INTERNATIONAL JOURNAL OF SOFT COMPUTING

MEDWELL PUBLICATIONS

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

[1]Mohamed Galal, [2]Tamer Salah, [3]Mostafa Mahmoud Aref and [4]Esam Elgohary
[1]Department of Predictive Analytics, National Bank of Egypt, Department of Computer Science Ain Shams University Cairo, Egypt
[2]Minimax Projects Manager, Mansoura, Egypt
[3]Department of Computer Science, Faculty of Computer Science and Information Systems, Ain Shams University, Cairo, Egypt
[4]Department of Information Systems, Institute of National Planning, CLIP project manager, Cairo, Egypt

Key words: Customer Segmentation, Hierarchical Clustering, Clustering as a Service (CaaS)

Corresponding Author:
Esam Elgohary
Department of Information Systems, Institute of National Planning, CLIP project manager, Cairo, Egypt

Page No.: 29-33
Volume: 16, Issue 03, 2021
ISSN: 1816-9503
International Journal of Soft Computing

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 segmenting 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 intelligent Customer Loyalty Intelligent Personalization (CLIP) system. The proposed expert system system has successfully implemented and tested over four months. The used data is a real customer data, it consists of 1659 customers, 146 products and 5685 orders. The clustering is applied using the hierarchical clustering and it reached a good results with high efficiency.

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, pro table relationships with customers.

Furthermore, the growing number of customers, the diversity of products on offer and the complexity of customer behavior has 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 lead 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 wide variety of industries with customizable and configurable customer's features and behavior parameters.

This research focuses on CLIP 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 study proposes a minor expert system to automate customer data segmentation. The paper sections were.

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 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 art of research in this field of study.

You *et al*.[2] 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 Chinese company. The applied techniques are RFM (Recency, Frequency and Monetary) 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*.[3] proposed 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.

Tripathi *et al*.[4] 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 better performance for a large number of observations while hierarchical clustering had the ability to handle fewer data points.

Dogan *et al*.[5] proposed two clustering models to segment 70032 customers by depending on their RFM values. First proposed model suggested that 42936 customers should have premium cards. This allowed companies to make customized promotions to gain customers loyalty. 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.

Yoseph *et al*.[6] 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.

Yosepha *et al*.[7] 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 2 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 5-9%.

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 behaviour to move within segments, reveal 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 study focused on customer segmentation where data scientists developed a minor expert system that automates clustering customer data.

MATERIALS AND METHODS

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 separated web application to receive sales transactional data from user or third-party applications in csv format

