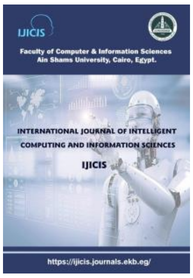
**IJICIS, Vol.22, No.2, 35-43**
**DOI: 10.21608/ijicis.2021.69041.1074**

International Journal of Intelligent Computing and Information Sciences

https://ijicis.journals.ekb.eg/


IJICIS
Faculty of Computer & Information Sciences Ain Shams University, Cairo, Egypt.
INTERNATIONAL JOURNAL OF INTELLIGENT COMPUTING AND INFORMATION SCIENCES
IJICIS
https://ijicis.journals.ekb.eg/

# Smart Support System for Evaluating Clustering as a Service: Behaviour Segmentation Case Study

M. Galal*
Predictive Analytics department, National Bank of Egypt, and Computer Science Deprtment
Ain Shams University,
Cairo, Egypt
mhdgalal@yahoo.com

T. Salah
Minimax Projects Manager
Mansoura, Egypt
tamer@minimax-soft.com

M.M. Aref
Computer Science, Faculty of computer science and information systems
Ain Shams University,
Cairo, Egypt
mostafa.aref@cis.asu.edu.eg

E. ElGohary
Information systems, Institute of National Planning, CLIP project manager
Cairo, Egypt
Esam.Elgohary@inp.edu.eg

Received 2021-03-22; Revised 2021-07-13; Accepted 2021-07-16

***Abstract:*** *Modern surveys reveal diminishing of socio-demographic segment descriptors, and evolution of dramatic increase of online services and customers. These conditions attract both researchers and decision makers to enhance market segmentation to gain customer loyalty and prevent customer attrition. This research contributes in developing a minor expert system to automate the evaluation of clustering process to enhance the Clustering as a Service (CaaS) through customer behavior segmentation case study. It comes as a part of the software development process to develop Customer Loyalty Intelligent Personalization (CLIP) system. The proposed expert system has been successfully implemented and tested over four months in two different dataset to proof the flexibility of implementation . The used data is a real customer data, it consists of 1659 customers, 146 products, and 5685 orders. The other datset consists of 668 transactions of real data in restaurant. The clustering is applied using the hierarchical clustering and it reached a good results with high efficiency. The proposed solution aims to be integrated with a plug and play product as it will be configured in different domains.*

***Keywords:*** *Customer Segmentation, Hierarchical Clustering, Clustering as a Service (CaaS).*

* Corresponding author: M. Galal
Predictive Analytics department, National Bank of Egypt, and Computer Science Deaprtmen,t Ain Shams University, Cairo, Egypt

E-mail address: mhdgalal@yahoo.com

## 1. Introduction

Recently, the relationship between companies and customers has become an indisputable aspect in business and thus, the presence of a governing mechanism to this relationship is essential. This controlling process of the interactions between organizations and customers is called Customer Relationship Management (CRM). Accordingly, the concept of CRM includes a set of methods and strategies to develop long-term, profitable relationships with customers.

Furthermore, the growing number of customers, the diversity of products on offer, and the complexity of customer behavior have made developing a tailored recommendation system for personal future needs a vital and challenging task. Herein, market segmentation is a powerful marketing technique solution, since it breaks down a target market audience into more manageable groups. It organizes customers based on demographic, geographic, behavioral, or psychographic categories or a combination of them. CRM streamlines this segmentation process so enterprises reach the customers who are most receptive to their products and services. Intelligent CRM uses data mining techniques within the marketing and sales sectors of business to improve analysis, increase revenues and save time.

Consequently, personalization of customer needs leads to better-targeted marketing campaigns and enhanced customer satisfaction with the ultimate aim of increased rates of customer retention, and improved competitive advantage [1]. Here is come CLIP role, CLIP is an intelligent, machine learning-based, real-time, personalized customer loyalty actions advisory system. CLIP will consider the wide variety of industries with customizable and configurable customers' features and behavior parameters.

This research focuses on CLIP's first phase which includes automating customer data segmentation, which is unlike traditional segmentation. It is based on both customer and purchases patterns formed by customers while they interact with an enterprise or make a purchasing decision. The paper proposes a minor expert system to automate customer data segmentation. The paper sections were: section 2 illustrated literature review, section 3 presented the proposed system, section 4 displayed a case study on client's data, and section 5 displayed the conclusions.

