ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

K-means clustering for the analysis of incomplete business data

Qi An[1], XiMing Ma[2]

[1]*School of Mathematical Sciences, Nanjing Normal Universtiy, Nanjing, China*
[2]*College of Data Science and Application, Inner Mongolia University of Technology, Huhhot, China*

Abstract: *Missing values can significantly reduce the accuracy and availability of business data. Usually, when clustering incomplete data, the data with missing values are deleted, and only the complete data are analyzed. However, this often leads to significant loss or deviation of information. This paper mainly studies how to use unsupervised machine learning techniques to deal with missing values. The combination of imputation method and clustering technology forms a new method to deal with missing values, which is helpful to overcome the problem of missing data. We propose a strategy based on the combination of K-means, big data K-means, p-k-means, and mean imputation method, singular value decomposition imputation method, k-nearest neighbor imputation method. By comparing the performance of nine methods in different business data sets. The experimental analysis was carried out on four benchmark data sets. The effectiveness of K-means clustering and imputation methods is verified on different data sets, and the results also have a certain application prospect.*

Keywords: *Missing data, Imputation, Clustering, Business*

1. Introduction

In data processing and analysis, missing data or missing values often appear in some observations in the data set without storing corresponding data values for specific variables [1]. XIONG et al [2] point out that Missing data often occurs in the process of data collection, transportation, storage. Due to the limitations of objective conditions such as historical conditions and equipment, complete information cannot be obtained, resulting in missing data in the collection process; Data transportation and transmission need to be completed by people. Human operation and misjudgment lead to missing data during transportation; During storage, data is missing due to storage medium failure and data compression damage. According to statistics, nearly 40% of the data in the UCI (a database for machine learning proposed by the University of California, Irvine) contains missing levels of varying degrees [3].

Missing data is an inevitable problem in cluster analysis. If missing values are allowed to perform cluster analysis directly, it will seriously reduce the effectiveness of the corresponding algorithm. Therefore, the processing of missing data is an indispensable data preprocessing process. Most of the workers are just deleting the missing data. This method is simple and easy to operate. However, Roderick et al [5] point out, the disadvantage is that when the proportion of missing data is high, the observations are regarded as entirely missing data, which will cause a large amount of missing part of the valuable information originally possessed, thereby reducing the effectiveness of data analysis. In recent years, according to XIONG et al [2], the estimation and imputation of missing data have received widespread attention. The zero-value imputation method retains, to a certain extent, the helpful information of the remaining part of the data that has missing data. However, the use of zero-value imputation will significantly affect the internal relationship between observations and other complete observations. In order to maintain the structural relationship between observations, more imputation methods are proposed, such as linear regression imputation, which is suitable for displaying the good linear relationship between variables. There are fewer observations of missing data. These methods effectively retain the internal structure between observations, but the practical range is narrow.

This article innovatively proposes a series of combined methods of interpolation and clustering. The advantage is that for different types of commercial data sets, different combination methods can be adopted to make the clustering effect as unaffected by missing data as possible. This article takes four datasets about Business from UCI as examples and proposes several new missing data clustering methods based on the combination of some different imputation and clustering methods. We use these methods to perform data analysis on the given data sets and compare their differences. Therefore, for different types

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

of data sets in practical applications, we need to adopt suitable processing methods so that the use of clustering algorithms to obtain a good and reliable result.

2. Proposed Methods

In this section, we mainly introduce several missing data analysis methods, mainly based on three types of commonly used imputation methods and three basic clustering methods. Different combination methods will be suitable for data sets with different characteristics [6].

2.1. Mean Imputation & k-means

Mean value imputation, similar to mode imputation, uses the corresponding attribute mean value of the existing complete data to impute the corresponding missing value. Mean value imputation requires data variables to obey or approximately obey a nearly normal distribution. Because unless otherwise stated, the variables of the general data set can be regarded as approximately obeying a normal distribution. The mean filling method is currently the most used method in the filling method and the most extended method based on this method [8].

