Segmentation of Services Provided by E-Commerce Platforms Using PAM Clustering

Divya Gaikwad and Hemraj Lamkuche*
1,2 Symbiosis Centre for Information Technology, Symbiosis International (Deemed University), Pune, Maharashtra, India
Email: *hemraj@scit.edu

Abstract. Today, data analytics have helped in solving many critical issues in various domains. From credit card fraud detection, detecting cancer to sentiment analysis, data analytics has come a long way. One of the important domains is e-commerce. With everything going digital, shopping has made its significant place in the digital world. A Partition around Medoids (PAM) clustering algorithm is discussed for grouping the services provided by e-commerce websites to their customers. The PAM clustering method helps to work upon real-time data with mixed data types. The clusters formed will provide better insights and patterns for the businesses to make better decisions in order to improve their customers’ engagement and multiply their profits using the historical data. With limited time and resources to invest, this method helps the company to proceed further by understanding the focus points of each service and their scope of development. Implementation of the proposed PAM clustering is done on the artificial dataset and is found to be effective.

Keywords: PAM Clustering, Silhouette, Gower distance, Medoids, Partition around Medoids.

1. Introduction
Many e-commerce websites are taking the lead all across the world. The facilities and services provided to these customers have advanced very rapidly. These e-commerce websites handle a large range of businesses like clothing, electronic gadgets, accessories, medicines, groceries, luxury items, etc. [2]. Each of these businesses provided the customers with multiple services to make their experience with the website or smart applications as smooth as possible. Few of these services that are provided to the customer play a major role in successfully running the business and are widely used by a major proportion of the customer base. Also, there are few services that do not majorly contribute to the functioning of the business. The crucial problem faced by the companies is understanding [1] which services are preferred more by the customers and which services are not liked by the customers and need to be enhanced in order to improve their customer’s experience and run a more profitable business [3].

The companies mainly need to identify the different types of services and the necessity to improve their performance in order to retain the existing customers and attract new customers. With limited resources and limited time, all the services and all the customers cannot be given equal importance. Different groups of customers have different requirements, and customized facilities provided to these customers will lead to great profits. The data needs to be divided into groups that need different measures
of attention, and insights from these clusters can be used for allotting the resources [4]. By making a more precise decision of forming the groups, the company will be able to achieve its goal.

Advanced analytics has made a prominent mark in the modern-day analytics field. With its technical approach and accurate results, advanced analytics helps to solve critical problems on huge data very quickly. For different issues faced by the company, different techniques of advanced data analytics can be used. PAM clustering is the Partition around medoids clustering. It helps in making segments of the data on the basis of the similarity index of the data points. PAM clustering is more suitable for real-time data as it can handle mixed data types, unlike many other clustering algorithms [5]. It uses the actual data points of the data set as medoids to form clusters. This type of clustering is more reliable to deal with outliers as compared to kmeans clustering. The clusters formed can provide good insights into the customer preference and usage of different services, which would further be useful for the company to make the decision for the resources invested in enhancing specific services to improve their customer's experience. With such major advantages and demonstrated better performance as compared to kmeans and rough kmeans, PAM clustering is highly recommended for the given scenario.

2. Literature Review

The author talks about a k-medoids algorithm used for clustering in huge data. To facilitate the local computation when searching for the optimal number of medoids, this algorithm uses the tin of medoids. The paper also states that this algorithm is more well-organized than most of the present k-medoids techniques but also retains a good quality of a k-medoid algorithm. The paper also discusses the applications of the new algorithm in road network extraction [6].

The author describes clustering as a procedure to group similar objects into one unit. The papers state that out of several types of clustering methods, the k-means clustering algorithm has comparatively less time complexity. This type of clustering is sensitive to outliers and would result in less accuracy [7]. They state that contrary to k-means clustering, the k-medoids clustering algorithm has fewer limitations. In this paper, they put forward a grid-based clustering method that will have higher or better accuracy than that of a k-medoids method. The Grid Multi-Dimensional k-medoids algorithm involves the thought of cluster validity index and is reflected from the outcomes of the experiments and shows that this new method has better and higher correctness than that of the k-medoids technique [8].

