Improving Clustering Quality Using Silhouette Score

Tin Tin Hmwe
University of Computer Studies, Pyay
tthmwe.cupyay@gmail.com

Nwet Yin Tun Thein
University of Computer Studies, Yangon
nwetyin2019@gmail.com

Khin Mar Cho
University of Computer Studies, Pyay
khinmarcho.07@gmail.com

Abstract

Understanding customers helps business to provide tailored services. Clustering analysis help all business owner to gain a coherent understanding of their customers. In order to maximize the value of each customer to the business, cluster customers into segments based on their income and spending score. For this purpose, K-means is used. K-means clustering is an unsupervised learning technique which mainly deals with identifying the structure or pattern of the data. In this type of algorithms, labeled data is absent (or the dependent variable is absent). Clustering in Customer data is the process of dividing a company's customers into groups that reflect similarity among customers in each group. A different number of clusters is tested with k-means classifier. In this experiment, up to 30 clusters is tested with the algorithm. This paper aims to improve the performance of the clustering results by measuring clustering quality with Silhouette scores. The best number of clusters is determined with Silhouette scores: the higher the better.

Keywords—clustering, analysis, customer segmentation

1. Introduction

In the field of e-commerce, knowing customer-specific data, such as customer profile, transaction, and web browsing data has led to a successful business. The main objective of every industry is to perceive each customer individually and use that to make it easier for the customer to do business with them rather than with competitors. In e-commerce sector, bucketing customers is a critical task for marketers. Customers segmentation groups the customers with frequent features into different clusters. Clustering algorithms, one of knowledge discovery techniques, is generally used in data or customer segmentation based on their resemblance. Analyzing and identifying low-risk customers, maintaining those customers are critical work of business organization.

The experiment in this paper aims to deploy clustering application on the fields of Business sector. To achieve this, two distinct features are examined. Firstly, an extensive number of categorical attributes are considered in the mall consumer data. Secondly, customers do not necessarily acquire a former insight of cluster analysis process. In this paper, K-Means clustering techniques is applied to identify groups of the customers. Computing the cluster quality process is continued after clustering. The process of measuring the cluster quality is complex. To solve this problem, various methods lied to decide the optimal number of clusters. Among these techniques, Silhouette measure is the utmost descriptive and used in this experiment for this reason.

Customers differ in terms of behavior, needs, wants and characteristics and the main objective of clustering is to identify different customer types and segment the customer base into buckets of similar profiles so that the process of target marketing can be carried out more competent. Hierarchical and k-means clustering techniques are extensively used in customer bucketing. Distance-based K-Means and Agglomerative Hierarchical Clustering are most broadly used among them.

Utilizing clustering algorithms in customer relationship management is conceptually discussed in [8]. It concentrates on behavioral segmentation. In [3], K-Means has been used as part of their clustering approach. G. Lefait,et al[4] also experimented K-Means for customer segmentation on their dataset. Despite hierarchical clustering techniques imply improper to many, [5] have used it for customer bucketing for their research and [6] have made use of it for applying clustering techniques on the supermarket's transaction data. K-means and Hierarchical Clustering algorithms are useful for clustering data and find comprehensive usage in customer segmentation. Hence, not only segmenting customers but improving cluster quality is the focus of this paper.

The paper is organized as follows. Section 3 provides detail information about the customer dataset used for customer segmentation and related concepts: Data wrangling or data preprocessing such as handling missing values, data feature transformation: data encoding and data standardization, K-Means clustering and Silhouette Score Methods for measuring Cluster quality. The Modelling of clustering in the context of customer segmentation and the major models of K-Means and Measuring Silhouette scores are described in Section 4. Finally, a results discussion and concluding remarks are provided in Section 5 and 6 respectively.

2. Clustering Techniques and Cluster Quality

2.1 Clustering Techniques

Clustering is an unsupervised classification algorithm which has no class label. Similar data points in the dataset are grouped to one of the output class. A

cluster is formed for each class. The number of output class becomes the clusters respectively. The data points within each cluster has high intra class similarity and low inter class similarity. Many applications today such as image segmentation, customer segmentation and grouping web pages, uses clustering techniques. Clustering techniques is broadly classified into four groups: partitioning methods, hierarchical methods, density-based methods and grid-based methods.

- **Partitioning algorithms:** Construct various partitions and then evaluate them by some criterion.
- **Hierarchy algorithms:** Create a hierarchical decomposition of the set of data (or objects) using some criterion.
- **Density-based:** based on connectivity and density functions.
- **Grid-based:** based on a multiple-level granularity structure

2.2 Clustering Quality

A good clustering method causes high grade clusters by high intra-class similarity and low inter-class similarity. The quality of a clustering result mainly depends on both the similarity measure used by the method and its implementation method. The quality of a standard clustering method is also calculated by its ability to explore some or all of the unknown patterns. There are mainly two types of measures to assess the clustering quality [9].

