Customer Behavior Clustering Based on Balance History Using Dynamic Time Warping Distance

Phan Duy Hung and Do Quang Dat

Abstract—Customer clustering, the division of customers into different groups, is a classical problem. It is especially important in banking as it serves multiple purposes in marketing, risk management, etc. Therefore, it has attracted the use of many modern machine learning models and techniques. But currently, most of them are only making use of “static” customer information. This paper proposes a new approach for customer clustering in banking based on the customers’ balance history. Basic Dynamic Time Warping (DTW) distance and Soft-DTW (SDTW) distance, an advanced form, are used to measure the difference between customers. To which, the two most popular strategies in time series clustering strategies, partitional and hierarchical, are applied which. In additional, some statistical features are given to prove the effectiveness of the proposed method.

Index Terms—Balance history, time series clustering, DTW, SDTW, partitional clustering, hierarchical clustering.

I. INTRODUCTION

With the evolution of big data and data mining, the classical problem of marketing and risk management in banking has attracted a lot of attention. One of the most common problems is how to effectively cluster customer, because customer information is very diverse, with age, gender, job, salary, spending patterns, etc.

Customer information can be divided into two groups named “static” and “dynamic”. The “static” customer information includes age, gender, job, salary, etc. This kind of information will stay consistent or rarely change. On the contrary, dynamic customer information will change frequently, for example, balance history, daily expenses.

There have been some studies regarding the matter. In [1], Jingyuan et al. combined Analytic Hierarchy Process (AHP) with K-means clustering to divide the Chinese weekly newspaper customers into four categories: platinum, gold, iron and lead. The indicators used for clustering were the customers’ current value and potential values such as discount rate, current purchase volume and service cost.

Yanan Gu et al. divided customer into two groups for commercial bank: redeemed customers and unredeemed customers [2]. This study focused on the implementation of reward program for current deposit customers of commercial banks, analyzing the differences in transaction behavior by using seven hypotheses: recency, frequency, monetary, CLV (Customer life value), number saving accounts, online banking and customer relationship.

One of the most used methods for analyzing customer value is RFM. RFM stands for the three dimensions of Recency, Frequency and Monetary. Ina Maryani and Dwiza Riana used the RFM model to determine potential customers and apply it to the CRM system (Customer Relationship Management) [3]. There are other authors who also make use of the RFM model. In [4], Mediana Aryuni et al. implemented clustering using methods K-Means and K-Medoids based on the RFM score of customer’s Internet Banking transactions. Ananthi Sheshasayee and L.Logeswari used use partitioning techniques, data mining tool and RFM parameters to segment customers. The resulting two groups were categorized based on frequency and the target customers would be offered some special offers to achieve the marketing goals of the organization [5]. An analysis with agglomerative, K-means and an advanced version of K-means clustering were carried out for RFM based market segmentation approach in [6]. Although aimed at direct marketing, Mohammad Amini et al. proposed a classification method which removes the imbalance of data using a combination of clustering and under-sampling. The results proved that their proposed method can improve the performance of the response models for bank direct marketing [7]. Contrary to Mohammad Amini, the authors in [8] explored the imbalance data in the banking context.

There are banks that have applied data mining technologies such as Hadoop, Spark for analyzing customer data. Shweta Yadav et al. analyzed the credit risk and the loan performance on data of the “Lending Club”, which is one of the biggest marketplaces for online credit in the Hadoop ecosystem [9].

While static customer information is frequently used, there is still a void in using dynamic customer information in studies with time-series data. In “Time-series clustering – A decade review” [10], Saeed Aghabozorgi et al. highlighted the importance and the need for clustering time-series datasets:

- Time-series datasets contain valuable information that can be obtained through pattern discovery.
- Time-series datasets are often very large and cannot be handled well by human inspectors.
- Time-series clustering is most used as an exploratory technique, and also as a subroutine in more complex data mining algorithms, such as rule discovery, indexing, classification, and anomaly detection.
- Representing time-series cluster structures as visual images (visualization of time-series data) can help users quickly understand the structure of data, clusters, anomalies, and other regularities in datasets.

