MI-K-MEAN ALGORITHM: A NEW APPROACH FOR FINANCIAL RISK ANALYSIS WITH MISSING DATA IMPUTATION IN BIG DATA

Ravindra Kumar[1], Diwakar Shukla[2], Kamlesh Kumar Pandey[3]
[1]Department of Computer Science and Application, Dr Harisingh Gour Vishwavidyalaya Sagar, M.P.,India
[2]Department of Mathematics and Statistics, Dr Harisingh Gour Vishwavidyalaya Sagar, M.P., India
[3]Department of Vocational Education, Indira Gandhi National Tribal University Amarkantak, M.P., India
[1]chakravarti.ravindra@gmail.com, [2]diwakarshukla@rediffmail.com, [3]kamleshamk@gmail.com

Abstract

The data mining is a tool of searching information from the data warehouse. Several mining algorithms exist in literature, one of the most common is the usual K-mean procedure. This generates centroids after every round of iteration. It is assumed that sample data is completely cleaned and noise free before the start of execution of the usual K-mean algorithm. If α% values are missing in sample data then after cleaning only (100-α)% values are available for the execution of the usual K-mean algorithm. Such bears a loss of information that affects the decision. This paper considers this problem and resolves such issue by replacing the missing data through imputed values calculated by the available values, called Mean Imputation (MI). It helps in financial risk analysis quite a lot because of risk prediction being taken on a larger sample (cleaned and imputed both). Several imputation procedures are available in literature. This paper considers the financial risk data as sample where the missing values of sample are imputed by the usual Mean-Imputation (MI) method and then on complete sample. Proposed MI-K-mean strategy is compared with no imputation usual procedure and found more efficient over the four-evaluation criterion of cluster formation while applying on risk data analysis.

Keywords: Missing Data, Mean Imputation (MI), Credit card risk, K-mean clustering, Big data

1. Introduction

Financial risk calculation is used to bifurcate the customer as per the account information in a bank. It is the possibility of potential losses in direct investments caused by the effects of corporate credit, tax financing, other economic factors and corresponding economic shocks. Risk computation is an important method to provide the general description about a customer as per the credit score, which helps to the bank manager for taking the decision about distributing the loan. Financial risk is a measure to identify and analyze the existing financial risk factors, determine the likelihood and severity of probable new risks, and it provide scientific basis for risk for evaluation prevention and control [18]. Loan risk analysis plays a vital role among banking system where bank can identify the customer those who are exposed with good and bad risk. For the decision-making process the human analysis is more complex for large amount of financial data. Financial risk [23] includes risk identification, risk assessment and risk treatment. Risk identification and assessment is a part relating to evaluate the account of a customer for the financial risks and their sources through account details. Moreover, qualitative and quantitative methods are important to measure the size of the risk and generating the risk warning.

Data mining provides the general description of the data for the analysis to predict the values and to forecast the solution for making the decisions on it [31]. Data pre-processing is an important step for data analysis used to clean the data for removing the noise. This is time-consuming task to perform some calculation over data.

Missing data calculation is very tedious task for finding the location and manage them. For missing data handling, the imputation is the process for substituting the value in place of missing value [9]. Imputation methods can be included with various statistical methods like mean, median and mode. Imputation [10] is commonly used in computer science and related fields for several reasons such as handling non responded data, preprocessing data, maintaining dataset integrity, improving model performance and data analysis for visualization. Maintaining Dataset Integrity, in many applications, maintaining the integrity of the dataset is vital. Removing rows or columns with missing values [14] might result in a significant loss of data, reducing the representativeness and potential insights that can be obtained. Imputation allows for retaining the maximum amount of information available in the dataset. Imputation can also lead to improved model performance by reducing the potential bias and noise introduced by missing data. By imputing missing values [26], models can utilize the complete dataset to learn patterns, relationships, and make more accurate predictions.

Clustering is a technique to find the homogeneity of the particular group of objects. Cluster analysis is very useful in big data to category of studied object is not known in advance, to group the similarities into a particular category based on the degree of affinity therefore the same category can achieve the maximum similarity and minimize the dissimilarities. Moreover, the different categories achieve the maximum homogeneity and minimum heterogeneity. Cluster model selection is the process involved as per objective of the problem [25]. Objective of the problem define as per domain and may be vary as per the model selection so that the model selection is important concern for the prediction. Clustering analysis method can be categorized into three different types: trying to calculate an optimal data partition [15] to divide the given data into a specific number of clusters; trying to find out a method for the cluster structure; and trying to find a method based on statistical model for potential cluster modelling.

The K-means [12] cluster analysis technique effectively ignore the subjective negative impact caused by the artificial threshold value and ignores the missing data aspect, therefore it can more accurately and objectively describe the state intervals of different financial risks. On the basis of previous summary and analysis, this paper provide the current research status and significance of financial risk using the imputation and k-means [17] algorithm, elaborated the development background, current status and future challenges of the K-means clustering algorithm using imputation method, introduced the related works of similarity measure and item clustering with imputation[13], proposed a financial risk indicator system based on the K-means[20] clustering algorithm, performed evaluation parameter and data processing, constructed a financial risk based model based on the K-means [28] clustering algorithm with imputation [19], the dataset stored the values in credit card [33]. Study results of this paper provide a reference for further researches on financial risk based on K-means [22] clustering algorithm with imputation and the removal of the data in big data mining.