In actual cluster analysis, k-means clustering as a basic algorithm is widely used [9]. The algorithm flow chart is shown in Figure 1.

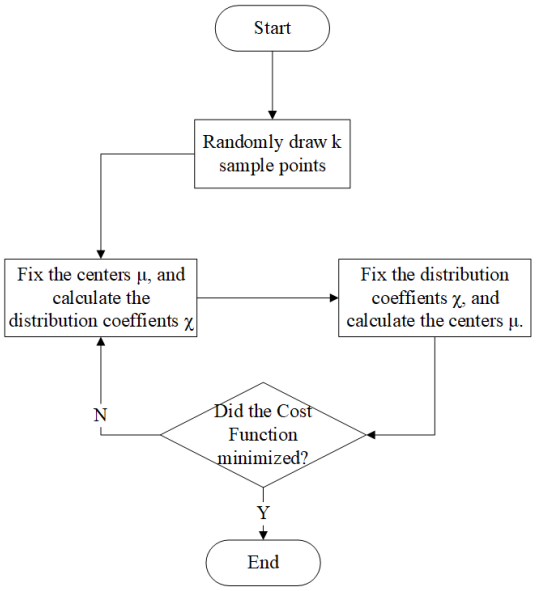

Figure 1: k-means flow chart

Given a data set with n sample points, n is the number of observations in the data set, where the format of each sample point $\{x_1, x_2, \cdots, x_d\} \in \mathbf{R}^d$, d is the number of variables in the data set. The task of k-means clustering is to divide these points into k different classes, so that the similarity of sample points in the same class is high, and the similarity of sample points between different classes is low.

Formally speaking, the k-means algorithm should find the distribution coefficients of all sample points $\chi_{ij} \in \{0,1\}(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant k)$, Where n is the number of sample points, and k is the number of clusters. And at the same time you should find the vector $\mu_j(1 \leqslant j \leqslant k)$, each vector represents the center of a class. If x_i belongs to the j th class, then $\chi_{ij} = 1$, otherwise $\chi_{ij} = 0$. Therefore, the size of the j th class is

$$n_j = \sum_{i=1}^{n} \chi_{ij}. \tag{1}$$

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

In the k-means algorithm, the cost function

$$C = \sum_{i=1}^{n} \sum_{j=1}^{k} \chi_{ij} \| x_i - \mu_j \|^2, \tag{2}$$

Needs to be minimized, where $\| \cdot \|$ refers to Euclidean distance, then

$$\chi_{ij} = \begin{cases} 1, & j = \underset{1 \leq j \leq k}{\operatorname{argmin}} \| x_i - \mu_j \|^2 \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

$$\mu_j = \frac{\sum_{i=1}^{n} \chi_{ij} x_i}{\sum_{i=1}^{n} \chi_{ij}} = \frac{\sum_{i=1}^{n} \chi_{ij} x_i}{n_j}, \tag{4}$$

Derived from eq.(2).

The Lloyd algorithm for k-means clustering is one of the most commonly used algorithms. The Lloyd algorithm uses a batch process to find the values of χ_{ij} and μ_j in two stages.

First fix the center μ_j of all classes, use the formula eq.(3) to update χ_{ij}; then fix all χ_{ij}, use the eq.(4) to update μ_j. The Lloyd algorithm repeats these two stages until the result converges, ensuring that eq.(2) converges to a local minimum [10].

The batch update method of the Lloyd algorithm is more computationally efficient, but the output of the algorithm is often unbalanced, and it will converge to a local minimum that does not meet the requirements due to improper selection of the initial point.

2.2. SVD Imputation "Big Data" k-means

The principle of SVD (Singular Value Decomposition) imputation is based on the SVD decomposition of the data set X

$$X = UDV' \tag{5}$$

To save the main information of the matrix in a low-dimensional structure, and then use

$$\min_{Z} \sum_{\text{Observed } (i,j)} (X_{ij} - Z_{ij})^2 \quad \text{subject to } \| Z \|_* \leq \tau \tag{6}$$

To calculate repeatedly for iteration, and the missing information in the iteration process is restored, so that the corresponding data of the restored matrix Z is the result of SVD imputation. Compared with mean value imputation, SVD imputation has better imputation effect for variables with specific properties, such as low-noise time series data [11]. In actual data analysis, similar dimensionality reduction methods can be selected for filling according to the situation, such as PCA[12] and LDA[13].

