Data Mining for Predicting Customer Satisfaction Using Clustering Techniques

Kartika Purwandari
Bioinformatics & Data Science Research Center
Bina Nusantara University
Jakarta, Indonesia 11480
kartika.purwandari@binus.edu

Join W. C. Sigalingging
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
joinwanchanlyn10@gmail.com

Muhammad Fhadli
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
muhammadfhadli20@gmail.com

Shinta Nur Arizky
Department of Computer Science and Information Engineering
National Central University
Taoyuan, Taiwan
shinta.nuraisya@gmail.com

Bens Pardamean
Computer Science Department, BINUS Graduate Program – Master of Computer Science
Bina Nusantara University
Jakarta, Indonesia 11480
bpardamean@binus.edu

***Abstract*— Managing customer satisfaction has become an important business trend, including restaurants business. This study aims to determine the application of the K-means, Spectral Clustering (SC), and Agglomerative Clustering (AC) method for measuring customer satisfaction on a family restaurant in Taiwan. We contribute the data collection process and application of data mining in a family restaurant. The clustering analysis based on agglomerative clustering approach performs as well as the K-means approach to cluster the same characteristics of the customers. At last, this study shows the measurement result of customer satisfaction and provides improvement suggestion to the restaurant concerned.**

Keywords*— agglomerative clustering, customer satisfaction, data mining, K-means, spectral clustering

I. INTRODUCTION

Restaurant industry emerges as one of the most important contributing business sectors in Asia [1]. Ranging from a small, cheap street restaurant to glamorous, expensive hotel restaurants or shopping malls, Asian-style restaurant business also have spread across the world. A wide range of new variants of food come out almost on daily basis. In this regard, in order to stand out in the intense business competitions, the restaurant's reliance on the quality of customer services needs to be maintained continuously [2]. Service preferences on very narrow segments of demography, for example, the middle age or elderly people, would put the restaurant business at risk. This shows that restaurants needs to be wise in assessing reviews from the customer to improve its business model and win the competition or at least to survive the fight.

Managing customer satisfactions has become a key issue for different types of businesses [3], particularly restaurants [4]. This would allow the restaurant to know which aspect they need to boost and to keep them fighting in the competition itself. There are a number of attributes that we can use to determine consumer satisfactions, e.g. food quality, price, and much more [5]. The choice of features also plays an important role in managing customer satisfaction because it will affect the outcome of the analysis.

Data mining is currently an option for researchers to extract information from the data and make an analysis [6]. numerous algorithms used in the data mining can help scientists to interpret patterns in the data, analyze it, and infer a phenomenon. This technique has been used to tackle several day-to-day problems in industries such as pharmacy, management, and safety [7]. Data mining systems rely on raw input databases and this causes problems, for example, database tends to be dynamic, incomplete, noisy, and large. Some missing data, outliers, and over fitting are important to implement data mining issues [8].

In this study, we examined customer service at a restaurant in Taiwan using data mining by clustering technique. We use clustering technique to read pattern in customers' opinion about the restaurant based on some measurement variables. Principal Component Analysis (PCA) was implemented to project highly dimensional data (number of question) into the lower dimension. We also looked for an attribute that played an important role in the results. This work is very relevant because maintaining customer service is something we seldom do in our everyday lives. We assessed how happy the customers were and provided input to the restaurant for their progress.

II. BACKGROUND

A. Predicting Using Clustering Techniques

Clustering is one of the unsupervised learning methods that is often applied in the area of text data mining. The idea of clustering is to do the data segmentation and gather them into partitions [9]. From a functional point of view, this method plays an important role in the applications of data mining such as scientific data discoveries. Data mining brings to clustering the complexities of the large datasets with various types of attributes [10]. This unsupervised method for clustering homogeneous elements (in this case restaurant data) into clusters may be done if we use information from the class to analyze the received clusters [11].

B. Data Mining on Restaurant Data

With the extensive use of digital information systems, a vast amount of time series can be obtained during daily business activities. Liu et al [12] demonstrate the application of time series in data mining by taking fast-food restaurants as the research object. Their work helps the franchise reap the reward.