This paper is organized in nine sections. Second contains technical part of background of big data, imputation, clustering and evaluation methodologies. The third section is based on main problem undertaken in the paper while fourth section is based on motivation and hypothesis creation of research. The solution as in the form of proposed procedure is in section 5 whereas section 6 is with a new MI-K mean algorithm which is crux of this study. Section 7 reveals the flow of execution of this new algorithm and Section 8 supports the outcomes with numerical data. The last section 9 contains conclusion of all findings in a nutshell.

2. Background Technical Aspect

2.1 Big Data

Big Data cluster is a term for a collection of datasets who are so large and complex that it becomes difficult to process using on-demand database management tools or using traditional data processing applications [21]. Big data can be classified mainly in three basic categories like Volume, Variety and Velocity. This large volume of data continuously increases day by day by using electronic gadgets and through web-based platforms. The social media is a major source of big data [24] and others sources are like medical, insurance, marketing, weather forecasting etc. The main source are social media like Facebook, WhatsApp, Twitter etc. where at every second the volume of data increases drastically. In a day data generate with different forms of text messages, audio-recording, images, videos, log files etc. Big data parameter is important to discuss along with challenges [13] of this technology.

2.2 Missing Data types, Techniques and Classification

Data collection and analysis is the major part in for research and development. This step be performed very carefully however due to the large data there is chance to miss any value for the entry or missing due to any other reason [7]. Missing observation has mainly three patterns MAR, MCAR, MNAR [32]. For handling the missing values in datasets various strategies exist like try to find out the missing data [9], leave out the incomplete data and go-ahead for the next step, replace the missing data [27] as per the mean value etc.

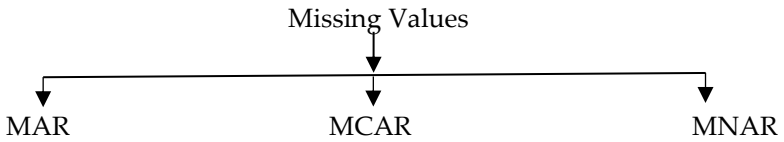

Figure 1. *Missing Data Types*

(a) MAR (Missing at Random): MAR [30] values give the same value in the particular group which it is belongs to the observed data.

(b) MCAR (Missing Completely at Random): In MCAR [32] finding the missing values in the same for all cases where the values are not available in the observation.

(c) MNAR (Missing Not at Random): MNAR is the missing value unknown to us, it is very difficult to finding in the observation.

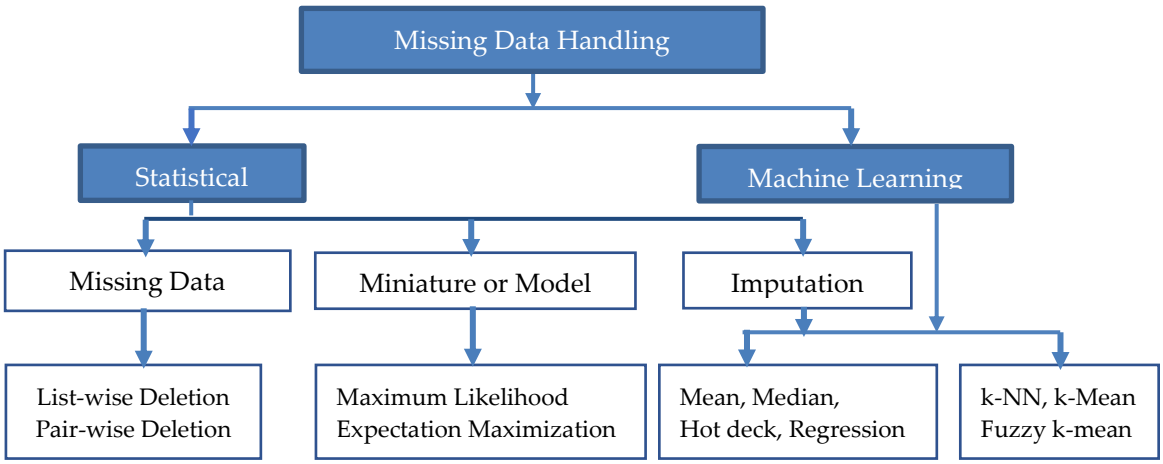

Figure 2: *Classification of Missing data handling techniques*

2.3 Data Clustering

Clustering is the method of a given data points, partition them into a set of groups which are as similar as possible. Data Clustering technique to find the complete or incomplete data clustering [8], is a data exploration technique used in various fields, including data science and machine learning. It involves grouping a set of data points into clusters, objects are similar to each other and performed the Big data clustering [11] in the dataset. The K-means [14] algorithm is the most frequently used clustering method. Moreover K-means [16] clustering algorithm is to select K data as the initial centroid of each category and divide them into K categories according to the principle of one category with the smallest distance, and then further the divided mean values are judged according to the square error criterion function.

Thakur and Shukla [1] proposed the missing data estimation based on the chaining technique in survey sampling. The method section included estimation, missing data, chaining, imputation, bias, mean squared error (MSE), factor type (F-T), chain type estimator, double sampling.

Thakur et al. [5] presented some new concept on mean estimation with imputation using two-phase sampling design. In this paper a imputation using in a sample survey in presence of missing data and one of the substitution techniques of missing observations is applied.

Shukla et al. [2] proposed some new aspects on imputation using sampling. Methods included estimation, missing data, imputation, bias, mean squared error (m.s.e.), compromised estimator, factor-type compromised imputation (FTCI). The number of causes that affect the quality of survey and missing data is one of such that keeps sample incomplete.