## 2. Literature Review

Researchers use various data mining techniques to figure out patterns in data. Their objective is to find customer segments or groups that permit enterprises to address the customer needs or desires, discover opportunities to optimize their customer journeys, and quantify their potential value to their business. This section illustrated the state of the art of research in this field of study.

Jo-Ting et al. 2013 [2] combined self-organizing maps and k-means methods to apply RFM (recency, frequency, and monetary) model to segment customers. It helped to segment customers into four types, loyal, potential, new, and lost customers.

Kristof et al. 2014 [3] investigated the influence of data accuracy and three segmentation techniques: RFM analysis, logistic regression, and decision trees. This research recommended the decision trees over the other techniques in the context of customer segmentation for direct marketing.

You et al. 2015 [4] proposed a model to accurately predict monthly supply quantity, using the RFM approach to select attributes to cluster customers into different groups. It used real data from a Chinese company. The applied techniques are RFM analysis, K means, and decision tree. The proposed model helped managers to identify the latent characteristics of different customer categories. The model was helpful in predicting marketing strategies that can reduce inventory for every customer category.

Abirami et al. 2016 [5] introduced a customer classification approach to analyze and estimate customer behavior using RFM analysis, K means, and association rules. It applied to the retail sector in India.

Jorge Everyman et al. 2017 [6] proposed a model based on recurrent neural network which used to predict the next event in a business problem. The research used two real datasets and showed that the proposed model surpassed the state-of-art and cross-validated precision in excess of 80%. It had limited number of unique values that limits feasibility of the approach.

Shreya et al. 2018 [7] explored the importance of K-means, hierarchical, and hybrid clustering models for customer segmentation. K-means clustering algorithm was relatively better in computational speed as compared to the hierarchical algorithms, it also required the full proximity matrix calculation for each iteration. K-means clustering gave a better performance for a large number of observations while hierarchical clustering had the ability to handle fewer data points.

Onur et al. 2018 [8] presented two clustering models to segment 70032 customers depending on their RFM values. The first proposed model suggested that 42936 customers should have premium cards. This allowed companies to make customized promotions to gain customers' loyalty. The second proposed model suggested four clusters. One of them contains 64081 customers. The company defined these customers as standard customers because their RFM scores are close to average scores. Thus, the company could choose not to give any card or any membership to these customers, since most of them are one-time buyers.

Fahed et al. 2019 [9] focused on maximizing Consumer Lifetime Value (LTV) to accommodate the dynamics in customer behavior for a medium-size retailer. It applied soft clustering Fuzzy C-Means (FCM) and hard clustering Expectation-Maximization (EM) algorithms to classify individual consumers who exhibit similar purchase history into segments. In the evaluation, cluster quality assessment (CQA) is applied. It showed EM algorithm scales much better than Fuzzy C-Means algorithms in the smaller dataset.

Hossein Abbasimehr et al. 2019 [10] presented clustering based on time series to extract dominant behavioral patterns of customers. It used bank customers' transactions data which are in the form of time series data. The data include the recency, frequency, and monetary (RFM) in business, that were supplied from the point-of-sale (POS) of a bank. The data accessed was of a seven-month time period. It would be preferred to analyze the data of a 12-month time period.

Fahed et al. 2020 [11] proposed three different market segmentation experiments using modified best fit regression using Expectation-Maximization (EM) and K-Means clustering algorithms were conducted and subsequently assessed using cluster quality assessment. The indicated analysis that the average lifetime of the customer was only two years, and the churn rate was 52%. Thus, a marketing strategy was devised based on these results and implemented on the departmental store sales. It was revealed in the marketing record that the sales growth rate increased from 5% to 9%.

Reviewing the listed literature indicates that researches focus on customer segmentation to discover the dominant patterns for customer movements, detect key factors that influence customer behavior to move within segments, reveal a relationship between brand and membership programs to increase customers loyalty, and improve decision making strategy to enhance marketing based on customer preferences. This research comes as part of a real E-commerce software implementation. This software aims to enable B2C institutions such as e-commerce and retail systems to manage their offers, discounts, bonuses, and other marketing tools to leverage customer loyalty. Customer segmentation is the first main component in CLIP system. This paper focused on customer segmentation where data scientists developed a minor expert system that automates clustering customer data.

