COMBINATION OF K-MEANS CLUSTERING AND K-NEAREST NEIGHBOR ON ECOMMERCE CUSTOMER SPENDING RATE PREDICTION

Boni Oktaviana Sembiring [1], Eka Rahayu [2], Mufida Khairani [3], Septiana Dewi Andriana [4], Arie Rafika Dewi [5], Khairunnisa [6]
[1,2,3,4,5,6] Universitas Harapan Medan
bonioktaviana@yahoo.co.id, eka.r0041@gmail.com, mufida.khairani@gmail.com, septianad89@gmail.com, arie.juny@gmail.com, khairunnisajv2@gmail.com

Abstract

Article Info
Received 10 June 2021
Revised 20 June 2021
Accepted 30 June 2021

K-Nearest Neighbor is a classification method that classifies new data into specific classes based on the proximity of characteristics to k members of existing classes. K-Nearest Neighbor relies heavily on training data. In actual circumstances such as the ecommerce customer spending rate dataset, there is no class label for each data. So that to be able to obtain datatraining required additional methods need to be added before the prediction process can be done. This research attempts to use K-Means Clustering to group datasets into multiple clusters which then each cluster will be given a class label according to the centroid characteristics of those clusters. The combination of KNN and K-Means Clustering methods in customer's spending rate predictions gives a fairly good result, where the accuracy of the prediction obtained is 89.6%.

Keywords: prediction, k-means, k-nearest neghbour

1. Introduction

Data classification is one of the most important fields in the current era of intelligent systems. The data classification process can be found in almost all smart systems. Classification has been used in various research fields such as diagnosis [1], detection [2] and prediction [3]. In general, the classification method uses the concept of feature similarity on objects in the classification process. An object will be classified into a class if it has the highest similarity value with the objects in that class. This concept can be found in one of the classifier methods such as K-Nearest Neighbor.

K-Nearest Neighbor (KNN) is a classifier method that uses a distance measurement of k objects whose class is known to determine the class of an object whose class is not known. Until now, many studies still use the KNN classifier because it is able to compete with other more complex methods [4]. Eftekhar Hossain et al classified leaf diseases in plants where KNN was able to detect disease with an accuracy of 96.76% [5]. The current K-Nearest Neighbor method has also undergone a lot of development and optimization [6] [7]. Classifier methods such as K-Nearest Neighbor require training data as a reference in the classification process. datatraining consists of object feature values whose class is known which is usually obtained from actual data that has occurred before. In some cases, this data does not yet have a class label so that the classification process using K-Nearest Neighbor cannot be carried out. Clustering is one approach that can be used to overcome this problem. Objects with similar characteristics will be grouped into clusters - clusters so that class determination can be easily done. On some researches, clustering can be implemented directly in the classification process. However, compared to classifier methods such as K-Nearest Neighbor, the clustering process requires a higher computational cost.

K-Means clustering is a clustering method that is widely used today. K-Means clustering has advantages in the simplicity of the process and the adjustment of the data in each iteration. K-Means clustering works by grouping data into k clusters based on the distance of the data to the cluster centroid. K-Means clustering is very sensitive to the selection of the initial centroid of the cluster [8]. Inappropriate selection of centroids can result in the clusters that are formed trapped in the optima locale [9]. Although currently optimization has been carried out in determining the initial centroid [10]. This research implements a combination of K-Means clustering and K-Nearest Neighbor in the process of predicting e-commerce customer classes on datasets that do not yet have classes so that they can be used for e-commerce customer prediction models. K-Means Clustering is used to group datasets into several clusters, then the cluster will be assigned a class according to the characteristics of the centroid. After the classes are obtained, then the data will be divided into training and test data to measure the prediction accuracy of the K-Nearest Neighbor classification.

2. Literature Reviews

2.1 K-Means Clustering

K-Means Clustering is a non-hierarchical clustering method that partitions a data set into several clusters where each cluster will consist of data with similar characteristics [11]. The purpose of the K-Means Clustering method is to group data to find a set of C clusters from the specified number of K clusters. The initialization step is carried out to determine the number of K clusters and the initial centroid value for each cluster. In each iteration, the data will be grouped into clusters with the closest distance value using the Euclidean distance equation (eq 1).

$$\|x - y\| = \sqrt{\sum_{i=1}^{v} |x_i - y_i|^2} \quad \text{...} \ (1)$$

The new centroid will then be computed using the mean value of the cluster members (eq 2). Clustering is then continued in the next iteration by repeating the grouping process using the new centroid value. Clustering will stop if the new centroid value does not change where the minimum SSE value [12] has been achieved or the maximum iteration has been reached.