Pandey and Shukla [3] deployed a new approach on stratified linear systematic sampling-based clustering approach for detection of financial risk group by mining of big data. Risk analysis is beneficial for taking the business decision for finding the unknown risks such as credit risk, debit risk, operational risk and financial risk.

Jager et al. [6] integrated a benchmark for data imputation methods. This paper provides the detailed information about the missing data and its categories such that MAR, MCAR and MNAR. The method section included data quality, data cleaning, imputation, missing data, benchmark, MCAR, MNAR, MAR. The data preprocessing method selection for automated data quality improvement.

Pandey et al. [4] employed max-min distance sort heuristic-based initial centroid method of partitional clustering for big data mining. The methods included big data clustering, Initial centroid algorithm, convergence speed, stratified sampling, K-means, K-means++, MDSHK-means.

2.4 Cluster Evaluation Parameters

(a) Silhouette Score: The silhouette score is a measurement of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). For the calculation using the mean intra-cluster distance (a) and the mean nearest-cluster distance (b) for each sample.

$$\text{Silhouette Coefficient} = (b - a) / \max(a, b)$$

$$\text{Percentage Gain} = \frac{|\text{Strategy B score} - \text{Strategy A Score}|}{\text{Strategy A}} \times 100$$

(b) Devies Bouldon Score (DBS): This Devies Bouldon Score is calculated as the average similarity measure of each cluster with its most similar cluster, where similarity is the ratio of within-cluster distances to between-cluster distances, which is simply the average of the similarity measures

of each cluster with a cluster most similar.

$$\bar{R} = \frac{1}{N}\sum_{i=1}^{N} R_i$$

$$\text{Percentage Gain} = \frac{|\text{Strategy A score} - \text{Strategy B Score}|}{\text{Strategy A}} \times 100$$

(c) Mutual Information Score (MIS): Mutual Information is a measure of the similarity between two labels of the same data. Where $|U_i|$ is the number of the samples in cluster U_i and $|V_j|$ is the number of the samples in cluster V_j, the Mutual Information between clustering U and V is given below.

$$MI(U,V) = \sum_{i=1}^{|U|}\sum_{j=1}^{|V|}\frac{|U_i \cap V_j|}{N}\log\frac{N|U_i \cap V_j|}{|U_i||V_j|}$$

$$\text{Percentage Gain} = \frac{|\text{Strategy B score} - \text{Strategy A Score}|}{\text{Strategy A}} \times 100$$

(d) Rand Index Score: Rand Index is calculating a similarity between two cluster results by taking all points identified within the same cluster. This value is equal to 0 when points are assigned into clusters randomly and it equals to 1 when the two cluster results are same.

$$\text{RI} = \frac{\text{Number of pairs in same cluster(actual) X Number of pairs in same cluster(predicted)}}{\text{Total number of possible pairs}}$$

$$\text{Percentage Gain} = \frac{|\text{Strategy B score} - \text{Strategy A Score}|}{\text{Strategy A}} \times 100$$

(e) Adjusted Rand Index (ARI): ARI is used to measure the similarity between two clustering by considering all the pairs of the n samples and calculating the counting pairs of the assigned in the same or different clusters in the actual and predicted. E is indicating Expected.

$$\text{ARI} = \frac{\text{Number of pairwise true positive prediction} - \text{E[RI]}}{\text{Average number of pairs in same cluster for actual and predicated} - \text{E[RI]}}$$

$$\text{Percentage Gain} = \frac{|\text{Strategy B score} - \text{Strategy A Score}|}{\text{Strategy A}} \times 100$$

2.5 Mean Imputation (MI)

Step I: Take sample of n observations.
Step II: Find missing values in dataset (out of n).
Step III: Let k (k<n) values of dataset are found missing.
Step IV: Find mean of (n-k) values in sample data. Let it is denoted as $\bar{x}^*$.
Step V: Replace all missing values in the dataset by $\bar{x}^*$.

3. Problem Undertaken

This paper aims to explore about the application of imputation techniques over clustering methods applicable to financial risk calculation in the big data environment. Cluster evaluation parameters evaluates the cluster accuracy and provide the efficient result for creating the clusters. This paper aspires to contribute the existing literature by providing efficient evidence and theoretical insights for data cluster calculation in the financial risk data setup when missing data is replaced by the imputed values. In view of combination of clustering and imputation need new algorithm which is a problem considered herein what follows.

4. Motivation

The data cleaning procedure reduces the sample size of financial risk data by eliminating noise presence therein. Noise may be in the term of missing values. One can think of that if such are replaced using the known values by an appropriate imputation method then larger sample size will be available for applying the usual K-mean algorithm which may produce efficient result of financial risk clustering. The financial risk is dangerous and require large data size for prediction.

4.1. Hypothesis

(a) Is there significant effect of imputed data against missing observations on the cluster evaluation parameters?
(b) Comparing risk reduction for imputed sample with the cleaned sample.
(c) Is the risk a decreasing function of imputed values in sample?

5. Proposed Procedure

Two strategies are given below.

Strategy A: A new algorithm is proposed named after "MI-K-mean algorithm" which considers entire sample data n (using imputation).

Strategy B: Usual K-mean algorithm applicable over only cleaned data which is less in sample size due to cleaning.