## 3. Proposed Customer Segmentation System

This section illustrates the data flow of the proposed customer segmentation component. The component consists of four sequential processes. Figure 1 shows the data flow diagram of the proposed system. The first process is called Extract, Transform, and Load (ETL), which is a separate web application to receive sales transactional data from a user or third-party applications in csv format files. The ETL transforms transactional data into the schema expected by CLIP including the customer features table. The second process is concerned with data preprocessing to do data verification and data cleaning. The third process aims to conduct data aggregations and normalization which make data better fit with the clustering method. Finally, the fourth phase focuses on segments clustering using hierarchical clustering. The input of these models is a customer dataset where the customer is represented by one record as a feature vector.  The final output is the customer segments which are stored in a database for further processes.

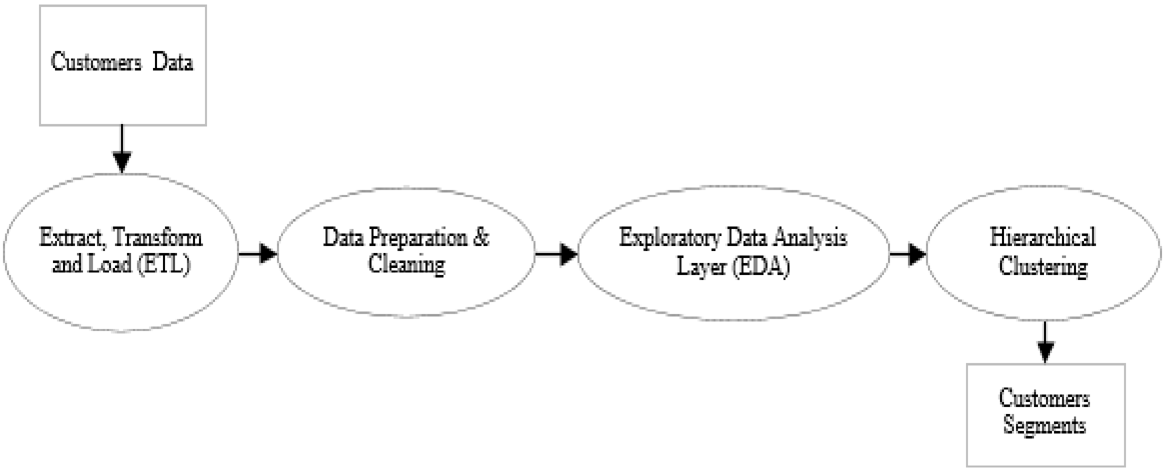

Customers Data
Extract, Transform and Load (ETL)
Data Preparation & Cleaning
Exploratory Data Analysis Layer (EDA)
Hierarchical Clustering
Customers Segments

Figure. 1: Customer Segmentation Expert System Dataflow.

The research uses real client data samples in the food sector. It consists of 1659 customers, 146 products, and 5685 orders. The development team selected and exported data that are related to sales only. The exported data consists of these main blocks, which were customer information, customer referrals, customer behavior, customer purchases, purchases per category, and loyalty offers. The research has aimed to automate the clustering phase to check if the data contains meaningful clusters or not. The exported data has included customer features with minimum requirements, which were categorized as: behavioral and purchase features only. Initially, the development team cleaned these data and dumped the empty columns. Afterward, hierarchical clustering is applied iteratively for defined sequential tasks in order to optimize and select the appropriate features to get customer clusters if they existed.