In recent years, with the increasing data storage and
processing powers, realworld applications have been given the chance to store data for a longer time. Shi yang You et al. were interested in finding the clustering with potential value in financial time series. An experiment on ten-time segments showed that the obtained clusters were effective, in which both the whole similarity and the trend similarity on training data were markedly higher than that of randomized clustering [11]. Hanaa Talei et al. presented an end-to-end real-time architecture for analyzing and clustering time-series sensor data using an IoT architecture [12]. The authors used the Euclidian distance to compute the distance between the time-series and AHC to cluster time-series. Another distance also used for the time-series data is DTW. Weizeng Wang et al. proposed a new time series distance calculation method that integrates the characteristics of DTW and ED as a distance calculation for K-means clustering [13].

In banking, customer clustering is a typical problem. However, most of the clustering methods for this are using static customer information, which may not provide an up-to-date picture of the customer. This paper uses the customer’s balance history to cluster customers based on two types of distances: DTW and SDTW. This clustering will be the input for marketing, risk management processes in banks. A dataset, which is stripped of customers’ private information is available at Github [14]. It can also be used for other fields of research regarding time series and customers’ behavior.

II. DATA DESCRIPTION

The data set is taken from a bank’s deposit account information. For each account, there is data about balance history in 48 months from January 2013 to November 2017. The total number of accounts collected is 976,289. However, 95% of the accounts are no longer active or the balance is very low, under 350 USD, and will not be used in this study because of their low significance. The number of remaining accounts with high balance is 5,869. These accounts represent only 0.6% of the total but they contain 46.55% of the total balance of all deposit accounts.

For privacy reasons, all customer information and bank information have been discarded. Only the account balance details remain. Hence, the data for each customer is a sequence of 48 values, each being the balance at the beginning of the month. Fig. 1 shows the balance history of three customers.

III. IMPLEMENTATION AND EVALUATION

A. Preprocessing

First, the original data is smoothed using the spline function to reduce sharp variations (Fig. 2). Next, the data is normalized to the range of [0,1] (Fig. 3).

The data corresponding to each account is a high-dimensional vector, reduced to two dimensions using principal component analysis (PCA). All the data can be observed as shown in Fig. 4.

The area with low density is undesirable. So, the KNN_SUM function, which returns the sum of distance to k-nearest objects, is used to filter out all anomalies [15] (Fig. 5).

From the result of the KNN_SUM function, at about 1300 and above, the outlier score don’t not change much. By taking the threshold of 1300, the distribution in 2D does not
lose too much data. The data distribution after filtering is more appropriate as all points in the sparse area were removed (Fig. 6). This preprocessed data will be the input of the next clustering process.

Fig. 5. Graph of the KNN_SUM function.

Fig. 6. Data distribution in 2D after removing outlier.

B. Methods

The experiments conducted in this section will use both methods, DTW and SDTW, and the time interval between them are used for comparison. Clustering strategies used are Partitional and Hierarchical. Finally, performance evaluations are based on four indicators, Silhouette, SF, DB, and elapsed time for clustering.

**DTW and SDTW.** Time-series distance can be divided into three types: Model-free, Shape-based and Feature-based. The DTW belongs to Shape-based class and is popular for measuring time series distance. To overcome the limitation of comparing directly two points between time-series and to be able to deal with local transformation such as warping, shifting and different series lengths, the most popular distance is DTW (Dynamic Time Warping) proposed by Donald J. Berndt and James Clifford in 1994 [16]. Although it overcomes the weakness of local transformation, it is affected by the scale of the two-time series. In “Soft-DTW: a Differentiable Loss Function for Time-Series”, M. Cuturi and M. Blondel proposed Soft-DTW (SDTW), a differentiable loss function, with both its value and gradient computed in quadratic time/space complexity (DTW has quadratic time but linear space complexity) [17]. The SDTW distance can return a negative value and the SDTW of X series and itself is not always equal zero. The SDTW is a promising method for clustering problem following the author’s experiment in [17].

**Clustering strategies.** In time-series clustering, two popular strategies are partitional (e.g. K-mean, K-medoids) and hierarchical (e.g. Agglomerative Hierarchical Clustering). In the partitional strategy, the dataset is decomposed into k partitions, where each partition represents a cluster. The cluster is then constructed by optimizing an objective partition criterion. The most well-known algorithms are K-mean and K-medoids. The partitional algorithms can be the basis for proposing a new model like HMM (Hidden Markov Model) [18] or optimize the result of clustering [19]. They also prove effective for video analysis in [20], image clustering [21] or epileptic seizure Detection [22]. A partitional strategy has some constraints such as the number of clusters and initial centroids. Its effectiveness also depends on the shape of the cluster, which is most effective in spherical shaped clusters. Hierarchical clustering is an approach of cluster analysis which makes a hierarchy of clusters using agglomerative or divisive algorithms. The agglomerative algorithm considers each item as a cluster, and then gradually merges the clusters (bottom-up) [11]. The hierarchical clustering method is weak in terms of computation and cannot be back-tracked after merging (bottom-up) or splitting (top-down) clusters.