This paper presents a comparison between Strategy A (Proposed) and Strategy B (usual method) for data mining. The step-wise execution of algorithm is as under:

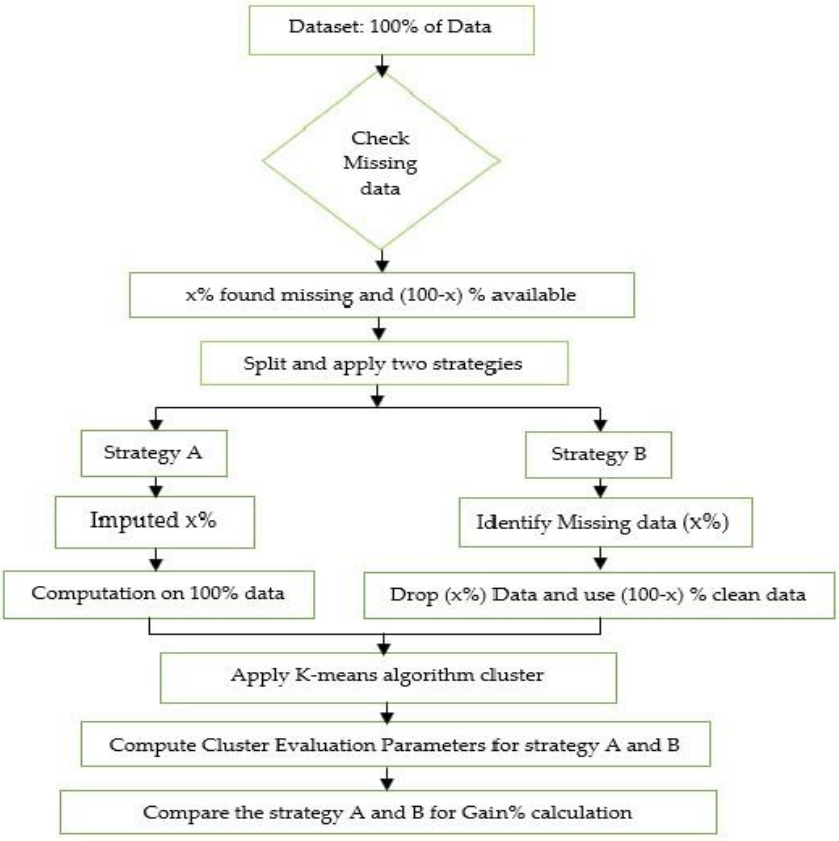

Figure 3: *Basic Model for workflow of proposed method with imputation and usual method*

The figure 3 contains two strategies A and B for the cluster formation by K-means algorithm on financial risk data. Strategy A uses replacement of missing values by an appropriate imputation method (MI-imputation) while strategy B contains cleaned data after eliminating the missing. The sample size for B is smaller that strategy A.

6. Proposed MI-K-mean Algorithm (Step-wise)

Imputation and clustering based proposed strategy A in order to find cluster is as under

(i) Input

1. N = {a_1, a_2, a_3, ……a_n} is the data points to the financial risk-based D dataset

2. K= Required number of clusters.

(ii) Output

C = {c_1, c_2, c_3 ……c_n}

(iii) Dataset Description

1 Dataset Head: data.head() [35]

2 Dataset Shape: data.shape() [35]

3 Dataset Statistical description: data.describe() [35]

(iv) Missing values Identification

Method for finding missing values (null values) in dataset: data.isnull.sum() [35]

(v) Dropping the Missing Values (Removal of the data)

Methods for deleting the missing data

(a) Row wise deletion: data.dropna(axis=0) [35]

(b) Column wise deletion: data.dropna(axis=1) [35]

(vi) Imputation: Mean Imputation

data['Column_name']= data['Column_name'].fillna(data['Column_name'].mean()) [35]

(vii) Clustering K-mean algorithm

1 K-Means Clustering (Un-supervise Clustering Method)

2 Select k random values

3 Find out the optimality of clusters using Elbow or Silhouette method

4 Calculate the cluster centers and centroid

5 Find out the clusters

(viii) Cluster Evaluation

1 Calculation of Silhouette Score (SC) for the cluster evaluation

2 Calculation of Devies Bouldon Score (DBS) for the cluster evaluation

3 Calculation of Mutual Information Score (MIS) for the cluster evaluation

4 Calculation of Rand Index (RI) for the cluster evaluation

5 Calculation of Adjusted Rand Index (ARI) for the cluster evaluation

7. Implementation Procedure of MI-K-algorithm

Model for implementation of the proposed Strategy A needs several steps in data cleaning process such as missing value identification and performed imputation methods to obtain clusters on financial risk data.