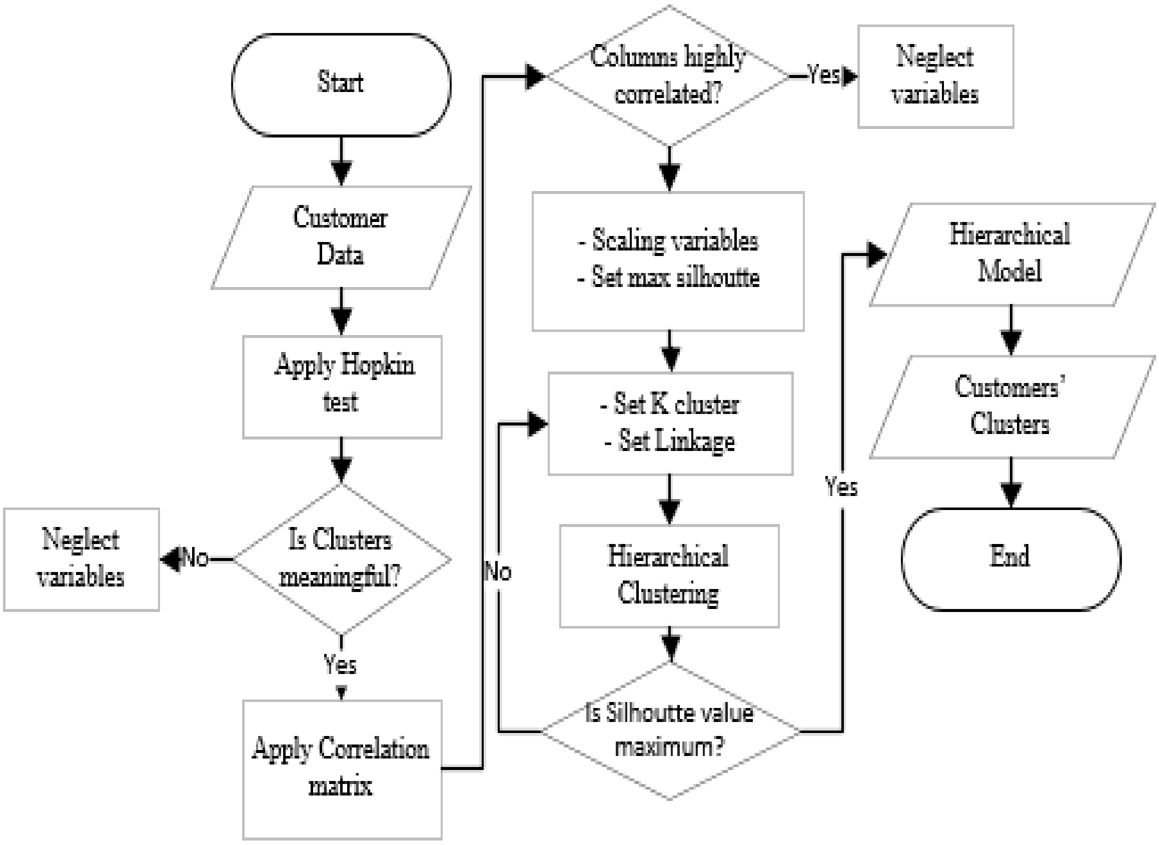


Figure. 2: Customer Segmentation Flowchart.

Figure 2 shows the customer segmentation expert system flowchart. It works in two main steps: checking clusters' tendencies and applying hierarchical clustering. Initially, the Hopkins test [12] is applied to assure cluster tendency between feature columns, by testing the spatial randomness of the data. Next, correlation is measured to remove highly correlated columns. The preprocessed output will be the feature columns that are not empty nor highly correlated and non-uniformly distributed. Thereafter, the hierarchical clustering process works on clustering the preprocessed features. These features are scaled using minmax scalar [13] to normalize input features into the range [0,1]. Consequently, the data scientists define the hierarchical clustering algorithm parameters, which are linkage [14] and k clusters. The linkage is used to measure similarity between clusters and k clusters is the number of defined clusters. The output segments/clusters are evaluated by silhouette score [15]. Repeatedly, those parameters and preprocessed features are varied and tuned to obtain the best combination to perform separable clusters. Briefly, the proposed minor expert system works as follows:

1. Apply Hopkin test on customer data to check features, if it forms meaningful clusters.
2. Apply correlation matrix on the input features to remove highly correlated features.
3. Scaling the selected feature columns.
4. Adjust hierarchical clustering algorithm parameters and train the feature columns.
5. Evaluating the built model using the silhouette score.
6. If the silhouette score is not maximum, then go to step 4.
7. If the model achieves the maximum silhouette score, then it is selected.

