HK. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Fan L. Research on precision marketing strategy of commercial consumer products based on big data mining of customer consumption. *Journal of the Institution of Engineers (India): Series C* 2023; 104: 163–168. doi: 10.1007/s40032-022-00908-7
2. Wang F. Analysis of user portraits in the cosmetics industry. In: Proceedings of the 2022 International Conference on Social Sciences and Humanities and Arts (SSHA 2022); 8 April 2022; pp. 973–977.
3. Chiu MC, Chuang KH. Applying transfer learning to achieve precision marketing in an omni-channel system—A case study of a sharing kitchen platform. *International Journal of Production Research* 2021; 59(24): 7594–7609. doi: 10.1080/00207543.2020.1868595
4. Ajay P, Nagaraj B, Jaya J. Bi-level energy optimization model in smart integrated engineering systems using WSN. *Energy Reports* 2022; 8: 2490–2495. doi: 10.1016/j.egyr.2022.01.183
5. Zhang M. Research on precision marketing based on consumer portrait from the perspective of machine learning. *Wireless Communications and Mobile Computing* 2022; 2022: 9408690. doi: 10.1155/2022/9408690
6. Wang T, Li N, Wang H, et al. Visual analysis of e-commerce user behavior based on log mining. *Advances in Multimedia* 2022; 2022: 4291978. doi: 10.1155/2022/4291978
7. Rajendran A, Balakrishnan N, Ajay P. Deep embedded median clustering for routing misbehaviour and attacks detection in ad-hoc networks. *Ad Hoc Networks* 2022; 126: 102757. doi: 10.1016/j.adhoc.2021.102757
8. Hou R, Kong YQ, Cai B, Liu H. Unstructured big data analysis algorithm and simulation of Internet of Things based on machine learning. *Neural Computing and Applications* 2020; 32: 5399–5407. doi: 10.1007/s00521-019-04682-z
9. Lakshmanaprabu SK, Shankar K, Ilayaraja M, et al. Random forest for big data classification in the internet of things using optimal features. *International Journal of Machine Learning and Cybernetics* 2019; 10: 2609–2618. doi: 10.1007/s13042-018-00916-z
10. Aldino AA, Darwis D, Prastowo AT, Sujana C. Implementation of K-means algorithm for clustering corn planting feasibility area in south lampung regency. *Journal of Physics: Conference Series* 2021; 1751(1): 012038. doi: 10.1088/1742-6596/1751/1/012038
11. Wu S, Liu J, Liu L. Modeling method of internet public information data mining based on probabilistic topic model. *The Journal of Supercomputing* 2019; 75: 5882–5897. doi: 10.1007/s11227-019-02885-8
12. Ajay P, Sharma A, Reddy PS, et al. Data analytics and cloud-based platform for internet of things applications in smart cities. In: Proceedings of the 2022 International Conference on Industry 4.0 Technology (I4Tech); 23–24 September 2022; Pune, India. pp. 1–6.
13. Rani LN, Defit S, Muhammad LJ. Determination of student subjects in higher education using hybrid data mining method with the K-means algorithm and FP growth. *International Journal of Artificial Intelligence Research* 2021; 5(1): 91–101. doi: 10.29099/ijair.v5i1.223
14. Wang W, Xia F, Nie H, et al. Vehicle trajectory clustering based on dynamic representation learning of internet of vehicles. *IEEE Transactions on Intelligent Transportation Systems* 2020; 22(6): 3567–3576. doi: 10.1109/TITS.2020.2995856
15. Sarowar G, Kamal S, Dey N. Internet of Things and its impacts in computing intelligence: A comprehensive review—IoT application for big data. *Big Data Analytics for Smart and Connected Cities*. IGI Global; 2019. pp. 103–136.
16. Cui Z, Jing X, Zhao P, et al. A new subspace clustering strategy for AI-based data analysis in IoT system. *IEEE Internet of Things Journal* 2021; 8(16): 12540–12549. doi: 10.1109/JIOT.2021.3056578