Enhancing Beer Recommendations through Clustering: A Comparison of Hierarchical and K-means Clustering Methods on Normalized Data

Tiansheng Zhu[1a,*], Yina Han[2b]

[a,*]B19090119185@cityu.mo,[b]1941065724@qq.com

[1]School of Business, City University of Macau of Business Administration, Macau 999078, China
[2]Business Administration, Hunan Normal University, Changsha, China

Abstract. The process of normalization confers notable advantages in optimizing gradient descent algorithms for machine learning applications and simplifying data processing. Through normalization, the scale of numeric types in the dataset can be adjusted, thereby facilitating the identification of patterns and trends that would otherwise remain obscured by variations in the magnitude of the features. This technique also reduces the risk of certain features having a disproportionate influence on the model's decision-making process. The scale function is one such means of normalization that effectively mitigates such issues.In the context of clustering, normalization plays a crucial role in enhancing the accuracy and efficiency of the process. It is imperative to note that in the case of an unclassified dataset, where the distribution of data is unknown, unsupervised learning models must be employed. Unsupervised learning models enable the identification of patterns and structures within the dataset without the need for prior classification. Clustering is one such approach that can extract meaningful insights from unstructured datasets.Through clustering data, the present research offer practical insights for businesses seeking to leverage data-driven approaches to enhance their operations and improve customer satisfaction.

Keywords: hierarchical method, k-means clustering method, beer recommendations, data-driven approache

1 Introduction

This present research endeavors to process a dataset consisting of 199 observations and 9 variables, including but not limited to 'Name', 'ABV', and 'IBU', among others. The objective is to leverage clustering techniques to impute the missing data and to recommend similar beers to the client. To achieve this aim, all the analytical procedures were conducted in the RStudio environment. Initially, it was observed that the missing data were confined to only two variables, namely 'ABV' and 'EBC,' and constituted an insignificant proportion of the overall dataset. To discern the missing pattern, a correlation test was executed, which revealed a weak correlation between the missing values [1]. As a result, it was deduced that the missing data was Missing Completely at Random (MCAR) and could be imputed using appropriate methods. Accordingly, Simple Imputation and Multiple Imputation techniques were employed, and the resultant imputed datasets were compared. After a comprehensive assessment of the outcomes, it was

MSEA 2023, May 26-28, Nanjing, People's Republic of China

DOI 10.4108/eai.26-5-2023.2334332

concluded that the Multiple Imputation technique was the most suitable approach to address the missing data problem.

Thereafter, the dataset was normalized to standardize the variables and to enhance the clustering process. K-means clustering was used to segregate the dataset into similar groups based on the respective observations' attributes. This approach facilitated a detailed understanding of the beer types in the dataset, and the resultant clusters served as the basis for recommending similar beers to the client [2].In conclusion, this report demonstrates how clustering techniques and imputation methods can be utilized to process datasets and recommend similar products. The findings offer practical insights for businesses seeking to leverage data-driven approaches to enhance their operations and improve customer satisfaction [3].

2 Imputation

When analyzing a dataset, the first thing we need to do is to process the dataset, i.e., data pre-processing. It is very important to fill in the missing data, as there are often errors or missing data in the dataset, instead of simply selecting random numbers or deleting missing data.

To analyze the situation of the data, first, need to know how much data exists in the data set and how much data is complete and incomplete. Using the VIM package aggr function to see how much data is missing and the situation [4].

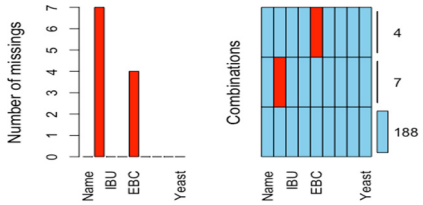

Figure 1. aggr function to view the distribution of missing values

Through figure 1 it can be found that 11 data are incomplete, including 'ABV' with 7 missing data, 'EBC' with 4 missing data, and 188 data are complete, with missing values accounting for 5.53% . A dataset with more than 1% missing values indicates that the data cannot be filled with Simple Imputation alone. too much missing data in the dataset, if Simple Imputation is used, it will easily cause too large a standard deviation and affect the final confidence of the data.To determine if there is a correlation between the missing data, correlation analysis is performed on the missing values by the corrgram package. If a strong correlation exists, it indicates that the data are not missing at random and other methods of imputation.

The results of missing value correlation analysis find that the correlation between missing values and all feature indicators is weak, and the missing data only exist in 'ABV' and 'EBC', indicating that the missing pattern of data is Missing Completely at Random (MCAR), the missing values can be filled by using Multiple Imputation.