The author highlighted that earlier research had concluded in producing an `all rules algorithm used to mine the data that contains all the rules of confidence and coverage thresholds. This paper discusses the applications of the K-medoids and PAM clustering algorithms with respect to these rules so that one may identify the similar and analogous sets of rules and understands them in a better way [9].

The author stated that exploration of the data is the uncovering of interesting patterns which may exist in spatial databases. [30] The paper discusses three important contributions. Firstly, CLARANS, a new clustering method, is proposed with the goal of identifying the spatial structures that might be present in the data. The experiment outcomes demonstrate that when checked and compared to the existing method of clustering, CLARANS is proved to be productive and capable. [10] Secondly, it is spoken about how CLARANS can manage the point objects and also the polygonal objects. Thirdly, two spatial data mining algorithms are developed that can understand the relationship between the spatial and non-spatial attributes.

The authors have presented an improvised algorithm of the k-medoids based on a CF-Tree. In this paper, a lot of improvement has been made in the k-medoids algorithm's concept. The training sample data is stored in a CF-Tree, and k-medoids algorithm is then performed to form clusters of the leaf nodes of the tree. In [11] is specific number of clusters are grouped from the root of the tree. The drawbacks of the k-medoids are improved in this algorithm. A few of drawbacks are stated are time complexity and scalability for huge datasets.

The authors talk about how clustering is very prominent in understanding the data, creating models for predictions on it, and dealing with the outliers in the provided data. The clusters that possess a similar nature or characteristics in the dataset are then grouped together using effective methods. [12] They also state that with increasing volumes of data in the real world the massive data with fewer background data...
can be divided into interesting insights and get insights about it with clustering. The two methods that the researchers shed light upon are K-means clustering and K-medoids clustering. The input that is given into the experimentation for these methods is data points that are distributed randomly and are further clustering based on how similar they are to one another. [13] The results are them compared to one another and have been proved that the time consumed in cluster head selection and space complexity of the overlapping cluster is better in the case of K-Medoids clustering compared to the K-Means clustering. This paper also states that K-Medoids clustering has better execution time and is not sensitive to outliers.

The authors put forth the importance of clustering methods and explain the technique and its necessity in companies. They discuss the different types of clustering methods that can be implemented. The paper also contains the properties that are necessary to consider a clustering type as a good clustering algorithm [15].

The authors have proposed an approach that helps online stores to provide customized marketing services to the customers by creating groups based on the customers’ data. This helps the stores focus on the important activities that generate profits by identifying segments of customers with the same intensions. For this segmentation process, the factors that affect customer behavior are considered. After the clustering is done, these online stores can promote and do marketing for the service for its online customers [14].

The authors state that by creating segments of the consumers into numerous categories, better insights into the customer behavior can be found. Real-world datasets consist of more than one type of data types that show the customer activity. This information of the customers in the form of segments helps companies to improve their customer relationships also improvise their marketing approaches to enhance their customers’ expectations. The paper proposes a soft clustering technique that uses a latent mixed-class membership clustering method to identify the customers depending on their information [17].

The authors use the credit card usage data for building the model samples and develop a framework for creating predictive models. These models use the pattern-based clustering method [18]. They have used the monetary matrix and fluctuate-rate matrix to study different modes. With the clustering done on both the matrixes, customer behavior and nature are identified. By using these characteristics, a two-dimension customer segmentation model is developed.

The authors propose an algorithm for K-medoids clustering. This algorithm runs similar to that of a K-means algorithm also tests a number of approaches for identification and selection of the initial medoids to form clusters. The distance matrix is calculated and used this calculation to find new medoids iteratively in every step [16]. Few artificial datasets that belong to real-time are used to check the correctness of the algorithm. These results are then compared to the other algorithms’ outcomes with regard to the adjusted Rand index. They also state that the proposed algorithm takes less time for computation as compared to other algorithms [19].