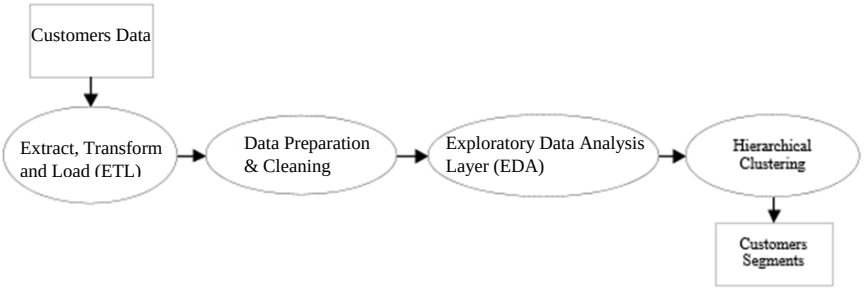

Fig. 1: Customer segmentation expert system dataflow

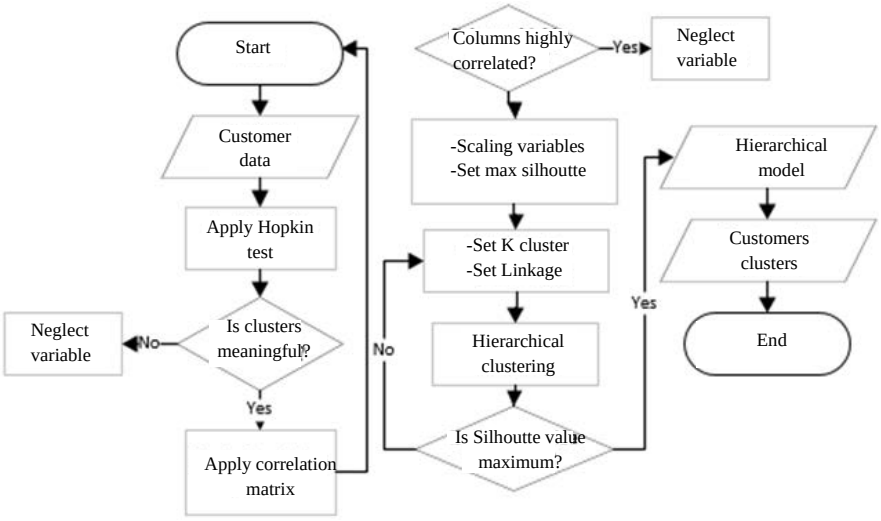

Fig. 2: Customer segmentation flowchat

files. The ETL transforms transactional data into the schema expected by CLIP including the customer features table. The second process is concerned about 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.

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 purchases features only. Initially, the development team cleaned these data and dumped the empty columns. Afterwards, 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 (Table 1).

Figure 2 shows the customer segmentation expert system flowchart. It works in two main steps: checking clusters' tendency and applying hierarchical clustering. Initially, Hopkins test[8] 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 which 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 min max scalar[9] to normalize input features into range [0,1]. Consequently, the data scientists define the hierarchical clustering algorithm parameters which are

Table 1: Sample results

Exp.	Category	Features name	Parameters	Evaluation
Exp0	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 Cluster 1 = 1 Linkage = average	Silhouette = 0.8532
	Purchases category	Category_1_Amount Category_2_Amount Category_3_Amount Category_4_Amount Category_5_Amount		
Exp1	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 Cluster 1 = 1 Cluster 2 = 1 Linkage = average	Silhouette = 0.7912
	Purchases category	Category_1_Amount, Category_2_Amount Category_3_Amount, Category_4_Amount, Category_5_Amount		
Exp2	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 Cluster 0 = 319 Cluster 1 = 1340 Linkage = ward	Silhouette = 0.6887
	Purchases category	Category_1_Amount Category_2_Amount Category_3_Amount Category_4_Amount Category_5_Amount		
Exp3	Customer purchases	Visit frequency, Total orders number Total weekend orders mumber Total purchase amount Customer lifespan Average order value	Cluster 0 = 319 Cluster 1 = 1340 Linkage = ward	Silhouette = 0.6993
	Purchases category	Category_3_Amount Category_4_Amount Category_5_Amount		

linkage[10] 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[11]. 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 following:

- Apply Hopkin test on customer data to check features, if it forms meaningful clusters
- Apply correlation matrix on features to remove highly correlated features
- Scaling the selected feature columns
- Adjust hierarchical clustering algorithm parameters and train the selected features
- The build model is evaluated using silhouette score
- If the silhouette score is not maximum, then go to step 4
- The model selected if it achieves the maximum silhouette score

RESULTS AND DISCUSSION

There have been different applied experiments on real customer data. This section illustrates an experiment using client data in the food sector. It consisted of 1659 customers, 146 products, and 5685 orders. The team has changed k clusters and linkage parameters to predict better clusters segments. The clustering has been evaluated using silhouette score. Table 1 represented the

experiments' results. The table displayed in its four columns the experiment name, the main category of features, the detailed feature/column names, the defined parameters, and the evaluation metric, respectively.