For the basic k-means clustering algorithm, since the initial point is randomly selected, if the initial point is not selected properly, it will easily cause the the situation where the best results are not achieved. In response to this problem, we introduce the "Big data" k-means in the first place. "Big data" k-means reduces the influence of random selection of initial points. Consider performing multiple experiments on the original algorithm, measure the clustering effect through appropriate indicators, and select the clustering result with the most occurrences as the result. The algorithm flow chart is shown in Figure 2.

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

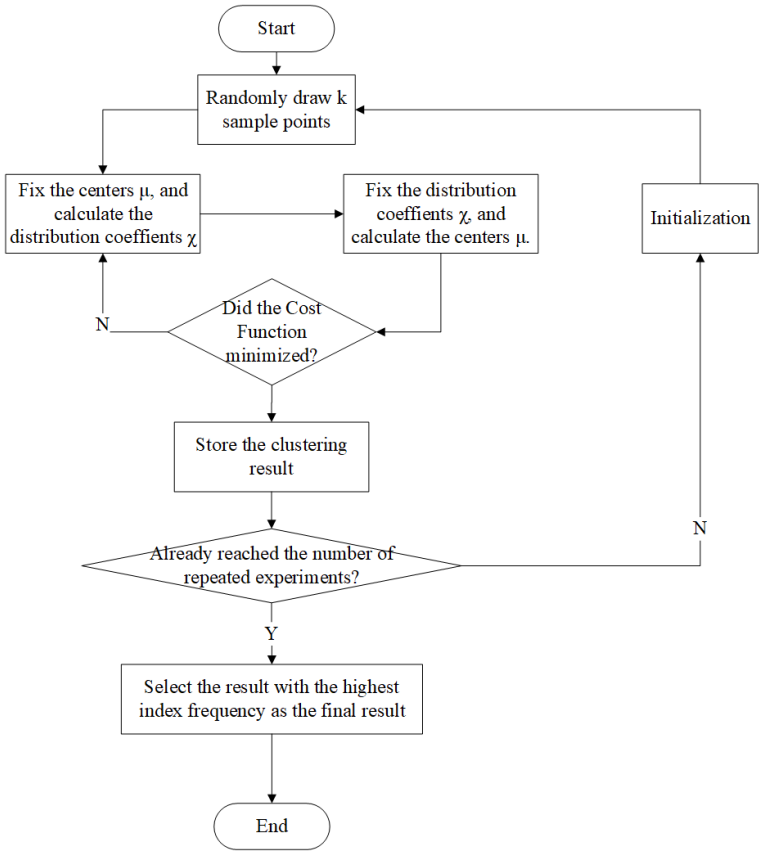

Figure 2: "Big data" k-means flow chart

2.3. kNN Imputation & p-k-means

The kNN (k-Nearest Neighbor) imputation method is a missing data imputation algorithm based on the kNN algorithm. According to the distance measurement or correlation analysis, select the k samples closest to the missing sample, and weight the k sample data to estimate the missing data of this sample is a common method to solve the problem of missing data.

Assuming that the distance between two samples x_a and x_b is $\mathrm{d}(x_a, x_b)$, it can handle both discrete variables and continuous Variable distance metric: heterogeneous Euclidean-overlap metric, which can be referred to as HEOM for short, $\mathrm{d}(x_a, x_b)$ can be defined as

$$\mathrm{d}(x_a, x_b) = \sqrt{\sum_{j=1}^{p} d_j (x_{aj}, x_{bj})^2}, \tag{7}$$

Where $d_j(x_{aj}, x_{bj})$ represents the distance of the j th variable between the sample x_a and x_b.