The authors discuss different clustering methods. They state that data mining is identifying the relationship and patterns that exist within a large database. One of the significant methods of information mining is the clustering technique. The gene expression data have high dimensions and are complex. Hierarchical clustering technique and partitioning clustering technique are used to identify patterns of gene expression, and this helps in classifying the collected samples. In the paper, the three portioning methods, namely k-means, PAM also rough k-means are studying and compared to one another to classify the cancer dataset. The researchers conclude that PAM clustering technique performs better when compared to the other clustering techniques [20].

The authors discuss the standing of e-services in b2c e-commerce. With no face-to-face contact, the clients judge the excellence by e-services provided to them [21]. This paper also deliberates the role of these e-services and their application to enhance client services. Two research projects are used to display the customer and business point of view.

The author states that advanced analytics helps the firms have a complete view of the operations and customers. The insights from the analysis done can be utilized in decision-making to achieve the goals.
The mining techniques are discussed for the utilization and problems related to their implementation [22].

The authors discuss the importance of the transformation of reformed enterprises to sustain in a competitive market. [23] The study includes the research of strategic management and corporate innovation. The importance of internal business processes mentioned in this paper.

The authors state that it is actually significant for e-commerce businesses to handle the satisfaction also the loyalty of their customers in order to sustain for the long term. A conceptual framework consisting of the relationship between e-services and customer behaviors in the hostel industry has been discussed in the paper [29].

The authors state that there are dissimilar methods also algorithms that can be used to get insights from given huge data. The state is clustering is one of the most significant methods of data mining. The definition and aim of clustering are discussed in the paper. One major aim of clustering, outlier detection, is discussed further. In this paper, research in health-related data sets is focused upon. PAM, CLARA, CLARANS, and ECLARANS are mentioned in this paper. Several performance measures are used to understand and recommend a clustering method for outlier detection[24].

The authors discuss the concept of outlier detection. They state that outlier detection is important to identify anomalies in the data set. A clustering method is proposed to detect outliers. PAM clustering is used to form small clusters that are considered outlier clusters. The remaining outliers are identified, and the distance between the medoids of current cluster and each point in the current cluster is calculated [25].

The authors talk about the need to update the current technologies in order to the changing data trends and patterns. This paper discusses the static and dynamic datasets and the optimal number of clusters for the respective types of datasets. For static datasets, the optimal number of clusters is said to be constant, while for dynamic datasets, the optimum number of clusters may vary from time to time. The paper involves the implementation of a method based on fuzzy silhouette on energetic data. This helps in finding the optimum number of clusters. This is done by associating conventional clustering techniques with artificial data in addition to dynamic customer segmentation.

The authors propose a multivariate control chart based on the Gower distance. This can manage both continuous and categorical data. A study was further proposed by the authors to compare the properties of the planned control chart with existing multivariate control charts. It was found that with an increasing number of categorical data, the proposed control chart performed better. A real case study was discussed with the implementation and applications of the proposed control chart [26].

The authors discuss the kmeans clustering algorithm in detail. Numerous metrics and similarity measures are discussed. The Gower similarity coefficient is stated in particular in this paper. These experiments are done on various datasets. The results of this experimentation have proved that the accuracy of kmeans clustering algorithm using the Gower similarity coefficient has outperformed the accuracy with the other metrics.

The authors talk about a kmeans clustering algorithm that performs well for both numerical too categorical data. A new cost function in addition to distance measure was proposed for the simultaneous occurrence of the values. A modified description of the cluster center is proposed to overcome the limitation of using only numerical data. This study has been implemented on real-world data groups and has been proved to be more effective as compared to many other clustering algorithms [27]. The authors discuss a k-medoids algorithm that deals with mixed feature type data. The algorithm creates partitions and a prototype for each class by optimizing the criteria. This proposed algorithm is implemented on many real symbolic datasets. The authors propose an algorithm for k-medoids clustering that runs similar to that of a k means clustering algorithm. The proposed algorithm finds new medoids using a distance matrix after every step. This algorithm is implemented on real and artificial datasets. The results of this clustering algorithm are compared with that of the other clustering algorithms.