978-1-7281-7071-8/20/$31.00 ©2020 IEEE 13-14 August 2020
2020 International Conference on Information Management and Technology (ICIMTech)

Data mining for recommending restaurant based on GPS technology was already implemented by Lee et al. That system required the location and the environment for the context of their model [13]. Recommendations from the system can help the mobile users come to the appropriate restaurants [14].

C. Data Mining for Predicting Customer Satisfaction

As stated by Hogan et al on Bayu's study [15], the ability of an enterprise to acquire and manage customer information is a key to maintaining a competitive advantage; however, managing customer relationships is difficult to achieve because the customer, either individually or in a group, has different preferences and expectations. Therefore, in order to overcome such problems, an information-based marketing strategy is needed to provide continuous responsive decision-making.

III. METHODOLOGY

A. Data Collection

This research is a quantitative study conducted in a family restaurant located at Zhongyang Lane, Zhongli District, Taoyuan Area. This restaurant is a family restaurant that was established on year 2000. The data used in this study were primary data obtained from respondents as many as 104 respondents. Respondents in this study were all visitors to the restaurant chosen randomly. Questionnaire distribution was carried out for three months, namely from October to December 2018. There are six key aspects addressed in the questionnaire that are measured in this research: reliability, tangibles, information system, responsiveness, empathy, and assurance. This process helps respondents to understand the detail of the measurement and gives more real feedbacks to our questionnaire. Table 1 shows the detail question and their corresponding aspects. During the data collection, each customer only obtained one piece of paper containing 16 questions to assess the quality of the restaurant on the basis of a number of features as shown in Fig. 1. The unit of measurement of variables in this study employs a 5 (five) point scale, namely 1 (poor), 2 (fair), 3 (satisfactory), 4 (good), and 5 (excellent) except for question 14, 15, and 16. We kept out answers of question number 14 in our analysis since they do not reflect the six aforementioned aspects. Moreover, question number 15 and 16 are measured using Boolean value.

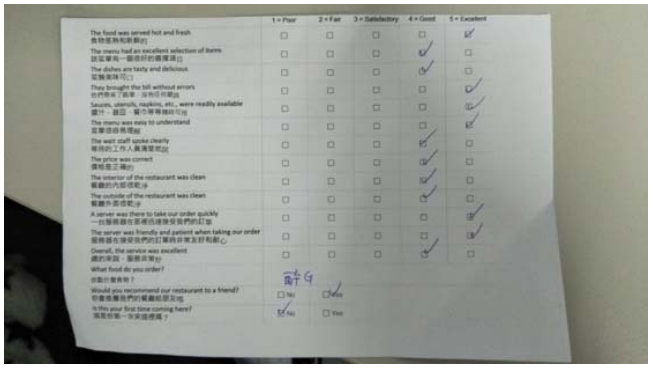

Fig. 1. Example of Survey Paper

TABLE I. CUSTOMER SATISFACTION'S ASPECT

No	Question	Aspect
1	The food was served hot and fresh	Reliability
2	The menu had an excellent selection of items	Reliability
3	The dishes are tasty and delicious	Reliability
4	They brought the bill without errors	Tangibles
5	Sauces, utensils, napkins, etc., were readily available	Tangibles
6	The menu was easy to understand	Information System
7	The wait staff spoke clearly	Responsiveness
8	The price was correct	Tangibles
9	The interior of the restaurant was clean	Tangibles
10	The outside of the restaurant was clean	Tangibles
11	A server was there to take our order quickly	Empathy
12	The server was friendly and patient when taking our order	Empathy
13	Overall, the service was excellent	Reliability
14	What food do you order?	None
15	Would you recommend our restaurant to a friend?	Assurance
16	Is this your first time coming here?	Assurance

B. Pre-processing

All the data on the document was moved to a CSV file with some null-values on the data. The answers to question number 1 to 13 were converted to a number from 1 to 5 that show the customer satisfaction level. Question number 14 was skipped because of the type of answer is "String" and it would not be processed for the next step. Questions number 15 and number 16 were translated to 0 for answer "No" and 1 for answer "Yes".