(i) ***External Measures*** which depend upon ground truth labels. Examples are conformed Rand index, Fowlkes-Mallows scores, information gain scores, similitude, integrity and V-measure.

(ii) ***Internal Measures*** that does not need ground truth labels. Silhouette Coefficient, Calinski-Harabasz Index, Davies-Bouldin Index etc., are used to measure the clustering performance.

3. Materials and methods

3.1 Customer Dataset Description

A Shopping Mall Dataset of Customers used in this experiment is from Kaggle[10]. In order to perform analysis, the Dataset of Customers categorized the customers into groups of individuals that are similar in specific ways. The dataset contains ~ 200 records with 5 features- Customer Id, Gender, Age, Annual Income (k$) and Spending Score (1–100). The last feature Spending Score refers to Score computed for each of their clients based on several criteria including for example their income and the number of times each week they appear in the mall and of course, the amount of cash they used in a year. the score ranges between 0 and 100 for low spend and high spend respectively. The description of the mall dataset is shown in Figure 1.

Input:
- k: the number of clusters,
- D: input dataset containing n objects.

Output: A set of k clusters.
Method:
(1) arbitrarily select k objects from D as the initial cluster centers;
(2) repeat
(3) (re)assign each object to the cluster to which the object is the most , based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, compute the mean value of the objects for each cluster;
(5) until no change;

	CustomerID	Gender	Age	Annual Income (k$)	Spending Score (1-100)
0	1	Male	19	15	39
1	2	Male	21	15	81
2	3	Female	20	16	6
3	4	Female	23	16	77
4	5	Female	31	17	40

Figure 1. Dataset Description

3.2 K-Means Clustering Techniques

Clustering process divisions a set of data instances into segments or buckets. Objects in a cluster are like one another, but The clusters resulted from the cluster analysis can be called as a clustering. Many applications such as business intelligence, image pattern recognition, web search, customer segmentation and information security has widely used clustering techniques.

In business intelligence, clustering can be used to construct a huge number of customers into groups, where customers within a group share similar attributes. This aids the development of business approaches for enhanced customer relationship management[1].

A variety of cluster analysis strategies have also been built into many numerical analysis software systems. The k-means algorithm is one of the data mining and machine learning tools and it is a centroid-based partitioning technique.

In machine learning, supervised learning is a process of classification where the information for class label is known for each training instance. In case of unsupervised learning such as clustering, the class label information is not presentFor this reason, clustering is a process of learning by observation, rather than learning by examples[1].

The centroid of a cluster is identified as the mean value of the instances within the clusters. It handles as follows. First, it inconstantly chooses k of the objects in D, each of which originally represents a cluster mean or center.

For the rest of the objects, an object is appointed to the cluster to which it is the most similar, based on the Euclidean distance between the object and the cluster mean. The k-means algorithm then iteratively develops

the within-cluster variation. For each cluster, it enumerates the new mean using the objects appointed to the cluster in the earlier iteration. All the objects are then reappointed using the updated means as the new cluster centers. The iterations go on until the assignment is unchanged, that is, the clusters formed in the current round are the same as those formed in the foregoing round. The k-means procedure is summarized in Figure 2.

Figure 2. K-Means Algorithm

3.3 Silhouette Coefficient

This is a better method to decide the optimal number of clusters to be formulated from the data. Silhouette specifies to a process of understanding and affirmation of consistency within clusters of data. The Silhouette measures the similarity of a data instance within a cluster comparing with another cluster. The score is computed for each data instance and the formula (1) shows for calculation of Silhouette coefficient.

Silhouette Coefficient = (x-y)/ max(x,y) (1)

where, y is the mean distance to the other instances in the same cluster. x represents mean distance to the instances of the next closest cluster.wwa

The coefficient ranges between -1 and 1. A value close to 1 denotes that the instance is near to its cluster is a part of the right cluster. Whereas a value near to -1 means that the value is appointed to the wrong cluster.

4. Modelling k-Means clustering for customer segmentation

In business sector, data mining and machine learning models are developed using the customer's data to target the right customer. Clustering analysis is used to segment customers into clusters based on their annual income. In order to achieve this, unsupervised learning techniques of Principal Component Analysis (PCA) and one of the finest clustering algorithms, k-means is used to identify customer segments of the shopping mall customer data. Segmenting customers can maximize the business sales.

4.1 Exploring the Data

The descriptive statistics for spending score are shown in Figure-3, mean, standard deviation, median and variance statistics is presented. If the variable is not numeric, it is got the counts in each category.