For the kNN imputation method, if the j variable of the sample x_a has a missing value, choose the k samples closest to the sample x_a. The j th variable of the k sample has no missing values, and the set of the distance from the sample x_a to the farthest of the k samples is set as

$$\vartheta_{x_a} = \{v_j\}_{j=1}^{k}, \tag{8}$$

Where v_1 is the sample closest to the sample x_a.

For the use of kNN algorithm for missing value imputation, it is mainly to select k nearest neighbor

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

samples from the training set, and estimate the imputation value of the missing sample by weighting the nearest neighbor samples[6].

The second idea to reduce the effect of initial centers is called the p-k-means algorithm. This method directly abandons the random initial points, but selects the most scattered points in the data set, so as to ensure that the initial cluster centers will not appear with two or more centers in the same family. The algorithm flow chart is shown in Figure 3.

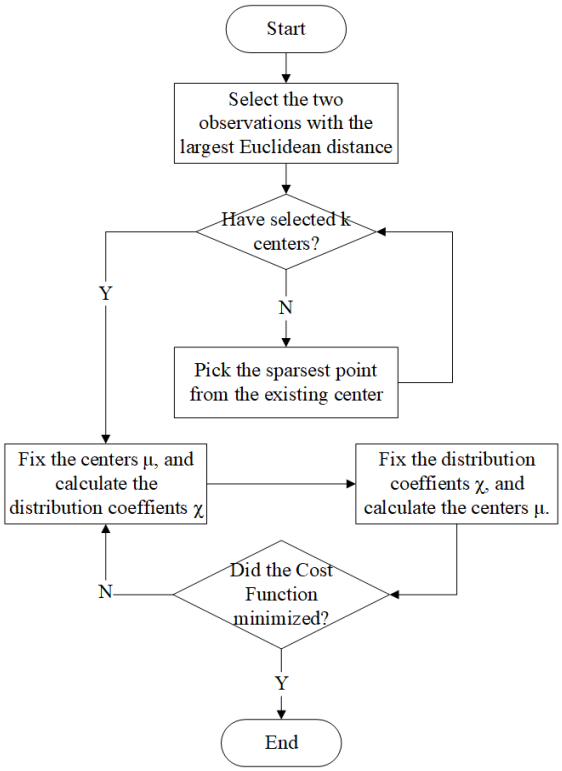

Figure 3: p-k-means algorithm flow chart

For $n(>k)$ observations, d variable data set $X=\{x_1,x_2,\cdots,x_n\}$, where $x_i=(x_{i1,},x_{i2},\cdots,x_{id})$, the steps for selecting the initial cluster centers are as follows.

The first step is to select the two observations with the largest Euclidean distance in the data set as the first and second clustering points, i.e.

$$\mu_1,\mu_2=\underset{1\leqslant i<j\leqslant n}{\operatorname{argmax}}\parallel x_i-x_j\parallel, \tag{9}$$

If the required number of clusters is 2, then the selection ends; otherwise, let

$$M^{(1)}=\{\mu_1,\mu_2\},X^{(1)}=\{x_1,x_2,\cdots,x_n\}\setminus M^{(1)}, \tag{10}$$

Which means to exclude the two selected centers, the third cluster center is

$$\mu_3=\frac{1}{2}\Big(\underset{x\in X^{(1)}}{\operatorname{argmax}}\{\parallel\mu_1,x\parallel+\parallel\mu_2,x\parallel\}+\underset{x\in X^{(1)}}{\operatorname{argmin}}\{\mid\parallel\mu_1,x\parallel-\parallel\mu_2,x\parallel\mid\}\Big); \tag{11}$$

Generally, for the selection of the j th $(4\leqslant j\leqslant k)$ center point, let

$$M^{(j-2)}=M^{(j-3)}\cup\{\mu_{j-1}\},X^{(j-2)}=X^{(j-3)}\setminus M^{(j-2)}, \tag{12}$$