C. Clustering Model

We used three models to analyze our pre-processed data, namely K-means Clustering, Spectral Clustering (SC), and Agglomerative Clustering (AC). The Principal Component Analysis (PCA) was implemented for the data visualization on every clustering result. PCA basically aims to simplify the observed variables by reducing (reducing) their dimensions. After obtaining several components of PCA results that are free from multicollinearity, these components become new variables that will be analyzed using cluster analysis according to the clustering objective function [16].

K-means is one of the most common clustering algorithms due to a simple and easy implementation, and has been used for many difficult tasks such as image segmentation [17]. The characteristic of K-means model is calculating the centroid of each cluster use the minimum value of squared errors as shown in equation 1. This process will be repeated until local minima was trapped [16].

$$J_K = \sum_{k=1}^{K} \sum_{i \in C_k} (x_i - m_k)^2 \quad (1)$$

where $(x_i, \ldots, x_n) = X$ as a data matrix of variable from the question and $m_k = \sum_{i \in C_k} x_i / n_k$ is the centroid of cluster C_k and n_k is the number of points in C_k.

Another efficient clustering algorithm is the spectral clustering In several occasions, spectral clustering surpasses the efficiency of K-means and operates efficiently by using only the regular linear algebra application [18]. The main idea of Spectral Clustering is to use eigenvectors from the matrix data. The use of eigenvector helps a lot in empirical successes of this approach. Assuming that we have s set points of $P =$

$\{p_1, \ldots, p_n\}$ with k cluster to be built, we need to compute the similarity matrix P. We utilize a similarity matrix to create similarity graph and compute unnormalized Laplacian L. Next, k eigenvectors are obtained from L. A matrix $X = \{x_1, \ldots, x_k\}$ is extracted by putting the eigenvectors into stacks. Afterward, the renormalized matrix of X denoted by Y can be obtained from the following equation

$$Y_{ij} = \frac{x_{ij}}{\sqrt{\sum_j x_{ij}^2}}. \quad (2)$$

Finally, clustering the matrix Y using K-means and hence we can assign P to the selected cluster [19]

Agglomerative clustering is well known for handling data with a collection of data points where each cluster will be clustered into a larger cluster until all clusters get to be one cluster [20]. If we denote N(x) as the set of vertices neighboring vertex x in a graph, a vertex y is "similar" to the vertex x if N(x) and N(y) have a large overlap. The agglomerative have calculated based on the similarity $\sigma(x, y)$ between vertices x and y by equation (3):

$$\sigma(x,y) = \begin{cases} \frac{N(x) \cap N(y)}{N(x) \cup N(y)}, if \mid N(x) \cup N(y) \mid > 0 \\ 0, \qquad otherwise \end{cases} \quad (3)$$

The similarity between two vertices lies in the range [0…1]: 0 if the vertices have no neighbors in common, and 1 if they have exactly the same neighbors [21]

IV. Result and Analysis

From this survey, 104 questionnaires were collected with 13 integer features and two Boolean features. There are many ambiguous responses from the respondent to the questionnaire. To solve this problem, the missing values were filled with the most frequent data [22]. Table 2 shows the size of the cluster in each method. In case of scale, we can see that K-means and Agglomerative Clustering (AC) divides the data equally enough between each cluster. On the other hand, Spectral Clustering (SC) have put less data into cluster 2 and 3.

TABLE II. Cluster size

Clustering Model	Cluster 1	Cluster 2	Cluster 3	Total
K-means	10	42	52	104
Spectral Clustering (SC)	100	3	1	104
Agglomerative Clustering (AC)	51	50	3	104

Fig. 2 shows the visualization of the resulting clustering using K-means with the first two principal components. Visually, the results of the data are already clustered well as we can observe from the fairly even distribution of the data in each cluster. Our finding is consistent with a statement of Teknomo in [23] that shows that the K-means method is able to group a large number of data points faster based on the number of predetermined groups.