Variable	Mean	Standard Deviation	Median	Variance
Spending Score (1-100)	50.2	25.758882	50.0	663.52

Figure 3. The descriptive statistics for spending score

The correlation between the numeric parameters is also analyzed and is shown in Figure-4.

	CustomerID	Age	AnnualIncome	SpendingScore
CustomerID	1.000000	-0.026763	0.977548	0.013835
Age	-0.026763	1.000000	-0.012398	-0.327227
AnnualIncome	0.977548	-0.012398	1.000000	0.009903
SpendingScore	0.013835	-0.327227	0.009903	1.000000

Figure 4. Correlation Measures

4.2 Data Wrangling

According to the figure, the dataset has zero null values in any column. The dataset has only one categorical feature: Gender which is encoded using one-hot encoding method. Data after encoding is shown in Figure 5. Feature scaling needs to be performed before applying dimensionality reduction techniques to the data so that the principal component vectors are not influenced by the natural differences in scale for features. A Standard Scaler is applied to standardize customer data.

	CustomerID	Age	Annual Income (k$)	Spending Score (1-100)	Gender_Female	Gender_Male
0	1	19	15	39	0	1
1	2	21	15	81	0	1
2	3	20	16	6	1	0
3	4	23	16	77	1	0
4	5	31	17	40	1	0

Figure 5. Dataset After Applying One-hot Encoding

Standardization is an important part of data preprocessing and it avoids a vast difference between the range of the feature. For this purpose, logarithmic scaling is applied on the data. The Standardized data is shown in Figure 6.

	Age	AnnualIncome	SpendingScore
0	2.944439	2.708050	3.663562
1	3.044522	2.708050	4.394449
2	2.995732	2.772589	1.791759
3	3.135494	2.772589	4.343805
4	3.433987	2.833213	3.688879
5	3.091042	2.833213	4.330733
6	3.555348	2.890372	1.791759
7	3.135494	2.890372	4.543295
8	4.158883	2.944439	1.098612
9	3.401197	2.944439	4.276666

Figure 6. Standardized Data

4.3 K-Means Clustering Analysis

In this section, the customer segmentation using unsupervised machine learning techniques with k-means algorithm is experimented and presented. K-Means algorithm is tested with a different number of clusters. There is no general ruling on this issue. It really depends on the data. In this experiment, up to 30 clusters are tested. The optimal number of clusters

is decided with Silhouette scores: the higher the score is, the better the cluster quality is.

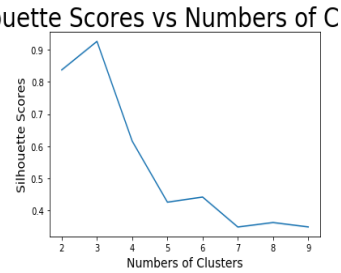

Figure 7. Measuring Silhouette Score

K-Means modelling starts from k=2 to k=30 and collect the silhouette scores for each of the cluster results. Then the best number of clusters is determined.

Table 1. Silhouette Scores for 30 Clusters

Number of Cluster	Silhouette Score Value
Cluster 2	0.8368483669134059
Cluster 3	0.928761990963092
Cluster 4	0.6119688865304789
Cluster 5	0.39691511122181805
Cluster 6	0.4049168004081001
Cluster 7	0.4106739544027617
Cluster 8	0.3208824240363465
Cluster 9	0.3194895215055452
Cluster 10	0.32499524484272224
Cluster 11	0.3234705692491315
Cluster 12	0.3388430409170668
Cluster 13	0.34035099104555366
Cluster 14	0.33511962883791424
Cluster 15	0.34791126685784307
Cluster 16	0.343649962765543
Cluster 17	0.3432787942914479
Cluster 18	0.34987111132715
Cluster 19	0.32514059439530607
Cluster 20	0.3382787867672019
Cluster 21	0.35031807261302006
Cluster 22	0.3490537503990409
Cluster 23	0.347557702223237
Cluster 24	0.347020986329717
Cluster 25	0.3388027853015341
Cluster 26	0.3283682069074598
Cluster 27	0.33662765903301345
Cluster 28	0.3403503932546018
Cluster 29	0.3451760362030432

The best number of clusters seem to be 6 (or maybe 9) or 2 in this experiment as shown in Figure 7 and Figure 8. Deciding the total of clusters regarding the optimal metric to be chosen for the number of cluster decision is to be made in line with the wants of the market or product.

The data are plotted on a spending score Vs annual income curve. According to the clustering results, the mall customers can be broadly grouped into 5 groups based on their purchases or spending scores made in the mall. Each cluster point is marked using a different sign, and the centroids of each cluster are marked using solid orange dots.

In Blued-colored cluster, people have average income and an average spending score, these people again will not be the choice targets of the markets, but again they will be deal with and other data mining techniques may be used to grow their spending score.