**Evaluation indexes.** Three indexes, Silhouette, DB (Davies-Bouldin) and SC (Score Function), are used to evaluate the ability of the clustering process. The clustering is optimal at the maximum Silhouette value [23] and the minimum SF [24] and DB values [25].

**C. Clustering Evaluation**

The experiments conducted in this section will use both methods to calculate the time interval between DTW and SDTW. Clustering strategies used are both Partitional (PAR) and Hierarchical (HIERAR). Two different centroid computation methods are explored: Partition Around Medoid (PAM) and DTW Barycenter Averaging (DBA) [26]. Performance evaluations are based on the three indexes, Silhouette, SF and DB.

The number of clusters in the study is limited to 8 so that it can be meaningful in practice. The results are shown in the following graphs.

Fig. 7. Silhouette value.

Fig. 8. DB value.
In Fig. 7, the SDTW distance has the best Silhouette index with 4 as the number of clusters. The Silhouette value will decrease slightly when the number of clusters increases. As opposed to the Silhouette value, the DB value vary differently with the number of cluster and different methods (Fig. 8). For example, the DB value of SDTW using the AHC strategy increases fast when number of clusters goes from 3 to 5 and decreased from 6. With the SF value, for easy visualization, the logarithmic scale is used (Fig. 9). SF values are almost unchanged when using the SDTW distance calculation method but will increase rapidly when using DTW.

From the analyses of the collected data set, the clustering is most optimal with the SDTW distance, the partitional clustering strategy and four clusters. The number of time series in the four clusters obtained with this method are: 1483, 1125, 512 and 2132.

Fig. 10 shows the patterns and centroids of each cluster. The centroid lines of the different clusters show that clustering is good. Next, the pre-standardized values of the accounts are restored and the average and variance of each cluster by month are calculated. Fig. 11 shows that the average of the monthly account balances of the groups is relatively equal, but the variance is quite different, especially between groups 1, and 2 & 3.

The total balance of all accounts is shown in Fig. 12 and is calculated according to the cluster in Fig. 13. Using the ARIMA method to model time series [27], the results are as shown in Table I. The total balance of all accounts has a complex structure with many peaks. The modeling results with ARIMA obtained \((p, d, q) = (0,1,1)\) mean that the next month's total balance depends only on the previous month. However, when divided into 4 clusters, the structure of the sub-series is simpler, especially in clusters 1 and 3. Clusters also have a more detailed ARIMA model, meaning that the

| TABLE I: AUTO ARIMA MODEL | p | d | q |
|---------------------------|---|---|---|
| CLUSTER 1                 | 1 | 2 | 1 |
| CLUSTER 2                 | 0 | 2 | 1 |
| CLUSTER 3                 | 2 | 2 | 1 |
| CLUSTER 4                 | 1 | 1 | 0 |
| ALL                       | 0 | 1 | 1 |
next month's balance will depend on many factors or will have more bases.

IV. CONCLUSION AND PERSPECTIVES

The paper proposes a new method for clustering customers based on the history of account balances and demonstrates the effectiveness of the method based on clustering data as well as statistics on real data. The result of customer clustering can be used in marketing strategies, forecasting of affordability, or customer deposits in the following month, etc. The paper is also a valuable reference for problems in areas such as Knowledge Representation [28], [29], Prediction [30], Time Series Analysis [31], etc. It can also be applied when dealing with time series data such as biomedical signals, insurance expenditures, etc.

CONFLICT OF INTEREST

The authors declare that they have no conflicts of interest in the subject matter or materials discussed in this manuscript.

AUTHOR CONTRIBUTIONS

Both authors discussed ideas and solutions for the problem; Do Quang Dat programmed, and Phan Duy Hung wrote the paper; all authors had approved the final version.

REFERENCES

[1] J.-Y. Han, L.-Q. Mei., and T. Wang, “Customer value management based on AHP and k-means clustering method,” in Proc. Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 2010, pp. 2216-2221.