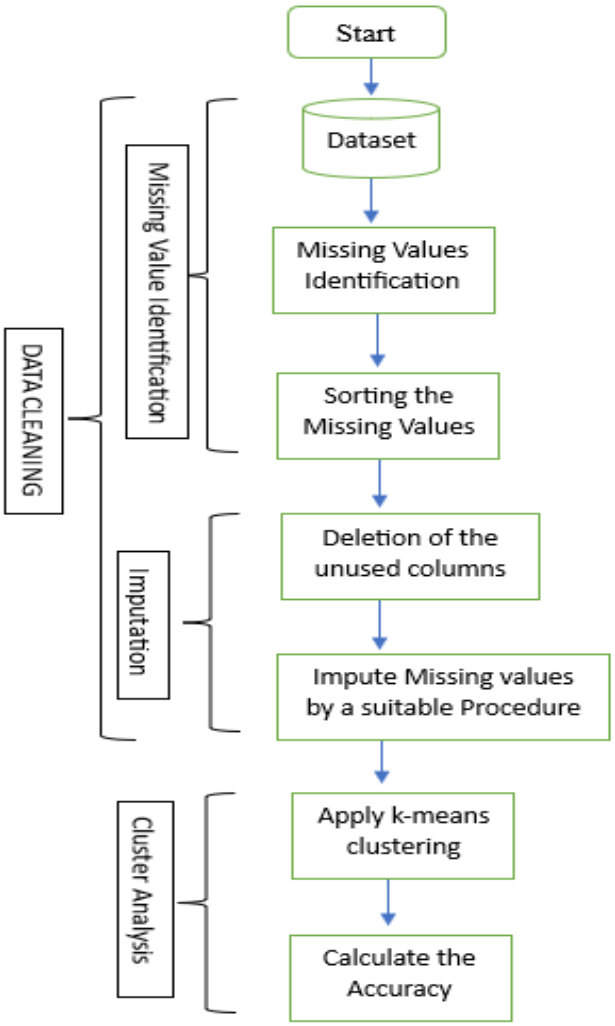

Figure 4: *Implementation model for proposed method (Strategy A)*

8. Empirical Analysis

An Empirical study has been performed for applying efforts on the computing environment, datasets, existing algorithm, evaluation criteria and results.

(A) Experiment Environment and Credit Card General Loan Risk Dataset [34]

Table 1: *Description of the Credit Card General Dataset*

ID	Dataset	Objects	Attributes	Class	Data source
data	CC General	8950	18	2	www.kaggle.com

(B) Computing Environment

The computing environment for the proposed clustering approach using mean imputation technique is developed in Anaconda Navigator (anaconda 3) Jupyter and Google Collab notebook. The experimental environment is configured with an Intel(R) Core (TM) i5-2430M CPU @ 2.40GHz, 256 GB SSD, 4GB DDR3 RAM, Windows 10 Pro, Python 3.10.11, Microsoft Edge browser.

(C) Results and Discussion

Table 2: *Original Dataset Credit Card General Data Analysis*

ind	CUST_ID	BALANCE	BALA	PURCHAS	ONEC	INST	CASH_A	PURC	ONEC	PURC	CASH	CA	PU	CREI	PAYME	MINIM	PRC_	TENURE
0	10001	40.90	0.82	95.40	0	95	0.00	0.17	0.00	0.08	0.00	0	2	1000	201.80	139.51	0.00	12
1	10002	3202.47	0.91	0.00	0	0	6442.95	0.00	0.00	0.00	0.25	4	0	7000	4103.03	1072.34	0.22	12
2	10003	2495.15	1.00	773.17	773.2	0	0.00	1.00	1.00	0.00	0.00	0	12	7500	622.07	627.28	0.00	12
3	10004	1666.67	0.64	1499.00	1499	0	205.79	0.08	0.08	0.00	0.08	1	1	7500	0.00	NaN	0.00	12
4	10005	817.71	1.00	16.00	16	0	0.00	0.08	0.08	0.00	0.00	0	1	1200	678.33	244.79	0.00	12

Table 2 shows the description of the sample dataset [Total 19 columns and five rows result using head() method]. Actual analysis performed over 8950 rows.

Table 3: *Data Size*

data.shape	Rows
Before Removal Shape Size	8950
After Removal Shape Size	8636

Table 3 shows that there are total 8950 row and after noice removal (cleaning) 8636 rows remained.

Table 4: *Reduced data for analysis*

index	CUST_ID	BALANCE	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	TENURE
0	10001	40.90	1000	201.80	139.50	12
1	10002	3202.46	7000	4103.03	1072.34	12
2	10003	2495.14	7500	622.06	627.28	12
3	10004	1666.67	7500	0	NaN (Missing)	12
4	10005	817.71	1200	678.33	244.791	12

Table 4 shows that only six columns have been taken for analysis besides that all area available.

Table 5: *Descriptive analysis of dataset*

index	CUST_ID	BALANCE	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	TENURE
count	8636	8636	8636	8636	8636	8636
mean	14477.9188	1601.225	4522.091	1784.478	864.3049	11.5343
std	2565.75979	2095.571	3659.24	2909.81	2372.566	1.3109
min	10001	0	50	0.0495	0.0191	6
25%	12267.75	148.0952	1600	418.5592	169.1635	12
50%	14469.5	916.8555	3000	896.6757	312.4523	12
75%	16698.25	2105.196	6500	1951.142	825.4965	12
max	18950	19043.14	30000	50721.48	76406.21	12

Table 5 shows descriptive analysis of data after removal of missing data.

Table 6: *Count of missing values in sample data*

CUST_ID	0
BALANCE	0
CREDIT_LIMIT	1 (Missing Values)
PAYMENTS	0
MINIMUM_PAYMENTS	313 (Missing Values)
TENURE	0

Table 6 provides information about the total number of missing fields(values) among six columns.

Table 7: *Statistic description of the data*

index	CUST_ID	BALANCE	CREDIT_LIMIT	PAYMENTS	MINIMUM_PAYMENTS	TENURE
count	8950	8950	8949	8950	8637	8950
mean	14475.5	1564.475	4494.449	1733.144	864.2065	11.517
std	2583.787	2081.532	3638.816	2895.064	2372.447	1.338
min	10001	0	50	0	0.019163	6
25%	12238.25	128.2819	1600	383.2762	169.1237	12
50%	14475.5	873.3852	3000	856.9015	312.3439	12
75%	16712.75	2054.14	6500	1901.134	825.4855	12
max	18950	19043.14	30000	50721.48	76406.21	12

Table 7 provides descriptive analysis statistics after the imputation of missing data (using mean imputation)