The author states that the division of data into homogenous clusters is one of the major challenges. The hierarchical clustering does not have a solution for this problem. The author also talks about the limitations of clustering algorithms like kmeans clustering that perform only on numerical data. The
The author in this paper proposes a clustering procedure that is analogous to the K-means clustering algorithm but overcomes the limitation of working on only numerical data. Decision tree induction algorithms are used to make rules for clusters [28]. The author discusses the Euclidean distance. In this paper, the complete statistical approach of the Euclidean distance is shown. All the basic properties of the Euclidean distance are listed.

The authors talk about the problem of missing observations in multivariate data. Clustering algorithms that can perform well on mixed data are well developed. The paper discusses the study, which shows that the mixed data types clustering algorithms can be extended to analyze the data with missing values. The authors discuss the importance of clustering algorithms as a part of data mining. In this paper, the authors propose a clustering algorithm that performs well with categorical information. The clusters are formed based on the dissimilarity between the data. The objects are then merged using the silhouette coefficient. The author’s contemporary an agglomerative fuzzy K-Means clustering algorithm. This is a postponement of the fuzzy kmeans clustering method. Comparatively, more consistent clusters are formed using this new proposed algorithm. This algorithm can help in deciding the number of optimal clusters, which is a limitation of the kmeans clustering algorithm. The experiments have been done on artificial datasets, and the results are found effective.

The author discusses the need for clustering methods across different domains. In this paper, different clustering algorithms are discussed that can be selected based on the criteria and requirements of the business problem. The author presents a clustering validity procedure. This evaluated and compares the results of the different clustering algorithms. The experiments have been performed on artificial and real datasets.

3. Proposed PAM Clustering Algorithm

The purpose of this learning is to propose a technique to form segments of the services provided by an e-commerce company to their customers in order to make a decision as to which of the provided services are to be improved in which aspect. This helps in identifying the services that will help in enhancing the customer experience. The segments formed will provide insights from each cluster. Each cluster group will have its specifications. A PAM clustering algorithm is proposed to segment the services to help achieve this goal. PAM stands for Partition around medoids. In this method, original data points known as the medoids are used, unlike kmeans clustering, in which the mean values of the data points are used. The medoids are representational objects of the data. While kmeans tries to minimize the total squared error between points, PAM clustering attempts to minimize the dissimilarity between the medoids. The clusters are formed based on the similarity between the medoids. To calculate the dissimilarity between the medoids, Gower distance is used. This distance is capable of measuring the dissimilarity between medoids that are mixed categorical and numerical values. It is the distance between the average of partial dissimilarities across the individuals. They range in [0, 1]. The Gower distance amongst the object i and j is given by

$$d(i,j) = 1/p \sum_{a=1}^{p} d_{ij}^{(f)}$$

In the case of numerical variables, partial dissimilarity is the ratio of absolute alteration between the explanations to the maximum range that is noticed from all individuals.

$$d_{ij}^{(f)} = \frac{|x_{if} - x_{ij}|}{R_{f}}$$

In the case of qualitative variables, if the values are different, then the partial dissimilarity is equal to 1, and if the values are the same, then the partial dissimilarity is equal to 0. The optimal number of clusters that are to be formed are based on the silhouette coefficient. The silhouette coefficient is a quantity of how identical the cluster is to its own cluster in judgment to the other clusters formed. So, the number of clusters with the highest silhouette coefficient will be considered as the most optimum number of clusters.

The algorithm consists of the following steps to form clusters:

Step 1: Selecting random points and forming clusters
Initially, m random points are selected as the medoids from the given n data points of the dataset. All these points are allocated to the closest cluster.

Step 2: Association to the closest cluster
Each of the data arguments is related to the closest medoid by using the Gower distance measure. The clusters formed to contain the most similar data points.

Step 3: Calculating the swapping cost
For every pair of objects k that are not selected and the selected object I, the swapping cost (TCik) is calculated.

Step 4: Replacing the new points with existing points.
If the value of TCik < 0, then object i is further replaced by k.