[2] Y. Gu, “The impact of reward programs on customer behavior of commercial bank,” in Proc. Third International Symposium on Information Processing, Washington, 2010, pp. 110-112.

[3] I. Maryani and D. Riana, “Clustering and profiling of customers using RFM for customer relationship management recommendations,” The 5th International Conference on Cyber and IT Service Management (CITSM), Denpasar, 2017, pp. 1-6.

[4] M. Aryuni, E. D. Madyatmadja, and E. Miranda, “Customer segmentation in XYZ Bank using K-means and K-medoids clustering,” in Proc. The International Conference on Information Management and Technology (ICIMTech), Jakarta, 2018, pp. 412-416.

[5] Sheshasayee and L. Logeshwari, “Implementation of clustering technique based RFM analysis for customer behaviour in online transactions,” in Proc. The 2nd International Conference on Trends in Electronics and Informatics (ICOEI 2018), Tamil Nadu, 2018, pp. 1166-1170.

[6] S. H. Shihab, S. Afroege, and S. Z. Mishu, “RFM based market segmentation approach using advanced K-means and agglomerative clustering: A comparative study,” in Proc. International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, 2019, pp. 1-4.

[7] M. Amini, J. Rezaeenour, and E. Hadavandi, “A cluster-based data balancing ensemble classifier for response modeling in bank direct marketing,” International Journal of Computational Intelligence and Applications, vol. 14, pp. 1-23, Dec. 2015.

[8] V. L. Migueils, A. S. Camanho, and J. Borges, “Predicting direct marketing response in banking: comparison of class imbalance methods,” Springer, vol. 11, no. 4, pp. 831-849, January 2016.

[9] S. Yadav and S. Thakur, “Bank loan analysis using customer usage data: a big data approach using Hadoop,” in Proc. the 2nd International Conference on Telecommunication and Networks (TEL-NET), Noida, 2018, pp. 1-8.

[10] S. Aghabozorgi and A. S. Shirkhorshidi, “Time-series clustering – A decade review,” Information System, vol. 53, pp. 16-38, May 2015.

[11] S. Y. You, Y. D. Wang, L. K. Luo et al., “Finding the clusters with potential value in financial time series based on agglomerative hierarchical clustering,” in Proc. The 11th International Conference on Computer Science & Education (ICCSE 2016), Nagoya, 2016, pp. 77-81.

[12] H. Talei, M. Essaaidi, and D. Benhaddou, “An end to end real time architecture for analyzing and clustering time series data: Case of an Energy management system,” in Proc. The 6th International Renewable and Sustainable Energy Conference (RESEC), Rabat, 2018, pp. 1-7.

[13] W. Wang, G. Lu, Y. Shi et al., “Time series clustering based on dynamic time warping,” in Proc. the IEEE 9th International Conference on Software Engineering and Service Science, Beijing, China, 2018, pp. 487-490.

[14] GITHUB. [Online]. Available: https://github.com/ziczacziczac/customer-clustering/tree/master/data

[15] The R Library for Distance and Density-Based Outlier Detection: A big data approach using Hadoop, “in Proc. the 2nd International Conference on Data Mining in Pattern Recognition, pp. 678-693, vol. 44, 2011, [Online]. Available: https://cran.r-project.org/web/packages/forecast/forecast.pdf

[16] F. Petitjean, A. Ketterlin, and P. Gancarski, “A global averaging method for dynamic time warping, with applications to clustering,” Pattern Recognition, pp. 243-256, vol. 46, no. 1, 2013.

[17] J. Santarcangelo, X-P. Zhang, “Dynamic time-alignment K-means kernel clustering for time sequences,” in Proc. the IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015, pp. 2532-2536.

[18] Y. Zhao, W. Lu, and Y. Meng, “Landsat Time Series Clustering under Modified Dynamic Time Warping,” in Proc. the Fourth International Workshop on Earth Observation and Remote Sensing Applications, Guangzhou, China, 2010, pp. 135-138.

[19] X. Yao and H.-L. Wei, “Improving K-means clustering performance using a new global time-series averaging using a new global time-series averaging,” in Proc. International Conference on Knowledge Discovery in Databases, Washington, 1994, pp. 359-370.

[20] M. Cuturi and M. Blondel, “Soft-DTW: A differentiable loss function for time-series,” arXiv preprint arXiv:1703.01541, 2017.