## 4. Case Study

This section illustrates how the proposed system performs on client data in the food sector to get customer segments. This data consisted of 1659 customers, 146 products, and 5685 orders. A combination of k-cluster and linkage parameters are used to form different models until better clusters segments are predicted. The silhouette score evaluates each built model. Table 1 represented the experiments' results. This table displays in four columns respectively: the iteration number, the main category of features, the detailed feature/column names, the defined parameters, and the evaluation metric. Table 1 shows the proposed expert system behavior on the customer data until proper customer clusters are revealed. The proposed system will work on automate checking the data existence in clusters until getting meaningful clusters. In the table below, the first three rows show no valid clusters formed, while the last two rows show more reliable clusters after refining hierarchical algorithm parameters as well as the feature columns. The category column shows the category of main features in customer data, which are their purchases and their categories. Those features are tuned based on both Hopkin test and correlation matrix. Repeatedly, the proposed expert system evaluates the feature columns as well hierarchical algorithm parameters until the most representative features and parameters are accomplished to get purposeful customers' clusters for further CLIP system phases.

To find out the system's performance with different business sectors, we used two client's data in Table 2. The first data belongs to a supermarket, and it consists of 1659 records. It achieves a 0.68 silhouette score. The second data belongs to a restaurant, and it consists of 668 records. It attains a 0.49 silhouette score. Ultimately, the proposed system has proved high scalability and stability to present clustering as a service in E-commerce.

Table 1 System Results on the Customer Data

| | Category | Features name | Parameters | Evaluation |
|---|---|---|---|---|
| 0 | Customer Purchases | Visit Frequency, Total Orders Number, Total Morning Orders Number, Total Weekend Orders Number, Total Order Items Number, Total Purchase Amount, Average Order Value, Customer Lifespan, Orders Frequency | Cluster 0 = 1658<br>Cluster 1 = 1<br>Linkage = average | silhouette = 0.8532 |
| | Purchases / Category | Category_1_Amount, Category_2_Amount, Category_3_Amount, Category_4_Amount, Category_5_Amount | | |
| 1 | Customer Purchases | Visit Frequency, Total Orders Number, Total Morning Orders Number, Total Weekend Orders Number, Total Order Items Number, Total Purchase Amount, Average Order Value, Customer Lifespan, Orders Frequency | Cluster 0 = 1657<br>Cluster 1 = 1<br>Cluster 2 = 1<br>Linkage = average | silhouette = 0.7912 |
| | Purchases / Category | Category_1_Amount, Category_2_Amount, Category_3_Amount, Category_4_Amount, Category_5_Amount | | |
| 2 | Customer Purchases | Visit Frequency, Total Orders Number, Total Morning Orders Number, Total Weekend Orders Number, Total Order Items Number, Total Purchase Amount, Average Order Value, Customer Lifespan, Orders Frequency | Cluster 0 = 319<br>Cluster 1 = 1340<br>Linkage = ward | silhouette = 0.6887 |
| | Purchases / Category | Category_1_Amount, Category_2_Amount, Category_3_Amount, Category_4_Amount, Category_5_Amount | | |
| 3 | Customer Purchases | Visit Frequency, Total Orders Number,Total Weekend Orders Number,Total Purchase Amount, Customer Lifespan, Average Order Value | Cluster 0 = 319<br>Cluster 1 = 1340<br>Linkage = ward | silhouette = 0.6993 |
| | Purchases / Category | Category_3_Amount, Category_4_Amount, Category_5_Amount | | |

Table 2 Other Datasets Results

| Data | Features | Clusters | Silhouette |
|---|---|---|---|
| Supermarket | VisitFrequency, TotalOrdersNumber, TotalMorningOrdersNumber, TotalWeekendOrdersNumber, TotalOrderItemsNumber, TotalPurchaseAmount, AverageOrderValue, CustomerLifeSpan, Category_1_Amount, Category_2_Amount, Category_3_Amount, Category_4_Amount, Category_5_Amount | 2 | 0.689 |
| Restaurant | VisitFrequency, TotalOrdersNumber, TotalMorningOrdersNumber, TotalWeekendOrdersNumber, TotalPurchaseAmount, AverageOrderValue, CustomerLifeSpan, Category_1_Amount, Category_2_Amount, Category_3_Amount, Category_4_Amount, Category_5_Amount | 2 | 0.4936 |