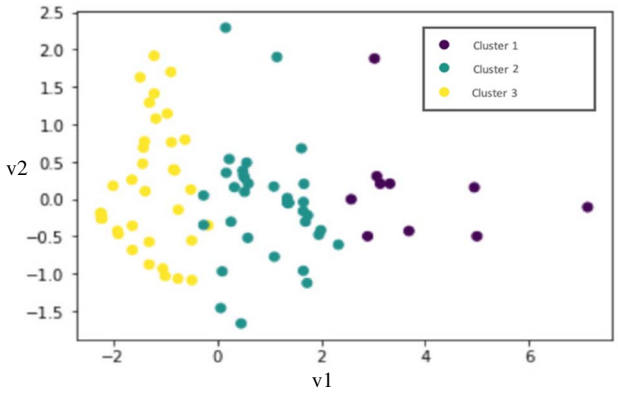

Fig. 2. Clustering Result using K-means

The evaluation of a cluster-predict methodology can be achieved on the basis of the different k clusters. We utilized k = 3 for our experiment. Fig. 3 displays the visualization of clustering results using Spectral Clustering using PCA. We plotted the samples in the first two principal components. Clustering results from cluster 1 provides the dominant results from the results of other clusters that means only one data point grouped into cluster 3 and only three points grouped into cluster 2.

The abstract model of the data distribution is not compulsory for a spectral clustering; however, it is important for a spectral analysis of the matrix of point-to-point similarities. Using spectral clustering, it can automatically measure the size and number of classes and can accommodate multi-scale data [24].

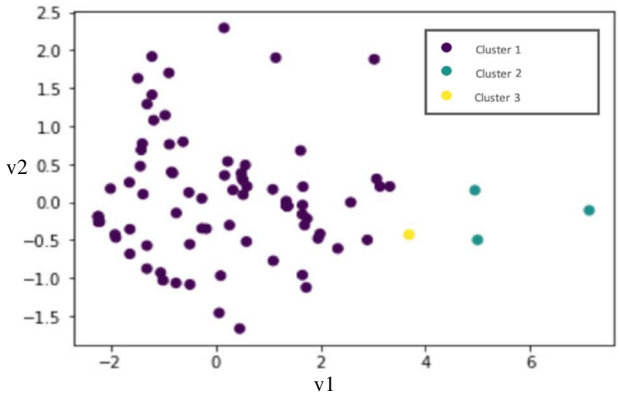

Fig. 3. Clustering Result using Spectral Clustering

The visualization of the clustering results from the agglomerative method using the first two principal components is shown in Fig. 4. Our findings suggest that clustered effects often appear equally distributed as the findings of the K-means clustering process. The clustering analysis based on this approach performs as well as the K-means approach that can be seen in cluster 1 and cluster 2. Although the performance seems plausible for the first two clusters, the agglomerative clustering method only gathers three data points to form cluster 3. However, it still suggests that the agglomerative method is proven to work most effectively in general-purpose configurations offered by modern standard applications [25].