[21] L.-L. Wei and J.-Q. Jiang, “A hidden Markov model-based K-means time series clustering algorithm,” in Proc. The IEEE International Conference on Intelligent Computing and Intelligent Systems, Xiamen, China, 2010, pp. 135-138.

[22] S. H. Shihab, S. Afroege, and S. Z. Mishu, “RFM based market segmentation approach using advanced K-means and agglomerative clustering: A comparative study,” in Proc. International Conference on Electrical, Computer and Communication Engineering (ECCE), Cox's Bazar, 2019, pp. 1-4.

[23] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, pp. 53-65, vol. 20, 1987.

[24] S. Saitta, B. Raphael, and I. F. Smith, “A bounded index for cluster segmentation approach using advanced K-means and agglomerative clustering: A comparative study,” in Proc. Ninth International Conference on Machine Learning and Cybernetics, Qingdao, 2010, pp. 2216-2221.

[25] J. Santarcangelo, X-P. Zhang, “Dynamic time-alignment K-means kernel clustering for time sequence,” in Proc. the IEEE International Conference on Image Processing (ICIP), Quebec City, QC, Canada, 2015, pp. 2532-2536.

[26] Y. Zhao, W. Lu, and Y. Meng, “Landsat Time Series Clustering under Modified Dynamic Time Warping,” in Proc. the Fourth International Workshop on Earth Observation and Remote Sensing Applications, Guangzhou, China, 2010, pp. 135-138.

[27] S. B. Muhammad, K. Fatih, and P. Kemal, “K-Means Clustering Based Time Series weighting with Epileptic Seizure Detection,” in Proc. Electric Electronics, Computer Science, Biomedical Engineering's Meeting, 2018, pp. 1-4.

[28] P. J. Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis,” Journal of computational and applied mathematics, pp. 53-65, vol. 20, 1987.

[29] S. Saitta, B. Raphael, and I. F. Smith, “A bounded index for cluster validation,” in Proc. International Workshop on Machine Learning and Data Mining in Pattern Recognition, 2007, pp. 174-187.

[30] O. Arbelaiz, I. Gururutxaga, J. Mugaerza, J. M. Perez & I. Perona, “An extensive comparative study of cluster validity indices,” Pattern Recognition, pp. 243-256, vol. 46, no. 1, 2013.

[31] F. Petitjean, A. Ketterlin, and P. Gancarski, “A global averaging method for dynamic time warping, with applications to clustering,” Pattern Recognition, pp. 678-693, vol. 44, 2011.

[32] The R library used for analysing univariate time series forecast. [Online]. Available: https://cran.r-project.org/web/packages/forecast/forecast.pdf

[33] P. D. Hung, N. T. T. Lien, and N. D. Ngoc, “Customer Segmentation Using Hierarchical Agglomerative Clustering,” in Proc. The 2nd International Conference on Information Science and Systems (ICISS 2019), ACM, New York, NY, USA, pp. 33-37.
[29] P. D. Hung, N. D. Ngoc, and T. D. Hanh, “K-means Clustering Using R A Case Study of Market Segmentation,” in Proc. The 5th International Conference on E-Business and Applications (ICEBA 2019), ACM, New York, NY, USA, pp. 100-104.

[30] P. D. Hung, T. D. Hanh, and V. T. Diep, “Breast cancer prediction using spark MLlib and ML packages,” in Proc. The 5th International Conference on Bioinformatics Research and Applications (ICBRA ’18), ACM, New York, NY, USA, pp. 52-59.

[31] P. D. Hung, “Estimating confident index of respiratory signal using an accelerometer,” in Proc. the IEEE 2nd International Conference on Signal and Image Processing (ICSIP), Singapore, 2017, pp. 413-417.

Copyright © 2020 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited (CC BY 4.0).

Phan Duy Hung received his Ph.D. degree from INP Grenoble France, in 2008. Since 2009, he has worked as a lecturer, the head of a department, director of master program in software engineering at FPT University in Hanoi, Vietnam. His current research interests include digital signal and image processing, internet of things, big data, artificial intelligence, measurement and control systems and industrial networking.

Do Quang Dat received his B. Sc from VNU University of Science, Vietnam. Now he works as a software expert in the fields of Software engineering and Data Science at Viettel Group. He is currently a master student at FPT University, Hanoi, Vietnam. His current research interests include internet of things, artificial intelligence.