$$C_i = \frac{\sum_{j=1}^{v} x_{ij}}{n} \quad \text{...} (2)$$

2.2 K-Nearest Neighbor

K-Nearest Neighbor (KNN) is a classification algorithm by finding the nearest K object in the training data [13]. The calculation of the distance in the KNN method can use the Euclidean distance equation (eq 1). If it is assumed that $z = (x', y')$ where z is the test data, x' is the feature of the test data and y' is the class of the unknown test data, then the classification process can be carried out using the following steps:

1. Calculates the distance between the test data to each training data whose value is then stored in D.
2. Choose $D_z \in D$, which is the k nearest neighbors of the test data.
3. Determine the class of test data using equation 3.

$$y' = argmax(\sum_{v(x_i, y_i) \in D_z} I(v = y_i) \quad \text{...} \ (3)$$

3. Research Methods

The dataset used in this research is an e-commerce customer dataset that does not yet have a class. K-Means clustering will be used to group the dataset into two clusters. The two clusters will be labeled with different classes, namely "High Spender" and "Low Spender" based on the Yearly Amount Spent value of the centroid of the two clusters. The member of the cluster with the largest Yearly Amount Spent centroid will be labeled as "High Spender" while the members of the next cluster will be labeled with the class "Low Spender". The dataset that already has a class label is then divided into two different datasets, namely the training dataset and the test dataset. The stages of the prediction process can be seen in Figure 1.

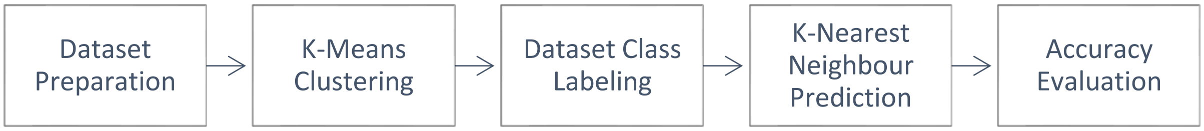

Figure 1. Research Stages

The training dataset will then be used as a data reference for the K-Nearest Neighbor method. The k parameter used in this research is k=3. The test data is then used as input to generate class predictions from the K-Nearest Neighbor method. The prediction results from the training data will then be compared with the class obtained from the K-Means Clustering process to obtain prediction accuracy using equation 4.

$$a = \frac{t}{n}x100\% \text{ .... } (4)$$

Where :

a = accuracy in percentage

t = correct predictions count

n = datatraining count

The dataset used in this research is an e-commerce customer dataset obtained from an online open dataset. The dataset consists of eight columns, namely email, address, avatar, average session length, time on app, time on website, length of membership and yearly amount spent. The raw dataset can be seen in table 1. Feature selection is then carried out to remove features from the dataset that do not affect the prediction results such as email, address and avatar so that the remaining features that will be used in the next process are average session length, time on app, time on website, length of membership and yearly amount spent.

Table 1. Raw Dataset

Email	Address	Avatar	Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent

mstephenson@fernandez.com	835 Frank …	Violet	34.497268	12.655651	39.577668	4.082621	587.951054
hduke@hotmail.com	4547 Arch..	DarkGreen	31.926272	11.109461	37.268959	2.664034	392.204933
pallen@yahoo.com	24645 Val …	Bisque	33.000915	11.330278	37.110597	4.104543	487.547505
riverarebecca@gmail.com	1414 Dav …	SaddleBrown	34.305557	13.717514	36.721283	3.120179	581.852344
mstephens@davidson-herman.com	14023 Ro…	MediumAquaMarine	33.330673	12.795189	37.536653	4.446308	599.406092
...	...	...	...	...	...	...	...
lewisjessica@craig-evans.com	4483 Jone...	Tan	33.237660	13.566160	36.417985	3.746573	573.847438
katrina56@gmail.com	172 Ow…	PaleVioletRed	34.702529	11.695736	37.190268	3.576526	529.049004
dale88@hotmail.com	0787 An...	Cornsilk	32.646777	11.499409	38.332576	4.958264	551.620145
cwilson@hotmail.com	680 Jen...	Teal	33.322501	12.391423	36.840086	2.336485	456.469510
hannahwilson@davidson.com	49791 R...	DarkMagenta	33.715981	12.418808	35.771016	2.735160	497.778642

Features selection is then carried out to remove features from the dataset that do not affect the prediction results such as email, address and avatar so that the remaining features that will be used in the next process are average session length, time on app, time on website, length of membership and yearly amount spent.