Then

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

$$\mu_j = \frac{1}{2}(\alpha_j + \beta_j),\ \text{where} \begin{cases} \alpha_j, \mu_j^{(1)}, \nu_j^{(1)} = \underset{x \in X^{(j-2)}, \mu, \nu \in M}{\operatorname{argmax}} \{ \| \mu, x \| + \| \nu, x \| \}, \\ \beta_j, \mu_j^{(2)}, \nu_j^{(2)} = \underset{x \in X^{(j-2)}, \mu, \nu \in M}{\operatorname{argmin}} \{ \| \mu, x \| - \| \nu, x \| \}). \end{cases} \quad (13)$$

The advantage of this improvement is that the amount of calculation is much less than that of the previous method, and the optimal solution can also be achieved [14].

2.4. Other combinations of imputation and clustering

Similarly, we can get the following methods through appropriate combinations: SVD Imputation & k-means, kNN Imputation & k-means, mean Imputation & "Big Data" k-means, kNN Imputation & "Big Data" k-means, mean Imputation & p-k-means and SVD Imputation & p-k-means. For the actual data set, we should analyze its internal structure and the characteristics of missing data, and choose an appropriate combination method of imputation and clustering.

3. Experiment Procedure

3.1. Description of the Dataset

We use R language for programming, and four datasets analyzing. The four datasets are Absenteeism at work [15], Facebook Live Sellers in Thailand [16], Online Shoppers Purchasing Intention [17], and Wholesale customers [18], which is obtained from the UCI database. The description of datasets can be seen in table 1.

Table 1: Dataset Description

Dataset	Observations	Variables	Classes
Absenteeism at work(D1)	740	21	4
Facebook Live Sellers in Thailand (D2)	7050	16	4
Online Shoppers Purchasing Intention (D3)	12330	18	2
Wholesale customers (D4)	440	8	2

3.2. Method comparison

In order to compare the effect, this article uses the complete data set for random missing. When the code is randomly lost, the number of missing variables for observations with missing data is strictly less than the total number of variables, otherwise the observation is completely lost, for data with too much data we only select part of the data for analysis. Then we use different combination methods of imputation and clustering.

3.3. Evaluation Index

Since in the experiment, the complete data is randomly lost and then the imputation and clustering is performed. But for general data sets with missing values, this indicator cannot be used obviously.

We can also use the SSE of the clustering for comparison. Generally, the smaller the SSE is, the better the clustering effect preforms [20].

The third indicator is Silhouette Coefficient [21]. The value of this indicator is close to 1, the better the clustering effect is. See the comparison of three indicators the experiment in Figure 4, note that the value in SSE is the ratio to Mean's SSE.

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

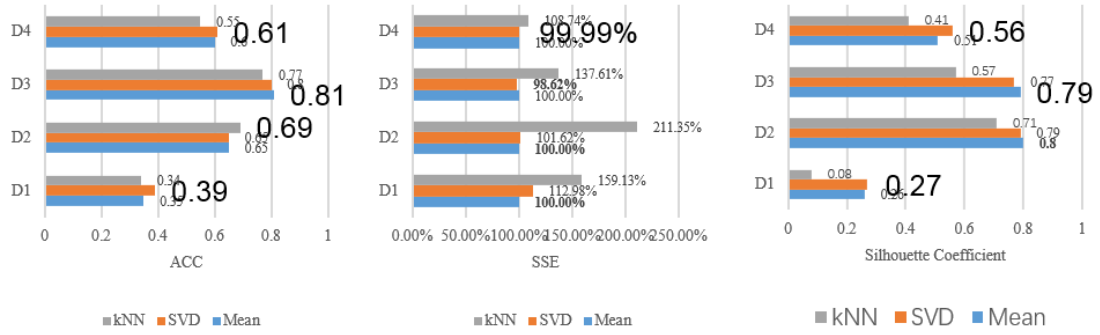

Figure 4: Accuracy, SSE and Silhouette Coefficient of different ways of imputation