Step 5: Repeating the procedure
Repeat the above steps 3 and 4 until no change in the medoids is observed.

The two possible ways of interpreting the results are by displaying and analyzing the summary of the clusters and by visualizing these clusters in lower-dimensional representation.

4. Implementation
To understand the uses and performance of this clustering algorithm, the PAM clustering is done on a sample data which includes few important attributes like the services provided by the company, the rating given to the service by the customer, volumes in which the service is used, mode of utilization of the services, etc. In the experimental data set, 1665 entries represent numerous activities done for different services provided by the company through different modes. The activities contain different ratings recorded from the customer. The data also describes the availability of the services in the smart app and if the services contribute to the revenue of the company. Table 1 represents data description; Table 2 represents the sample observations of an experimental dataset.

| Attributes   | Data Types |
|--------------|------------|
| Volumes      | Int        |
| Services     | Character  |
| Rating       | Factor     |
| Mode         | Factor     |
| Availability | Logical    |
| Revenue      | Logical    |

Table 2: The sample six observations of the experimental dataset

| Volumes | Services         | Rating    | Mode      | Availability | Revenue |
|---------|------------------|-----------|-----------|--------------|---------|
| 110     | Forgot Password  | Very Good | Smart App | No           | No      |
| 129     | Forgot Password  | Very Good | Smart App | No           | No      |
| 80      | Forgot Password  | Good      | Phone     | No           | No      |
| 82      | Forgot Password  | Good      | Smart App | No           | No      |
| 83      | Forgot Password  | Good      | Smart App | No           | No      |
| 60      | Delete Account   | Average   | Website   | Yes          | No      |

The PAM clustering has been performed on this experimental data. Each data point is considered as a medoid. The distance between the medoids is calculated using the Gower distance. The Gower distance...
also helps in identifying the two most similar medoids and two most dissimilar medoids. Clusters are formed based on the similarity between the medoids.

Table 3 consists of the data of the two medoids that are most similar to one another in the dataset. The values of the attributes Services, Rating, Mode, Availability, and revenue are the same for both the medoids. This makes the categorical attribute values the same for both the medoids. The volume attribute differs by a value of 1. This is the least possible difference between two variables with regard to the numerical attribute. This makes the medoids the most similar ones.

Table 3: Two most similar medoids

| Volumes | Services       | Rating | Mode   | Availability | Revenue |
|---------|----------------|--------|--------|--------------|---------|
| 80      | Forgot Password| Good   | Smart App | No           | No      |
| 81      | Forgot Password| Good   | Smart App | No           | No      |

Table 4 consists of the data of the two medoids that are most dissimilar to one another in the dataset. The values of the attributes Services, Rating, Mode, Availability, and revenue are all different for both the medoids. This makes the categorical attribute values dissimilar for both the medoids. The volume attribute differs by a value of 110. 19 is the least possible value in the dataset and 129 is the highest value in the dataset. This makes it the highest possible difference between two variables with regard to the numerical attribute. Altogether this pair of medoids is the most dissimilar one.

Table 4: Two most dissimilar medoids

| Volumes | Services       | Rating | Mode   | Availability | Revenue |
|---------|----------------|--------|--------|--------------|---------|
| 129     | Upgrade to Premium | Very Good | Website | Yes          | Yes     |
| 19      | Verification   | Bad    | Smart App | No           | No      |

Figure 1: Plot to find an optimal number of clusters using Silhouette Index

To form clusters, first, the optimal number of clusters needs to be found. While kmeans algorithm uses the elbow curve to find the optimum number of clusters, the PAM clustering algorithms use silhouette index for the same. Silhouette analysis helps in understanding the separation distance between the resulting clusters. It shows how close each point in a cluster is to the points in the neighboring clusters. The higher the value, the tighter the clusters are. Using this concept, the optimal number of clusters can be found. In Figure 1, the plot of the silhouette values is shown. The point with the maximum
value shows the high similarity of each point in one cluster as compared to the neighboring clusters. So, the value with the highest silhouette index value is the optimum number of clusters value.