## 5. Conclusion

This research comes as a part of implementing an E-commerce system called CLIP. The research goal is to build a standalone and portable Clustering as a Service product (CaaS). It is a minor expert system that evaluates the client data until meaningful clusters are accomplished in an automated way. This portable product could be used from one client to another by the minimal need of administration and engineering processes. Its main objective is to automate the evaluation process of clustering to enhance the engineering lifecycle. A team of data scientists and researchers have worked together to build and evaluate the proposed CLIP system to assure the best model choice for the client data. This team does many experiments on various clients' data to ensure the system's reliability. The research includes a case study performed on client data in the food sector. It consists of 1659 customers, 146 products, and 5685 orders. Unlike the state-of-art of traditional clustering approaches, the proposed system accomplished 0.69 silhouette accuracy measurement. The client clusters will involve in the proceeding phases in the CLIP system to boost customer loyalty and curb customer churn.  In the future work, the proposed system will be applied in different domains. The automated model evaluation concept will be tested on a supervised data mining technique.

## References

1. Eric W. T. Ngai, Li Xiu, Dorothy C. K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Systems with Applications, 36(2) (2009) 2592-2602.
2. Jo-Ting Wei, Ming-Chun Lee, Hsuan-Kai Chen, Hsin-Hung Wu. Customer relationship management in the hairdressing industry: An application of data mining techniques. Expert Systems with Applications. 40(18) (2013) 7513-7518.
3. Kristof Coussement, Filip A.M. Van den Bossche, Koen W. De Bock. Data accuracy's impact on segmentation performance: Benchmarking RFM analysis, logistic regression, and decision trees. Journal of Business Research. 67(1) (2014) 2751-2758.
4. You, Z., Si, Y.-W., Zhang, D., Zeng, X., Leung, S.C.H. and Li, T., A decision-making framework for precision marketing, Expert Systems with Applications, 42(7) (2015) 3357-3367

5. Abirami, M. and Pattabiraman, V. , Data Mining Approach for Intelligent Customer Behavior Analysis for a Retail Store, Springer International Publishing, Cham, (2016) 283-291.
6. Joerg Evermann, Jana-Rebecca Resheb, Peter Fettkeb, Predicting Process Behaviour using Deep learning, Elsevier Decision Support System, (2017) 129-140.
7. Shreya Tripathi, Aditya Bhardwaj, Poovammal E, Approaches to Clustering in Customer Segmentation, International Journal of Engineering & Technology, 7(3) (2018) 802-807.
8. Onur Doğan, Ejder Ayçin, Zeki Atıl Bulut, Customer Segmentation by using RFM Model and Clustering Methods: a case study in retail industry, International Journal of Contemporary Economics and Administrative Sciences, 8(1) (2018) 1-19.
9. Fahed Yoseph, Nurul Hashimah Ahamed Hassain Malim and Mohammad AlMalaily, New Behavioral Segmentation Methods to Understand Consumers in Retail Industry, International Journal of Computer Science & Information Technology (IJCSIT), 11(1) (2019).
10. Hossein Abbasimehr, Mostafa Shabani, “A new methodology for customer behavior analysis using time series clustering: A case study on a bank’s customers”, Journal Emerald Insigh, 2019
11. Fahed Yosepha, Nurul Hashimah Ahamed Hassain Malimb, Markku Heikkiläc, Adrian Brezulianud, Oana Gemane, and Nur Aqilah Paskhal Rostamf, The Impact of Big Data Market Segmentation Using Data Mining and Clustering Techniques, Journal of Intelligent and Fuzzy Systems, (2020).
12. Banerjee A, Dave RN. Validating clusters using the Hopkins statistic, IEEE International Conference on Fuzzy Systems, 2004, p. 149-153.
13. Patro, S Gopal & Sahu, Kishore Kumar. Normalization: A Preprocessing Stage. International Advanced Research Journal in Science, Engineering, and Technology (IARJSET), (2015).
14. O. Yim, K. Ramdeen, Hierarchical Cluster Analysis: Comparison of three linkage measures and application to psychological data. The Quantitative Methods for Psychology, 11(1) (2015) 8-21.
15. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, (20) (1987) 53-65.