Table 8: *Silhouette Method for finding the optimal cluster*

k(clusters)	**(Strategy A Proposed)**	**(Strategy B)(Usual)**
k=2	0.4155	0.4000
k=3	0.3681	0.3691
k=4	0.2788	0.2810
k=5	**0.2792**	**0.2812**
k=6	0.1937	0.1919
k=7	0.1875	0.18771
k=8	0.1311	0.1089
k=9	0.1358	0.1234
k=10	0.1147	0.1245

Table 8 shows the optimal cluster is obtained at k=5 using Silhouette Method

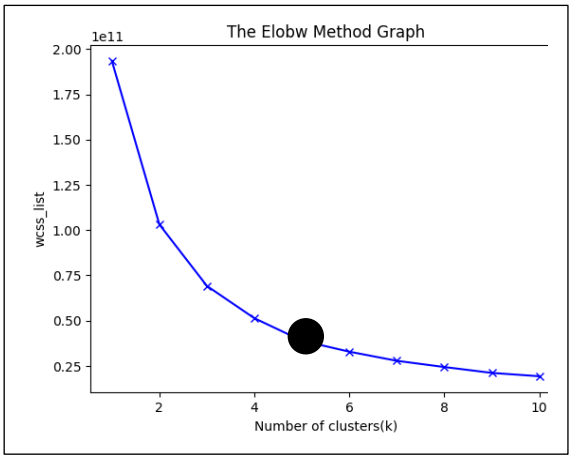

Figure 1. *Strategy A (Elbow method)*

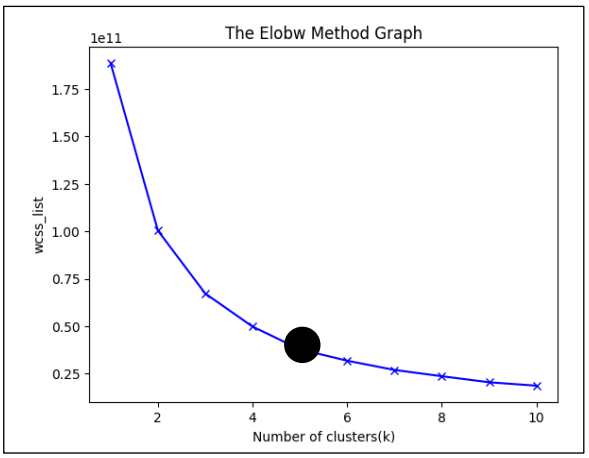

Figure 2. *Strategy B (Elbow method)*

Fig.1 and 2 reveals that graphical representation of the cluster optimality using strategy. Using Elbow method at k=5 the proposed strategy A is better than strategy B.

Table 9: *Centroid or cluster center calculation*

S.No	Strategy A	Strategy B
	Cluster Centers or Centroids	Cluster Centers or Centroids
1	[12751.36, 2347.60]	[6487.49, 1460.97]
2	[2113.72, 892.23]	[13822.78, 22720.13]
3	[13775.00 , 22802.04]	[7121.49, 7406.53]
4	[13775.00 , 22802.04],	[12731.04, 2403.48]
5	[6482.10, 1394.93]	[2103.22 , 923.39]

Table 9 reveals that Centroid obtained for five clusters (k=5) where k denotes the optimal number of clusters

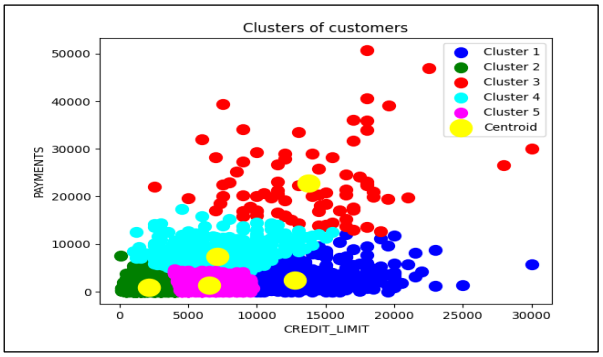

Figure 3: *Strategy A (Cluster representation)*

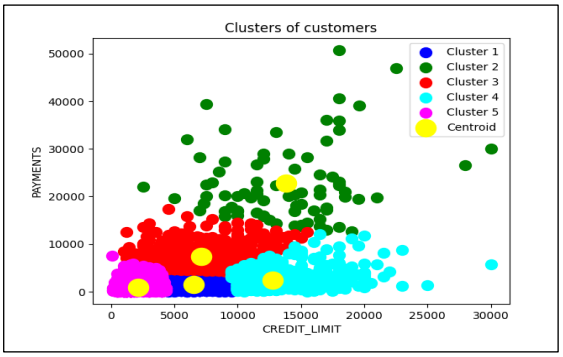

Figure 4: *Strategy B (Cluster representation)*

Comparing figure 3 and 4 one can observe that figure 3 is showing betterment (strategy A) in comparison to figure 4(strategy B) plotted using Matplotlib library.