Multiple Imputations can be performed on the dataset by using the mice package, and since the data format type of 'Name' and 'Yeast' in the dataset is a string, it can be excluded when creating a subset. When using the mice function for Imputation, determine m=5, maxit=10, and finally

use the complete function for filling. To ensure the reliability of Multiple Imputations, they can be compared with Simple Imputations. Here, Simple Imputation uses the method of inserting the average value for data interpolation.

To compare the effectiveness of the two interpolation methods, the frequency plots of 'ABV' and 'EBC' are drawn separately using the hist function, where mirep is using Multiple Imputations and si is using Simple Imputations.

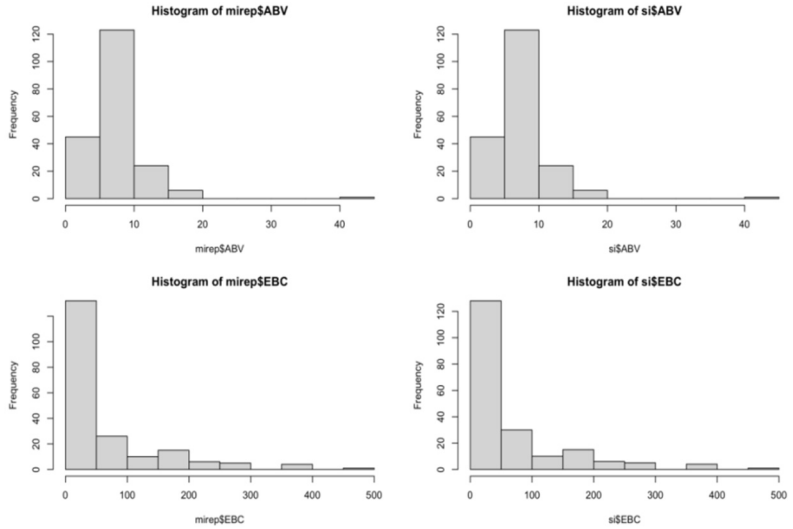

Figure 2. Comparison of the effect of two imputation methods

Through figure 2, we can find that the effect after the two imputations is not obvious, probably because the difference between the data is too small or the original distribution of the data is in line with the law of average imputations. Therefore, to obtain more accurate results, it is necessary to use the summary function and the sd function for a more refined comparison.

Table 1. Comparison of two imputations values with the original data mean and standard deviation

	Brewdog$ABV	Si$ABV	Mirep$ABV	Brewdog$EBC	Si$EBC	Mirep$EBC
Mean	7.675	7.675	7.669	71.66	71.03	71.66
sd	3.946238	3.875854	3.875989	90.85139	89.92902	90.0393

Through table 1, we can find that the difference between the standard deviation of the imputation data and the standard deviation of the original data shows that Mirep$ABV (0.070249) < Si$ABV (0.070384), Mirep$EBC (0.81209) < Si$EBC (0.92237), which indicates that the overall effect of Multiple Imputation is better than Simple Imputation. It is also found that the difference between the mean and standard deviation of the imputation data and the original data is particularly small, so it is more beneficial to adopt Multiple Imputation when dealing with missing values [5], as can be seen in figure3.

	ABV	IBU	OG	EBC	PH	AttenuationLevel	FermentationTempCelsius
1	7.50	50.0	1070.0	40.0	4.4	81.40	21
2	9.00	50.0	1084.0	20.0	4.4	82.10	21
3	10.00	85.0	1098.0	130.0	4.4	79.60	21
4	7.80	70.0	1074.0	90.0	4.4	79.70	18
5	5.00	30.0	1050.0	60.0	4.4	76.00	19
6	4.90	30.0	1047.0	12.0	4.4	80.70	10
7	18.00	70.0	1150.0	57.0	4.4	93.30	22
8	10.50	14.0	1093.0	79.0	4.4	80.00	19
9	15.00	80.0	1113.0	400.0	4.0	84.10	21
10	11.20	150.0	1098.0	70.0	4.4	87.00	17
11	10.43	65.0	1095.0	23.0	4.4	83.20	21
12	11.50	80.0	1096.0	115.0	4.4	79.20	20
13	12.80	70.0	1108.0	79.0	4.4	81.50	18
14	11.30	50.0	1098.0	164.0	4.4	79.60	20
15	12.80	50.0	1096.0	111.0	4.4	79.17	21
16	10.70	100.0	1105.0	300.0	4.3	76.20	21
17	11.80	80.0	1096.0	115.0	5.2	79.20	20
18	14.20	20.0	1025.0	67.0	4.0	75.60	21
19	4.50	40.0	1045.0	18.0	4.2	75.60	19
20	4.50	40.0	1045.0	18.0	4.2	75.60	19

Figure 3. Data after using Multiple Imputation

3 Clustering

After processing the data, it is still necessary to normalize the data. Normalization speeds up the gradient descent to find the optimal solution and makes it easier to process the data. By using the scale function, the values of numeric types can be normalized.