Table 2. Cleaned Dataset

Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent
34.497268	12.655651	39.577668	4.082621	587.951054
31.926272	11.109461	37.268959	2.664034	392.204933
33.000915	11.330278	37.110597	4.104543	487.547505
34.305557	13.717514	36.721283	3.120179	581.852344
33.330673	12.795189	37.536653	4.446308	599.406092
...	...	...	...	...
33.237660	13.566160	36.417985	3.746573	573.847438
34.702529	11.695736	37.190268	3.576526	529.049004
32.646777	11.499409	38.332576	4.958264	551.620145
33.322501	12.391423	36.840086	2.336485	456.469510

33.715981	12.418808	35.771016	2.735160	497.778642

4. Results and Discussions

The dataset that has been prepared will first be clustered using K-Means Clustering which is intended to produce two different clusters for class labeling on the data contained in the dataset. The results of the clustering process using K-Means Clustering can be seen in table 3. From the results of the clustering, two different clusters with the centroid cluster of consecutive features [Avg. Session Length, Time on App, Time on Website, Length of Membership, Yearly Amount Spent] for cluster C1 is [33.37427499, 12.39869746, 37.07244144, 4.17299573, 561.35877976] and the centroid of cluster C2 is [32.73467047, 11.70903705, 37.0485444799, 437.99023277799].

Table 3. Clustering Result

Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent	Cluster
34.497268	12.655651	39.577668	4.082621	587.951054	1
31.926272	11.109461	37.268959	2.664034	392.204933	2
33.000915	11.330278	37.110597	4.104543	487.547505	2
34.305557	13.717514	36.721283	3.120179	581.852344	1
33.330673	12.795189	37.536653	4.446308	599.406092	1
...	...	...	...	...	...
33.237660	13.566160	36.417985	3.746573	573.847438	1
34.702529	11.695736	37.190268	3.576526	529.049004	1
32.646777	11.499409	38.332576	4.958264	551.620145	1
33.322501	12.391423	36.840086	2.336485	456.469510	2
33.715981	12.418808	35.771016	2.735160	497.778642	2

Because the value of the Yearly Amount Spent feature of C1 is greater than C2, each member of C1 will be assigned a class label of "High Spender" and "Low Spender" for members of C2 cluster. The results of class labeling can be seen in table 4. The dataset that already has a class label is then divided into two different datasets, namely the training dataset and the test dataset.

Table 4. Class Labeling

Avg. Session Length	Time on App	Time on Website	Length of Membership	Yearly Amount Spent	Label
34.497268	12.655651	39.577668	4.082621	587.951054	High Spender
31.926272	11.109461	37.268959	2.664034	392.204933	Low Spender
33.000915	11.330278	37.110597	4.104543	487.547505	Low Spender
34.305557	13.717514	36.721283	3.120179	581.852344	High Spender
33.330673	12.795189	37.536653	4.446308	599.406092	High Spender
...	...	...	...	...	...
33.237660	13.566160	36.417985	3.746573	573.847438	High Spender
34.702529	11.695736	37.190268	3.576526	529.049004	High Spender
32.646777	11.499409	38.332576	4.958264	551.620145	High Spender

33.322501	12.391423	36.840086	2.336485	456.469510	Low Spender
33.715981	12.418808	35.771016	2.735160	497.778642	Low Spender

The training dataset and the test dataset are then used together in the prediction process using the K-Nearest Neighbor method. The features used in the training dataset and the test dataset are all features of the clustering results except for the Yearly Amount Spent feature. The results of the predictions can be seen in table 5. The class prediction results from the test dataset are then compared with the original class to obtain the accuracy value of the prediction results which can be seen in table 6.

Table 5. Prediction Results

Avg. Session Length	Time on App	Time on Website	Length of Membership	Prediction
32.672944	12.276057	37.192794	3.982472	High Spender
32.728521	10.131712	34.845612	3.287702	Low Spender
33.409923	12.026942	36.133894	2.313350	Low Spender
31.724203	13.172287	36.199753	3.557814	Low Spender
32.711119	12.326291	36.673878	3.350279	Low Spender
...	...	...	...	...
33.237660	13.566160	36.417985	3.746573	High Spender
34.702529	11.695736	37.190268	3.576526	High Spender
32.646777	11.499409	38.332576	4.958264	High Spender
33.322501	12.391423	36.840086	2.336485	Low Spender
33.715981	12.418808	35.771016	2.735160	Low Spender

Table 6. Prediction Evaluation

Avg. Session Length	Time on App	Time on Website	Length of Membership	Label	Predict.	Status
32.672944	12.276057	37.192794	3.982472	High Spender	High Spender	True
32.728521	10.131712	34.845612	3.287702	Low Spender	Low Spender	True
33.409923	12.026942	36.133894	2.313350	Low Spender	Low Spender	True
31.724203	13.172287	36.199753	3.557814	High Spender	Low Spender	False
32.711119	12.326291	36.673878	3.350279	Low Spender	Low Spender	True
...	...	...	...	...	...	...
33.237660	13.566160	36.417985	3.746573	High Spender	High Spender	True
34.702529	11.695736	37.190268	3.576526	High Spender	High Spender	True