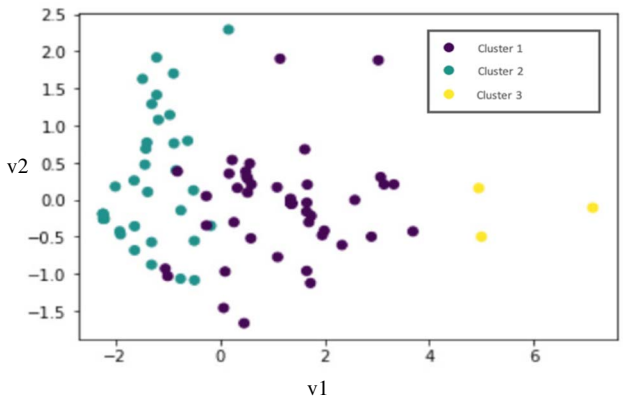

Fig. 4. Clustering Result using Agglomerative Clustering

TABLE III. CLUSTERING RESULTS

No	Question	K-means			SC			AC		
		1	2	3	1	2	3	1	2	3
1	The food was served hot and fresh	3	4	5	4	4	3	4	5	3
2	The menu had an excellent selection of items	3	4	5	4	3	4	4	5	3
3	The dishes are tasty and delicious	3	4	5	4	3	4	4	5	3
4	They brought the bill without errors	4	4	5	5	3	4	4	5	3
5	Sauces, utensils, napkins, etc., were readily available	4	4	5	5	3	4	4	5	3
6	The menu was easy to understand	3	4	5	4	**2**	4	4	5	**2**
7	The wait staff spoke clearly	3	4	5	4	3	3	4	5	3
8	The price was correct	3	4	5	4	**2**	3	4	5	**2**
9	The interior of the restaurant was clean	3	4	5	4	**2**	3	4	5	3
10	The outside of the restaurant was clean	3	4	5	4	3	3	4	5	3
11	A server was there to take our order quickly	3	4	5	5	3	3	4	5	3
12	The server was friendly and patient when taking our order	3	4	5	5	3	3	4	5	3
13	Overall, the service was excellent	3	4	5	5	3	3	4	5	3
14	Would you recommend our restaurant to a friend?	1	1	1	1	1	1	1	1	1
15	Is this your first time coming here?	0	0	0	0	0	0	0	0	0

Table 3 was calculated from the mean of grouped data from each cluster. As can be seen in Table 3, K-m\means method is capable of putting the data with the most score 3-4 into cluster 1, score 4 into cluster 2, and score 5 into cluster 3. The performance of the SC method still has a range of 2 to 4 in the cluster 2, and does not appeared score 5 in cluster 2 and cluster 3. The AC method, still have score 2 in the cluster 3, but cluster 1 can get a score of 4 for all questions, and all score 5 for cluster 2.

Moreover, qquestion number 1 to 5, 7, and 10 to 13 earned score with a range of 3-5 as the most preferred value. This implies that the majority of customers are already satisfied with the restaurant's current service. To further improve the quality of the customer service, the restaurant needs to evaluate the variables in question 6, 8, and 9, since several customers gave them a score of 2. For the other aspect, the performance of the restaurant is not bad, although some customers do want some changes, e.g. the menu information, the cleanliness of the restaurant outside and the communication skills of the staff.

V. CONCLUSSION

Data mining process using three clustering methods is able to identify which aspects of the restaurant that satisfy the expectation of the customers as well as aspects that need to be improved from the restaurant. The results of this study can effectively provide suggestions and input on the priority of developing and improving restaurants in the future. In the future study, to improve the quality of clustering result, we can employ a strategy to gather the data using online forms to fill in the questionnaires so that it can reduce missing values of certain questions [26]. Further, samples selection and designing a more relevant questionnaire aiming for a particular type of restaurant business models will be helpful to improve the proposed models

VI. RESEARCH PART

K.P. contributed to interpreting and expanding theory according to the analyzed data and structured the literature review. M.F., J.W.S., and S.N.A. conducted data computation and data analysis. B.P. supervised and verified the results of this work. All authors actively discussed the final results and contributed accordingly to the manuscript writing.

ACKNOWLEDGMENT

This study was supported by "27" Restaurant in Zhongyang Lane, Zhongli District, Taoyuan Area, Taiwan.

REFERENCES

[1] H. V. Boo, Service Environment of Restaurants: Findings from the Youth Customers, Journal of ASIAN Behavioural Studies, vol. 2(2), pp. 67-77, 2017.

[2] S. Auty, Consumer Choice and Segmentation in the Restaurant Industry, The Service Industries Journal, vol. 12(3), pp. 324-339, 1992.

[3] Zairi M. Managing customer satisfaction: a best practice perspective. The TQM magazine, 2000 Dec.

[4] B. Adhi Tama, Data Mining for Predicting Customer Satisfaction in Fast-food Restaurant, Journal of Theoretical and Applied Information Technology, vol. 75(1), pp. 18-24, 2015.

[5] J. Hanaysha, Testing the Effects of Food Quality, Price Fairness, and Physical Environment on Customer Satisfaction in Fast Food Restaurant Industry, Journal of Asian Business Strategy, vol. 6(2), pp. 31-40, 2016.

[6] K. V. R. and R. K., A Study on Application of Spatial Data Mining Techniques for Rural Progress, CoRR, vol. abs/1303.0447, 2013.

[7] A. Jain, G. Hautier, S. P. Ong, and K. Person, New Opportunities for Materials Informatics: Resources and Data Mining T echniques for Uncovering Hidden Relationship, Journal of Materials Research, vol. 31(8), pp. 977-994, 2016.