From the Silhouette Plot, it was found that the optimal number of clusters to be formed is six. The 2-dimensional representation of the six clusters formed using the PAM clustering is shown in Figure 2. A t-SNE (t-Distributed Stochastic Neighbor Embedding) graph for the partitions is represented in the below figure.

![Figure 2: The 6 Clusters formed using PAM clustering](image)

The dimensions are reduced to lower space for a better understanding of the clusters. From Figure 2, six clusters are represented with six different colors. The clusters are mostly confined within themselves. Each cluster consists of medoids of the data set and is formed by the similarity between these medoids. Each cluster has a particular criterion that they describe.

5. Results
The clustering was applied, and the data was divided into 6 clusters. Each cluster has different definitions. These clusters will help in deriving insights from the data.

![Figure 3: Availability and revenue insights derived from cluster 1](image)

5.1 Insights derived from Cluster 1
This cluster consists of the data points, which are mostly not available on the Smart App and do not contribute to the Revenue. From the graph, we can see that the two services that are used to a large extent are the Verification service and Forgot Password service. Based on these insights, the company can further proceed and make decisions as to how the services need to be made available on Smart Apps
too in order to enhance their customers experience on Smart Applications too. Figure 3 shows availability and revenue insights derived from cluster 1.

5.2 Insights derived from Cluster 2
Cluster 2 consists of the data points that majorly have an average rating provided by the customer. The services in this cluster are available but do not contribute to the revenue. The insights from this cluster can be used to enhance the services by doing deep research on the customers’ requirements and complaints. Figure 4 shows availability and revenue insights derived from cluster 2.

![Figure 4: Availability and revenue insights derived from cluster 2](image)

5.3 Insights derived from Cluster 3
Cluster 3 consists of the data points that majorly have bad ratings. It consists of the services that are available to the customers and do not contribute to the revenue. The insights from this cluster are very important as they put forth the weak aspects of the company’s business and help in making decisions for improving the services provided to the customers. Figure 5 shows availability and revenue insights derived from cluster 3.

![Figure 5: Availability and revenue insights derived from cluster 3](image)

5.4 Insights derived from Cluster 4
Cluster 4 consists of the data points that majorly have an average user experience. The services are available to the customer and contribute to the revenue. The insights found from this cluster also play a major role in decision-making as all the services listed contribute to the revenue of the company’s business. Focus and enhancement of these services are very important to improve the customer service and also to make better profits. Figure 6 shows availability and revenue insights derived from cluster 4.

![Figure 6](image_url)

**Figure 6**: Availability and revenue insights derived from cluster 4

### 5.5 Insights derived from Cluster 5

Cluster 5 consists of the data points that mostly have a good user experience. The services found in this cluster are available to the customers and contribute to the revenue. With very high good ratings, the services seem to do fine. Such services can be sustained as they are customer-preferred as well as profitable. Figure 7 shows availability and revenue insights derived from cluster 5.

![Figure 7](image_url)

**Figure 7**: Availability and revenue insights derived from cluster 5

### 5.6 Insights derived from Cluster 6

All these data points in this cluster are available and do contribute to the revenue. From the graph as we can see that the ratings of these services are neither too good nor too bad. Many of the services fall into this cluster, and better services, and innovative ideas can be implemented in order to retain a huge number of customers through the numerous services. Figure 8 shows availability and revenue insights derived from cluster 6.
5.7 Description of the Clusters formed

The bubble graph shows the distribution of volumes within the clusters. The transaction volume distribution ranges from 15,447 to 24,108. From the figure, we can see that cluster 6 has the highest volumes with 24,108. This cluster is followed by cluster 5 with a volume 20,825. Further, cluster 2 has volumes 19,849, which are a little lesser than cluster 5. Cluster 4 and cluster 1 have volumes 15,447 and 14,988, respectively. The cluster with the least volume is cluster 3 with 10,390 volumes. Figure 9 represents the volumes present in each of the 6 clusters.