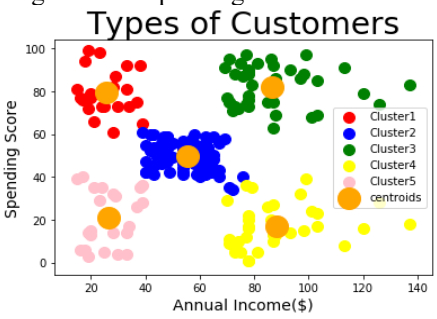

Figure 8. Clusters of Customers

In pinked-colored cluster, people have low annual income and low spending scores, this is quite reasonable as people having low salaries prefer to buy less, The marketers will be least attentive in people belonging to this cluster.

In yellowed-colored cluster. people have large income but little spending scores. These people may be dissatisfied with the marker's services. These can be the core objectives of the market. Therefore, marketers should do additional new facilities so that they can attract these people and can meet their needs.

In red-colored cluster, people have low income but higher spending scores, these are those people who for some reason love to buy products more often even though they have a low income. This may be that these customers are appeased with the marker's services.

In green-colored cluster, people have high income and high spending scores, this is the ideal case for the mall or shops as these people are the prime sources of profit. Those customers might be the ordinary customers of the business organization.

5. Results and Discussion

The analysis of Silhouette score is an examination of consistency between the clusters. This paper tested the Silhouette scores up to 30 clusters and select the best scores among them. The analysis of Silhouette score is examination of consistency between the clusters. This paper tested the Silhouette scores up to 30 clusters and select the best scores among them. The Figure-7 shows that the optimal average score of 0.93 for cluster 3 and the average Silhouette measures decreased obviously to about 0.40 to 0.32 respectively.

Compared with the clustering approaches presented in [11], Tripathi et al., compares the clustering results of customer segmentation using K-Means and Hierarchical Clustering. They examined pros and cons of each clustering methods they presented and considered to bring out a hybrid approach by combining the K-Means and Hierarchical clustering to perform better than the individual method.

Finally, Using knowledge mining techniques, to escalate the profits of the business, the marketers should focus customers associated with cluster 3 and cluster 5 and should also control its standards to keep the customers associated with cluster 1 and cluster 3. To conclude, it is incredible to apply machine learning and

data mining techniques in businesses to augment benefit.

6. Conclusion

Knowledge discovery methods and applications in customer clustering has been analyzed to identify high profit customers. Shopping mall dataset is used for analysis in this paper. Then, K-Means techniques was applied on standardized mall dataset for grouping similar behavior customers. The optimal k-number of clusters are chosen by using Silhouette score method. The greater the score is the better the cluster quality is. This analysis can be applied in businesses to divide customers into clusters based on annual income and spending score.

References

[1] Jiawei Han Micheline Kamber Jian Pei, "*Data Mining: Concepts and Techniques", 3rd Edition,* **ISBN:** 9780123814807.

[2] "Algorithm of Principal Component Analysis (PCA)", https://iq.opengenus.org/algorithm-principal-component-analysis-pca/

[3] Y. Chen, et al., "Identifying patients in target customer segments

using a two-stage clustering-classification approach: A hospital-based assessment", Computers in Biology and Medicine, vol. 42, no. 2, pp. 213-221, 2012.

[4] G. Lefait and T. Kechadi, "Customer segmentation architecture based on clustering techniques", in Fourth International Conference on Digital Society, Sint Maarten, 2010, pp. 243-248.

[5] M. Namvar, M. Gholamian and S. KhakAbi, "A two-phase clustering method for intelligent customer segmentation", in International Conference on Intelligent Systems, Modelling and Simulation, Liverpool, 2010, pp. 215-219.

[6] D. Gaur and S. Gaur, "Comprehensive analysis of data clustering algorithms", in Future Information Communication Technology and Applications. Dordrecht: Springer Netherlands, 2013, pp. 753-762.

[7] .Sajana, C. Sheela Rani and K. Narayana, "A Survey on Clustering Techniques for Big Data Mining", Indian Journal of Science and Technology, vol. 9, no. 3, 2016.

[8] H. Ziafat, M. Shakeri, "Using Data Mining Techniques in Customer Segmentation", SSN : 2248-9622, Vol. 4, Issue 9, Int. Journal of Engineering Research and Applications, 2014, pp.70-79

[9] "Clustering Evaluation Strategies", https://towardsdatascience.com/clustering-evaluation-strategies-98a4006fcfc.

[10] https://www.kaggle.com/akram24/mall-customers

[11] S. Tripathi, A. Bhardwaj, "Approaches to Clustering in Customer Segmentation", International Journal of Engineering &Technology, 7 (3.12) (2018) 802 -807.