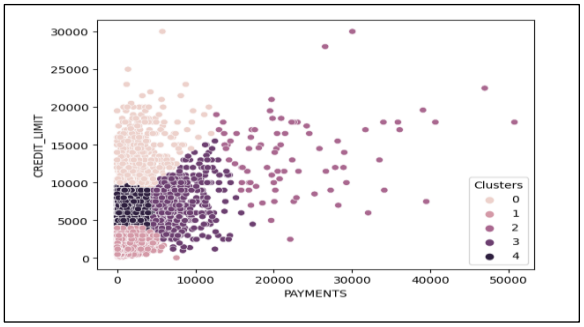

Figure 5: *Strategy A*

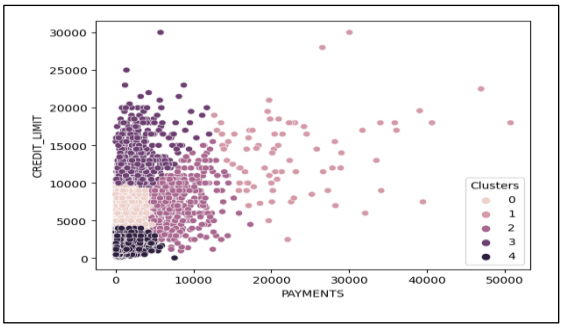

Figure 6: *Strategy B*

Considering figure 5 and figure 6, the strategy A is better than B plotted using Seaborn library.

Table 10: *Cluster evaluation parameters Strategy A and Strategy B*

Cluster Evaluation	Strategy A	Strategy B	Percentage Gain (%)
Silhouette Score	0.287495857	0.303354945	5.51%
Devies Bouldon Score	1.283730623	1.013587896	21.04%
Mutual Information Score	1.075392591	0.381224234	64.55%
Rand Index Score	1	0.655464085	34.45%
Adjusted Rand Index	1	0.276573174	72.34%

Table 10 shows percentage gain due to five evaluation parameters of clusters by Silhouette, Devies Bouldon, Mutual Information, Rand Index and Adjusted Rand Index Score. There is significant percentage gain in four cluster evaluation criterions.

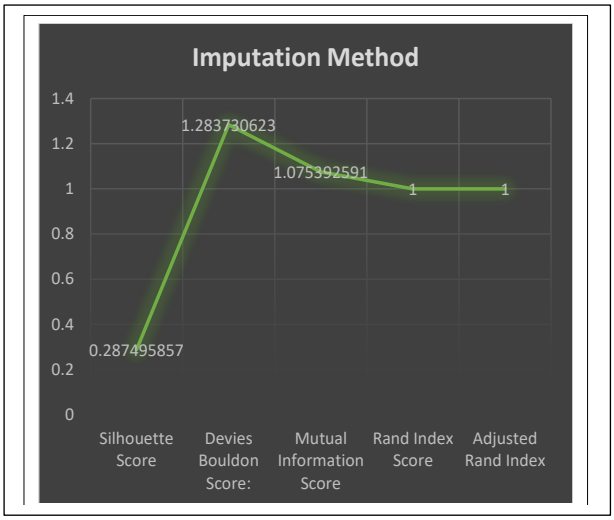

Figure 7: *Strategy A*

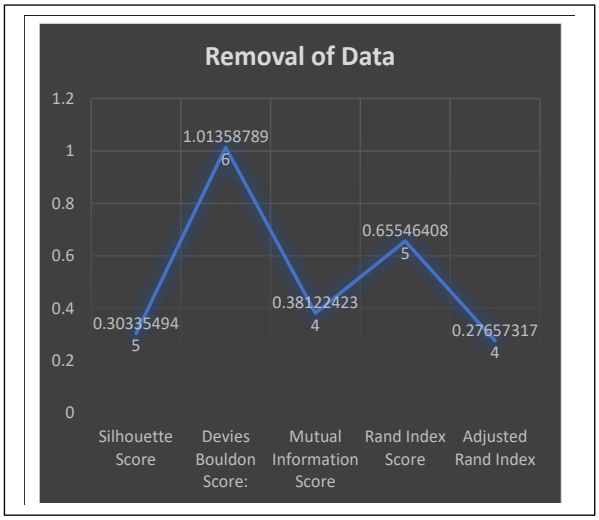

Figure 8: *Strategy B*

Figure 7 and 8 have graphical representation of the evaluation criterion on five parameters. Only first criteria (Silhouette score) bearing the low value but all other four evaluations showing gain, so strategy A is better than strategy B.