Each of the services provided by the e-commerce company belong in different proportions to different clusters, as shown in the below figure. While few services make a significant meaning in few clusters, few services can be neglected in few clusters due to very less volume of transactions or usage of those services. The decision as to what enhancements need to be made in each of these services can be made depending on the volumes of transactions and the important criteria like revenue, customer experiences etc. Figure 10 shows each service belonging to different clusters.
6. Conclusion

With e-commerce businesses having limited resources and limited time, decision-making plays a very crucial role. To handle such issues and have a supporting mechanism or procedure for decision-making, advanced analytics should be used. The PAM clustering algorithm is one such advanced data analytics technique that helps in forming groups based on the similarity between the actual data objects called the medoids. The groups formed will contain specific criteria and will help gain insights. These insights or criteria can be utilized by the decision-makers to understand the types of customer patterns. The clustering algorithm helps in narrowing down the focus onto lesser points with maximum scope of profits. It helps to make different decisions for different services and keep their business intact and multiple profits. The PAM clustering proposed is appropriate for real-time data sets with mixed data types. The clustering method has performed efficiently on the artificial dataset.

References

[1] Ahmad, A., & Dey, L. (2007). A k-mean clustering algorithm for mixed numeric and categorical data. *Data and Knowledge Engineering*, 63(2), 503–527. https://doi.org/10.1016/j.datak.2007.03.016.

[2] Aranganayagi, S., & Thangavel, K. (2007). Clustering Categorical Data using Silhouette Coefficient as a Relocating Measure. *Proceedings - International Conference on Computational Intelligence and Multimedia Applications, ICCIMA 2007, 1*, 97–102. https://doi.org/10.1109/ICCIMA.2007.328.

[3] Arora, P., Deepali, & Varshney, S. (2016). Analysis of K-Means and K-Medoids Algorithm for Big Data. *Physics Procedia*, 78(December 2015), 507–512. https://doi.org/10.1016/j.procs.2016.02.095.

[4] Ben Ali, B., & Massmoudi, Y. (2013). K-means clustering based on gower similarity coefficient: A comparative study. *2013 5th International Conference on Modeling, Simulation and Applied Optimization, ICMSAO 2013*. https://doi.org/10.1109/ICMSAO.2013.6552669.

[5] BinAlAbid, F., & Mottalib, M. A. (2012). An Accurate Grid-based PAM Clustering Method for Large Dataset. *International Journal of Computer Applications*, 41(21), 1–6. https://doi.org/10.5120/5821-7808.

[6] Bose, R. (2009). Advanced analytics: opportunities and challenges. *Industrial Management &
De Assis, E. C., & De Souza, R. M. C. R. (2011). A K-medoids clustering algorithm for mixed feature-type symbolic data. *Conference Proceedings - IEEE International Conference on Systems, Man and Cybernetics*, 527–531. https://doi.org/10.1109/ICSMC.2011.6083737.

Gower, J. C. (1985). *Properties of Euclidean and Non-Euclidean Distance Matrices*. 97, 81–97.

Halkich, M., & Vazirgiannis, M. (2001). A data set oriented approach for clustering algorithm selection. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2168, 165–179. https://doi.org/10.1007/3-540-44794-6_14.

Hong, T., & Kim, E. (2012). Segmenting customers in online stores based on factors that affect the customer’s intention to purchase. *Expert Systems with Applications*, 39(2), 2127–2131. https://doi.org/10.1016/j.eswa.2011.07.114.

HUANG, Z. (2004). Clustering algorithm for large data sets with mixed numerical and categorical values. *Dianzi Yu Xinxi Xuebao/Journal of Electronics and Information Technology*, 26(8), 1203–1209. https://doi.org/10.1117/12.538864.

Hunt, L., & Jorgensen, M. (2003). Mixture model clustering for mixed data with missing information. *Computational Statistics and Data Analysis*, 41(3–4), 429–440. https://doi.org/10.1016/S0167-9473(02)00190-1.

Kumar, P., & Wasan, S. K. (2011). Comparative Study of K-Means, Pam and Rough K-Means Algorithms Using Cancer Datasets. In *Proceedings of CSIT: 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)*, 1(2009), 136–140.