4. Conclusion

This paper proposes a series of imputation and clustering combination methods through the processing and analysis of business datasets. These combined methods have different effects in overcoming the impact of missing values on cluster analysis on specific business datasets. This article also evaluates the classification effect of the combination methods through three different cluster evaluation indicators. In the actual application of unsupervised machine learning to cluster analysis of business data, the impact of missing values on the information structure of the data set cannot be ignored. Therefore, first we need to analyze the characteristics of the missing values of the dataset and predict which methods may achieve good imputation effects. Secondly, try several times of imputation and clustering combination methods and visualize the clustering effect. Finally select the best combination method through appropriate indicators to ensure the accuracy and availability of business data.

References

[1] MA, Zongfang, LIU Zhe, et al. "Credal Transfer Learning With Multi-Estimation for Missing Data." IEEE Access 8 (2020), pp. 70316-70328.
[2] XIONG Zhongmin, GUO Huaiyu, and WU Yuexin. "Review of Missing Data Processing Methods". In: Computer Engineering and Applications 57.14 (2021), pp. 27–38.
[3] Pedro J Garcı´ea-Laencina et al. "K nearest neighbours with mutual information for simultaneous classification and missing data imputation". In: Neurocomputing 72.7-9 (2009), pp. 1483–1493.
[4] Roderick JA Little and Donald B Rubin. Statistical analysis with missing data. Vol. 793. John Wiley & Sons, 2019.
[5] CHEN Wanjiao. "Research on Application of Missing Data Imputation in Medical Field". In: South China University of Technology (2019).
[6] HAN Jiawei, PEI Jian, and Kamber Micheline. Data mining: concepts and techniques. Elsevier, 2011.
[7] James MacQueen et al. "Some methods for classification and analysis of multivariate observations". In: Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. Vol. 1. 14. Oakland, CA, USA. 1967, pp. 281–297.
[8] ZHOU Wang, ZHANG Chenlin, and WU Jianxin. "Qualitative balanced clustering algorithm based on Hartigan-Wong and Lloyd". In: Journal of Shandong University (Engineering Science) 46.05 (2016), pp. 37–44.
[9] T. Olga et al. "Missing value estimation methods for DNA microarrays". In: Bioinformatics 17.6 (2001), pp. 520–525.
[10] Julie Josse and Fran¸cois Husson. "Handling missing values in exploratory multivariate data analysis methods". In: Journal de la Soci´et´e Fran¸caise de Statistique 153.2 (2012), pp. 79–99.
[11] Jaap Brand. Development, implementation and evaluation of multiple imputation strategies for the statistical analysis of incomplete data sets. 1999.
[12] ZANG Chuanyu et al. "Research on K-Means Algorithm Analysis and Improvement". In: Computer Science and Application 6.9 (2016), p. 14.
[13] A. Martiniano et al. "Application of a neuro fuzzy network in prediction of absenteeism at work". In: Information Systems and Technologies (CISTI), 2012 7th Iberian Conference on. 36 vols. 18. 2012.
[14] N. Dehouche. "Dataset on usage and engagement patterns for Facebook Live sellers in Thailand". In: Data in Brief 30 (2020), p. 105661.
[15] C. O. Sakar et al. "Real-time prediction of online shoppers' purchasing intention using multilayer

ISSN 2616-5775 Vol. 4, Issue 6: 35-42, DOI: 10.25236/AJCIS.2021.040606

perceptron and LSTM recurrent neural networks". In: Neural Computing and Applications (2018).
[16] Nuno Gon¸calo Costa Fernandes Marques de Abreu et al. "Analise do perfil do cliente Recheio e desenvolvimento de um sistema promocional". PhD thesis. 2011.
[17] HONG Qing et al. "Video user group classification based on barrage comments sentiment analysis andclustering algorithms". In: Computer Engineering & Science 40.06 (2018), pp. 1125–1139.
[18] Peter J Rousseeuw. "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis". In: Journal of computational and applied mathematics 20 (1987), pp. 53–65.