32.646777	11.499409	38.332576	4.958264	High Spender	High Spender	True
33.322501	12.391423	36.840086	2.336485	Low Spender	Low Spender	True
33.715981	12.418808	35.771016	2.735160	Low Spender	Low Spender	True

Evaluation of predictions from the prediction model resulted in 224 predictions with correct class values and 26 predictions with incorrect values. Based on the evaluation of these predictions, the accuracy of the prediction model used is 89.6%. This shows that the combination of clustering and classification using K-Means Clustering and K-Nearest Neighbor provides good accuracy in predicting the spend rate of e-commerce customers.

5. Conclussions

Classification on datasets that do not yet have a training dataset can be done by the use of clustering method as an initial stage before the classification process. This research use an e-commerce customer dataset that does not yet have a class label so that it requires additional methods to be able to produce a training dataset. The K-Means Clustering method is able to produce two different clusters from the dataset so that it is possible to assign class labels to each data contained in the dataset. The dataset that has been prepared can then be used for testing predictions using the K-Nearest Neighbor method which provides a fairly good accuracy of 89.6%.

Reference

[1] L. Kou, C. Liu, G.-W. Cai, J.-n. Zhou, Q.-d. Yuan and S.-m. Pang, "Fault diagnosis for open-circuit faults in NPC inverter based on knowledge-driven and data-driven approaches," IET Power Electronics, vol. 13, no. 6, pp. 1236-1245, 2020.

[2] Y. Wu and S. Liu, "A review of data-driven approaches for burst detection in water distribution systems," Urban Water Journal, vol. 14, no. 9, pp. 972-983, 2017.

[3] Y. Wei, X. Zhang, Y. Shi, L. Xia, S. Pan, J. Wu, M. Han and X. Zhao, "A review of data-driven approaches for prediction and classification of building energy consumption," Renewable and Sustainable Energy Reviews, vol. 82, pp. 1027-1047, 2018.

[4] H. A. A. Alfeilat, A. B. Hassanat, O. Lasassmeh, A. S. Tarawneh, M. B. Alhasanat, H. S. E. Salman and V. S. Prasath, "Effects of distance measure choice on k-nearest neighbor classifier performance: a review," Big data, vol. 7, no. 4, pp. 221-248, 2019.

[5] E. Hossain, M. F. Hossain and M. A. Rahaman, "A color and texture based approach for the detection and classification of plant leaf disease using KNN classifier," in 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE), 2019.

[6] J. Gou, H. Ma, W. Ou, S. Zeng, Y. Rao and H. Yang, "A generalized mean distance-based k-nearest neighbor classifier," Expert Systems with Applications, vol. 115, pp. 356-372, 2019.

[7] W. M. Shaban, A. H. Rabie, A. I. Saleh and M. Abo-Elsoud, "A new COVID-19 Patients Detection Strategy (CPDS) based on hybrid feature selection and enhanced KNN classifier," Knowledge-Based Systems, vol. 205, p. 106270, 2020.

[8] K. P. Sinaga and M.-S. Yang, "Unsupervised K-means clustering algorithm," IEEE Access, vol. 8, pp. 80716-80727, 2020.

[9] M. E. Celebi, H. A. Kingravi and P. A. Vela, "A comparative study of efficient initialization methods for the k-means clustering algorithm," Expert Systems with Applications, vol. 40, p. 200, 2013.

[10] M. A. Syakur, B. K. Khotimah, E. M. S. Rochman and B. D. Satoto, "Integration k-means clustering method and elbow method for identification of the best customer profile cluster," IOP Conference Series: Materials Science and Engineering, vol. 336, no. 1, p. 012017, 2018.

[11] I. Lubis, H. A. Simamora, S. Rahman, R. Siregar and H. Lubis, "Aplikasi Edit Foto Background Dengan Menggunakan Metode K-Means Clustering," Query: Journal of Information Systems, vol. 3, no. 1, pp. 12-21, 2019.

[12] A. F. Jahwar and A. M. Abdulazeez, "Meta-heuristic algorithms for k-means clustering: A review," PalArch's Journal of Archaeology of Egypt/Egyptology, vol. 17, no. 7, pp. 12002-12020, 2020.

[13] Okfalisa, I. Gazalba, Mustakim and N. G. I. Reza, "Comparative analysis of k-nearest neighbor and modified k-nearest neighbor algorithm for data classification," in 2017 2nd International conferences on Information Technology, Information Systems and Electrical Engineering (ICITISEE), IEEE, 2017.