Kumar, V., Kumar, S., & Singh, A. K. (2013). *Outlier Detection : A Clustering-Based Approach*. 7, 16–19.

Li, M. J., Ng, M. K., Cheung, Y. M., & Huang, J. Z. (2008). Agglomerative fuzzy K-Means clustering algorithm with selection of number of clusters. *IEEE Transactions on Knowledge and Data Engineering*, 20(11), 1519–1534. https://doi.org/10.1109/TKDE.2008.88.

Li, Z. (2011). Research on customer segmentation model by clustering. *2011 International Conference on Computer Science and Service System, CSSS 2011 - Proceedings*, 3437–3440. https://doi.org/10.1109/CSSS.2011.5974496.

Meng, Y., Luo, K., & Liu, J. (2012). An improved K-medoids clustering algorithm. *Advanced Materials Research*, 562–564, 2106–2110. https://doi.org/10.4028/www.scientific.net/AMR.562-564.2106.

Murimbika, M., & Urban, B. (2013). Strategic management practices and corporate entrepreneurship: A cluster analysis of financial and business services firms in South Africa. *African Journal of Business Management*, 7(16), 1522–1535. https://doi.org/10.5897/AJBMI12.13666.

Ng, R. T., & Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering*, 14(5), 1003–1016. https://doi.org/10.1109/TKDE.2002.1033770.

Park, H. S., & Jun, C. H. (2009). A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications*, 36(2 PART 2), 3336–3341. https://doi.org/10.1016/j.eswa.2008.01.039.

Rai, P., & Singh, S. (2010). A Survey of Clustering Techniques. *International Journal of Computer Applications*, 7(12), 1–5. https://doi.org/10.5120/1326-1808.

Reynolds, A. P., Richards, G., & Rayward-Smith, V. J. (2004). The application of K-medoids and PAM to the clustering of rules. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 3177(January 2004), 173–178. https://doi.org/10.1007/978-3-540-28651-6_25.

Singh, M. (2002). E-services and their role in B2C e-commerce. *Managing Service Quality: An International Journal*, 12(6), 434–446. https://doi.org/10.1108/09604520210451911.

Sobihah, M., Mohamad, M., Ali, N. A. M., & Wan Ismail, W. Z. (2015). E-commerce service
quality on customer satisfaction, belief and loyalty: A proposal. Mediterranean Journal of Social Sciences, 6(2), 260–266. https://doi.org/10.5901/mjss.2015.v6n2p260.

[25] Subbalakshmi, C., Rama Krishna, G., Krishna Mohan Rao, S., & Venketeswa Rao, P. (2015). A method to find optimum number of clusters based on fuzzy silhouette on dynamic data set. Procedia Computer Science, 46(Icict 2014), 346–353. https://doi.org/10.1016/j.procs.2015.02.030.

[26] Tuerhong, G., & Kim, S. B. (2014). Gower distance-based multivariate control charts for a mixture of continuous and categorical variables. Expert Systems with Applications, 41(4 PART 2), 1701–1707. https://doi.org/10.1016/j.eswa.2013.08.068.

[27] Vijayarani, S., Nithya, S., & Research Scholar, M. P. (2011). An Efficient Clustering Algorithm for Outlier Detection. International Journal of Computer Applications, 32(7), 975–8887.

[28] Wu, R. S., & Chou, P. H. (2011). Customer segmentation of multiple category data in e-commerce using a soft-clustering approach. Electronic Commerce Research and Applications, 10(3), 331–341. https://doi.org/10.1016/j.elerap.2010.11.002.

[29] Zhang, Q., & Couloigner, I. (2005). A new and efficient K-medoid algorithm for spatial clustering. Lecture Notes in Computer Science, 3482(III), 181–189. https://doi.org/10.1007/11424857_20.

[30] Saravanan, N. (2013). Hand Geometry Recognition based on optimized K-Means Clustering and Segmentation Algorithm. International Journal of MC Square Scientific Research, 5(1), 11-14.