Table 1 showed the proposed expert system behavior on the customer data until proper customer segments revealed. The proposed system will work on automate checking the data existence in clusters until get meaningful clusters. The first three iterations showed no valid clusters formed while the last two iterations showed more reliable clusters after refining hierarchical algorithm parameters as well as the feature columns. In the table below, the category column showed the main tables where the main features were extracted from customer data, these tables were customer purchases and purchases categories. The features column consisted of the selected columns from the pre-mentioned category tables, these features chosen 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 accomplished to get purposeful customer's clusters for further CLIP system phases.

CONCLUSION

This research comes as a part of implementing a real E-commerce system called CLIP. The research goal is to build a standalone and a portable Clustering as a Service product (CaaS) using minor expert system that evaluates the features as well hierarchical algorithm parameters until the most representative features and parameters accomplished to get purposeful customers' clusters in an automated way. This portable product could be used from one client to another by minimal need of administration and engineering processes. Its main objective is to automate the evaluation process of clustering to enhance the engineering lifecycle. The teams of data scientists and researchers have worked together to build and evaluate the proposed CLIP system to assure the best model choice for each client's data. The data scientists' team have done many experiments on different clients' real data to assure the proposed expert system reliability. This research introduced an experiment done on real client data in the food sector. It consists of 1659 customers, 146 products, and 5685 orders. Unlike state-of-art of traditional clustering approaches, the proposed system accomplished 0.69 silhouette accuracy measurement.

In the future work, the proposed system will be applied to different domains and the feature engineering and extraction processes will be enhanced. The customer's' segments will involve the precedent phases in the CLIP system to boost customer loyalty and curb customer churn. The automated model evaluation concept will be tested on the supervised data mining techniques.

REFERENCES

01. Ngai, E.W.T., L. Xiu and D.C.K. Chau, 2009. Application of data mining techniques in customer relationship management: A literature review and classification. Expert Syst. Applic., 36: 2592-2602.
02. You, Z., Y.W. Si, D. Zhang, X. Zeng, S.C. Leung and T. Li, 2015. A decision-making framework for precision marketing. Expert Syst. Appl., 42: 3357-3367.
03. Abirami, M. and V. Pattabiraman, 2016. Data mining approach for intelligent customer behavior analysis for a retail store. Proceedings of the 3rd International Symposium on Big Data and Cloud Computing Challenges (ISBCC-16), March 10-11, 2016, Springer, Cham, Switzerland, pp: 283-291.
04. Tripathi, S., A. Bhardwaj and E. Poovammal, 2018. Approaches to clustering in customer segmentation. Int. J. Eng. Technol., 7: 802-807.
05. Dogan, O., E. Aycin and Z. Bulut, 2018. Customer segmentation by using RFM model and clustering methods: A case study in retail industry. Int. J. Contemp. Econ. Administrative Sci., 8: 1-19.
06. Yoseph, F., N.H.A.H. Malim and M. AlMalaily, 2019. New behavioral segmentation methods to understand consumers in retail industry. Int. J. Comput. Sci. Inf. Technol. (IJCSIT.), Vol. 11, No. 1.
07. Yoseph, F., A.H.N.H. Malim, M. Heikkila, A. Brezulianu, O. Geman and P.N.A. Rostam, 2020. The impact of big data market segmentation using data mining and clustering techniques. J. Intell. Fuzzy Syst., 38: 6159-6173.
08. Banerjee, A. and R.N. Dave, 2004. Validating clusters using the Hopkins statistic. Proceedings of the 2004 IEEE International Conference on Fuzzy Systems (IEEE Cat. No.04CH37542) Vol. 1, July 25-29, 2004, IEEE, Budapest, Hungary, pp: 149-153.
09. Patro, S. and K.K. Sahu, 2015. Normalization: A preprocessing stage. Int. Adv. Res. J. Sci. Eng. Technol., Vol. 2, No. 3.
10. Yim, O. and K.T. Ramdeen, 2015. Hierarchical cluster analysis: Comparison of three linkage measures and application to psychological data. Quantitative Methods Psychol., 11: 8-21.
11. Rousseeuw, J.P., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Applied Math., 20: 53-65.