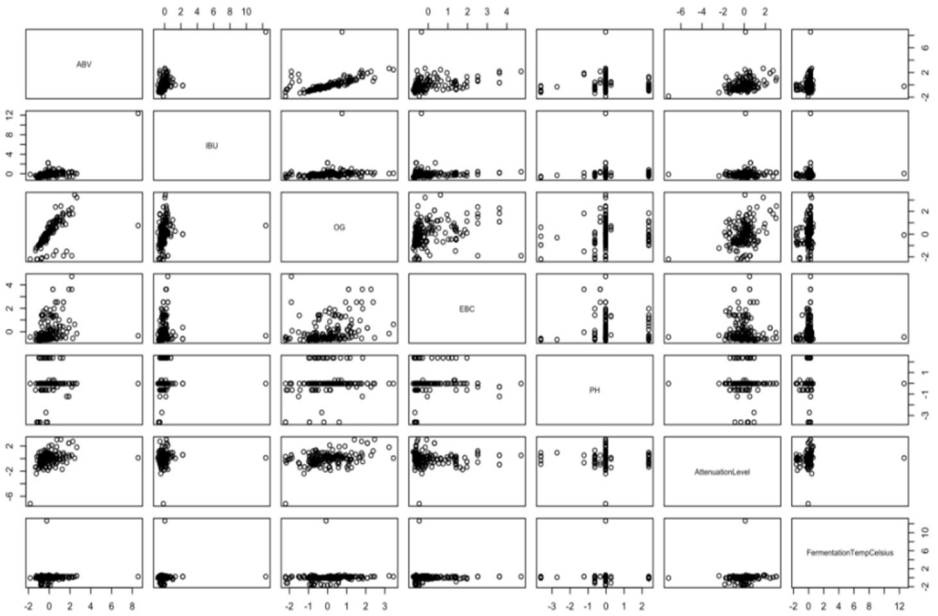

Figure 4. Normalized data

Since the dataset is not accurately classified, an unsupervised learning model must be chosen for clustering. As can be seen in figure 4, one of the values in the normalized data is significantly higher than all the values, so pay special attention to this data when clustering.

The first clustering method can be considered hierarchical clustering. Hierarchical clustering often involves the following steps: to specify the separation or variation between clusters. Make each point a separate cluster (n points, n clusters). Up until there is just 1 cluster, repeat the

following: Measure the separation between each cluster. Combine the two groups that are closest to one another. maintains the cluster operations' order [6].

Using the hclust function for hierarchical clustering and setting the k value of hierarchical clustering to 3, we can get the dendrogram. The method used for hierarchical clustering is Ward. The ward approach is where the cluster membership of a data point is assessed by computing the sum of squared total deviations from the cluster mean. The criterion for merging clusters is that the merging should minimize the increase in the sum of squared errors.

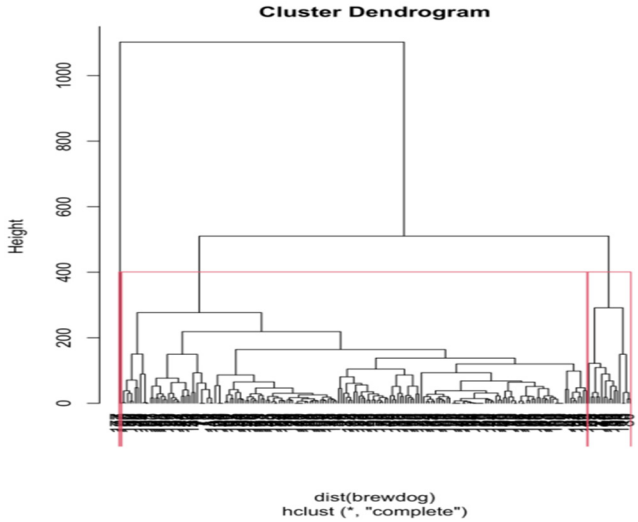

Figure 5. dendrogram

Figure 5 shows that all the data are divided into three categories, and the size difference of each category is obvious.

Looking at the dendrogram, find a horizontal rectangle of maximum height from top to bottom that does not intersect any horizontal vertical dendrogram lines. The portion of the dendrogram where the rectangle with the maximum height can be cut out was chosen because it represents the maximum Euclidean distance between the optimal number of clusters. From this dendrogram, there exist three large matrices representing the optimal number of clusters as three clusters.