[8] R R, B S, Sofia VSA. Data Mining Issues and Challenges: A Review. Ijarcce, vol. 7(11), pp. 118-121, 2018.

[9] Budiarto A, Mahesworo B, Baurley J, Suparyanto T, Pardamean B. Fast and Effective Clustering Method for Ancestry Estimation. Procedia Computer Science. Vol. 1(157), pp. 306-12, 2019 Jan.

[10] Aggarwal, Charu C., and ChengXiang Zhai, eds. Mining text data. Springer Science & Business Media, 2012.

[11] Lopez MI, Luna JM, Romero C, Ventura S. Classification via clustering for predicting final marks based on student participation in forums. International Educational Data Mining Society. 2012 Jun.

[12] Berkhin, Pavel. "A survey of clustering data mining techniques." In Grouping multidimensional data, pp. 25-71. Springer, Berlin, Heidelberg, 2006.

[13] Lee, Bae-Hee, Heung-Nam Kim, Jin-Guk Jung, and Geun-Sik Jo. "Location-based service with context data for a restaurant recommendation." In International Conference on Database and Expert Systems Applications, pp. 430-438. Springer, Berlin, Heidelberg, 2006.

[14] Lee, Bae-Hee, Heung-Nam Kim, Jin-Guk Jung, and Geun-Sik Jo. "Location-based service with context data for a restaurant recommendation." In International Conference on Database and Expert Systems Applications, pp. 430-438. Springer, Berlin, Heidelberg, 2006.

[15] Tama BA. Data mining for predicting customer satisfaction in fast-food restaurant. J Theor Appl Inf Technol, vol. 75(1), pp. 18-24, 2015.

[16] Ding, Chris, and Xiaofeng He. "K-means clustering via principal component analysis." In Proceedings of the twenty-first international conference on Machine learning, p. 29, 2004.

[17] N. Dhanachandra, K. Manglem, and Y. J. Chanu, Image Segmentation using K-means Clustering Algorithm and Subtractive Clustering Algorithm, Eleventh International Conference on Image and Signal Processing, ICISP 2015 (August 21-23, 2015).

[18] U. von Luxburg, A Tutorial on Spectral Clustering, CoRR, vol. abs/0711.0189, 2007.

[19] Von Luxburg U. A tutorial on spectral clustering. Stat Comput, vol. 17(4), pp. 395-416, 2007.

[20] I. Crawford, S. Ruske, D. O. Topping, and M. W. Gallagher, Evaluation of Hierarchical Agglomerative Cluster Analysis Methods for Discrimination of Primary Biological Aerosol, Atmospheric Measurement Techniques, vol. 8(11), pp. 4979-4991, 2015.

[21] Beeferman D, Berger A. Agglomerative clustering of a search engine query log. Proceeding Sixth ACM SIGKDD Int Conf Knowl Discov Data Min, pp.407-416, 2000.

[22] Shyu, M-L., Indika Priyantha Kuruppu-Appuhamilage, S-C. Chen, and L. Chang. "Handling missing values via decomposition of the conditioned set." In IRI-2005 IEEE International Conference on Information Reuse and Integration, Conf, 2005., pp. 199-204. IEEE, 2005.

[23] Lathifaturrahmah L. Perbandingan Hasil Penggerombolan K-Means, Fuzzy K-Means, dan Two Step Clustering. Jurnal Pendidikan Matematika. Vol. 2(1), pp. 39-62, 2017 Apr.

[24] Zelnik-Manor, Lihi, and Pietro Perona. "Self-tuning spectral clustering." In Advances in neural information processing systems, pp. 1601-1608, 2005.

[25] Müllner, D. (2011). Modern hierarchical, agglomerative clustering algorithms. arXiv preprint arXiv:1109.2378.

[26] Toy, A. and Supriyanti, W.. Evaluasi Usability Aplikasi Jadwal Terpadu Universitas Muhammadiyah Surakarta Dengan Metode Kuisioner. SEMNASTEKNOMEDIA ONLINE. Vol. 2(1), pp. 1-10, 2014.