The dendrogram shows that one of the data has a height of about 1100, which is significantly higher than the height of the other data, so it is classified into a separate category. A separate class of data ids can be found in the graph as 177. For the second category, according to the rules of hierarchical clustering, the distance between sample points should be calculated, and the data will become the same cluster because they are basically at about the same height. For the third category, the data with lower heights in the second category also had obvious gaps, so it became the third category. The data after hierarchical clustering was divided into three categories, of which the first category was 1 data, the second category was 181 data, and the third category was 17 data.

K-means is another unsupervised clustering algorithm with excellent results. K-means has a

better clustering algorithm than hierarchical clustering and can handle larger dimensions and larger data sets [7]. It is sensitive to outliers, and a data point that is far out of the majority may affect a cluster or create its own cluster. An imbalance in the data or say a large difference in the data between categories can make the clustering ineffective. In addition, K-means clustering uses an iterative approach that can handle outliers well. In contrast to hierarchical clustering, if data has been divided, then no other calculations can be performed, which will affect the subsequent calculations of the algorithm.Since Brewdog has a total of 199 data, it is very unsuitable for using hierarchical clustering. For hierarchical clustering, a number of data less than 150 is best, and the structure of the hierarchical clustering algorithm leads to less accurate results that may be recommended.To select the number of clusters, the NbClust function is used, where the method chosen is 'kmeans', and the results are shown in the figure below [8].

```
* Among all indices:
* 3 proposed 2 as the best number of clusters
* 4 proposed 3 as the best number of clusters
* 2 proposed 6 as the best number of clusters
* 1 proposed 7 as the best number of clusters
* 1 proposed 8 as the best number of clusters
* 4 proposed 9 as the best number of clusters
* 1 proposed 12 as the best number of clusters
* 2 proposed 13 as the best number of clusters
* 2 proposed 14 as the best number of clusters
* 3 proposed 15 as the best number of clusters

                   ***** Conclusion *****

* According to the majority rule, the best number of clusters is  3
```

Figure 6. NbClust recommends the best number of clusters

Figure 6 shows that the best number of clusters is three, for which we can use k-means clustering to perform clustering.

```
> model1.k = kmeans(brewdog, centers=3, iter.max = 80, nstart=50)
> brewdog.id2 = model1.k$cluster
> table(brewdog.id2)
brewdog.id2
  1   2   3
162   5  32
```

Figure 7. K-means clustering display

Through figure 7, it is found that although the number of each class differs greatly, the number of each class changes a lot compared to hierarchical clustering, so it can be used as a way of recommendation in a more practical way [9].

4 Conclusion

The present research compare two clustering methods - hierarchical clustering and K-means clustering to recommend similar beers to the client. The hierarchical clustering method involves finding the separation between clusters and combining the closest groups. The dendrogram is used to determine the optimal number of clusters. In this study, the data was divided into three categories using hierarchical clustering. However, hierarchical clustering is not suitable for large datasets like the one in this study because it can lead to less accurate results. On the other hand, K-means clustering is better for larger data sets and is more sensitive to outliers. The authors used the NbClust function to select the number of clusters and found that K-means clustering is

a more practical way to recommend similar beers. The results showed that although the number of each class differed greatly, K-means clustering could still be used to recommend similar beers in a more practical way [10].

References

[1] Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. Science, 349(6245), 255-260.

[2] Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. Electronics, 9(8), 1295.

[3] Mahesh, B. (2020). Machine learning algorithms-a review. International Journal of Science and Research (IJSR).[Internet], 9, 381-386.

[4] Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. Advances in neural information processing systems, 30.

[5] Baraldi, A. N., & Enders, C. K. (2010). An introduction to modern missing data analyses. Journal of school psychology, 48(1), 5-37.

[6] Madhulatha, T. S. (2012). An overview on clustering methods. arXiv preprint arXiv:1205.1117.

[7] Mohamad, I. B., & Usman, D. (2013). Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 6(17), 3299-3303.

[8] Murtagh, F., & Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1), 86-97.

[9] Saxena, A., Prasad, M., Gupta, A., Bharill, N., Patel, O. P., Tiwari, A., ... & Lin, C. T. (2017). A review of clustering techniques and developments. Neurocomputing, 267, 664-681.

[10] Lee, H., Kang, H., Chung, M. K., Kim, B. N., & Lee, D. S. (2012). Persistent brain network homology from the perspective of dendrogram. IEEE transactions on medical imaging, 31(12), 2267-2277.