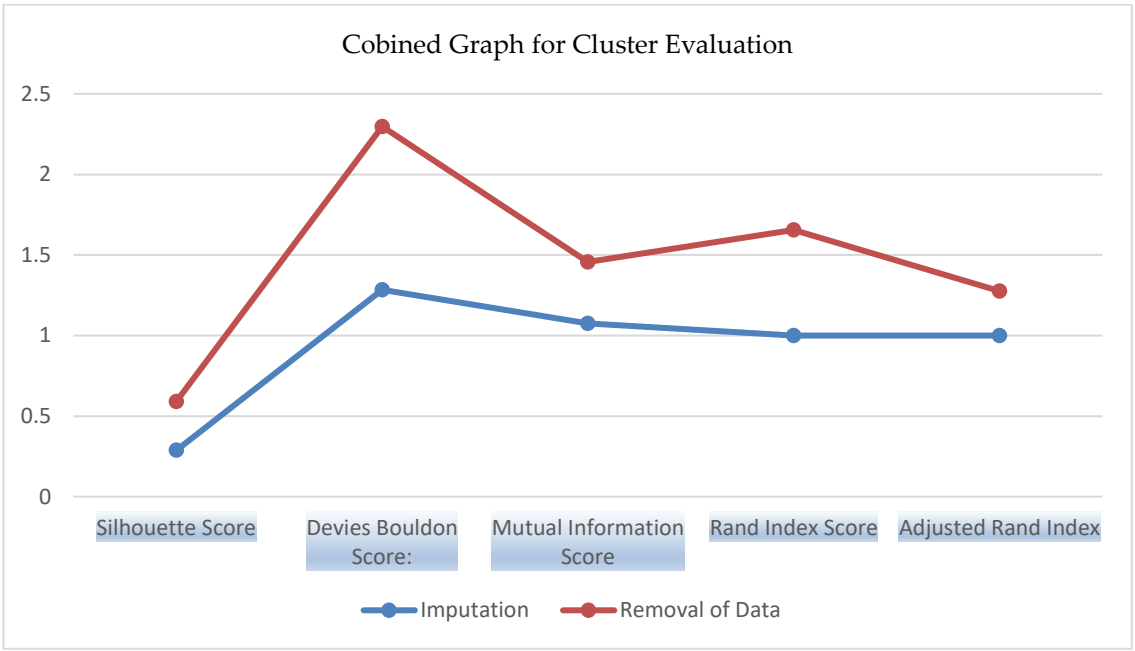

Figure 9: *Combined Graph for cluster evaluation parameters*

Figure 9 have a combined graphical representation of the evaluation criterion showing betterment for the proposed strategy A.

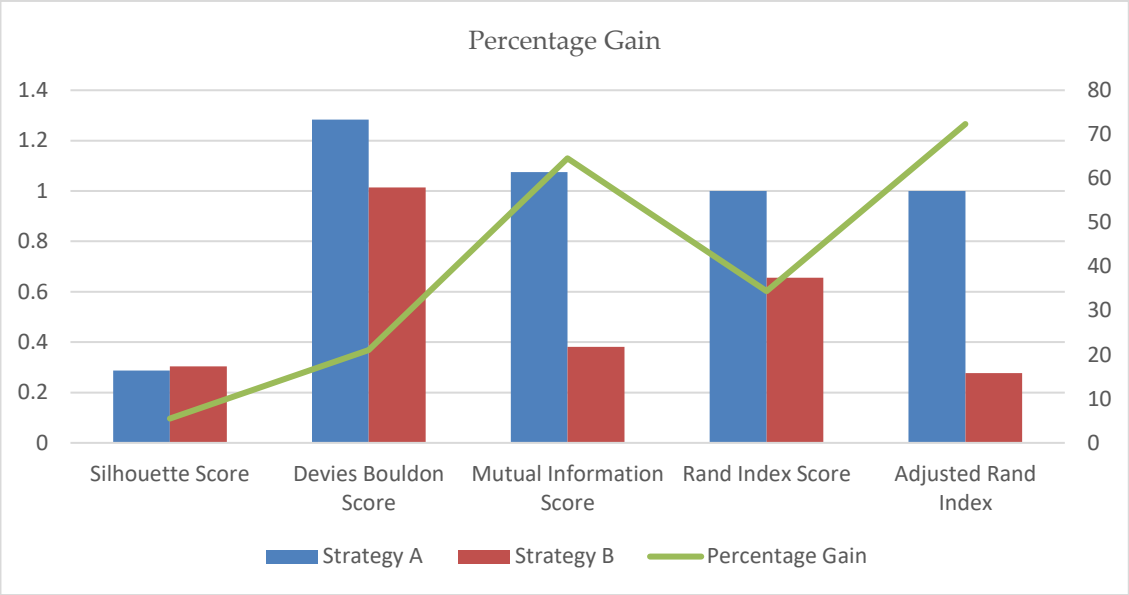

Figure 10: *Percentage Gain between the Strategy A and Strategy B*

Figure 10 reveals the percentage gain due to strategy A over the strategy B. So using the mean imputation (MI) of missing data one can find better result by using proposed algorithm (Strategy A)

9. Conclusion

This paper has presented a new algorithm named after MI-K-mean algorithm for the analysis of financial risk data. When risk factor exists then more input data values are required to reach the better decision. So, for such situation, the usual K-mean algorithm fail to form creating the efficient clusters. The proposed algorithm MI-K-mean (Strategy A) found efficient over the four cluster evaluation parameters while applying over the financial risk data. Table 10, figure 7, figure 8, figure 9 and figure 10 are supporting this fact. The MI-K-mean algorithm contain a new imputation-based approach which is unique feature. It opens bright avenues for further research while dealing with the risk data.

References

[1] Thakur, N. S., & Shukla, D. (2022). Missing data estimation based on the chaining technique in survey sampling. *Statistics in Transition new series*, 23(4):91-111.

[2] D. Shukla, N. S. Thakur, and S. Pathak. (2013). Some new aspects on imputation in sampling. *African Journal of Mathematics and Computer Science Research*, 6(1):5–15.

[3] Pandey, K.K., Shukla, D. (2022). Stratified linear systematic sampling-based clustering approach for detection of financial risk group by mining of big data. *Int J Syst Assur Eng Manag*, 13:1239–1253.

[4] Pandey, K. K., & Pradhan, N. (2014). An analytical and comparative study of various data preprocessing method in data mining. *International Journal of Emerging Technology and Advanced Engineering*, 4(10):174-180.

[5] Shukla, D., Thakur, N. S. (2008). Estimation of mean with imputation of missing data using Factor Type Estimator. *Statistics in Transition*, 9(1):33–48.

[6] Jäger, S., Allhorn, A., & Biebmann, F. (2021). A Benchmark for Data Imputation Methods. *Frontiers in Big Data*, 4:693674- 88.

[7] Ali, A., Emran, N. A., & Asmai, S. A. (2021). Missing values compensation in duplicates detection using hot deck method. *Journal of Big Data*, 8(1):1-19.

[8] Jung Wun Lee & Ofer Harel (2023). Incomplete clustering analysis via multiple imputation. *Journal of Applied Statistics*, 50(9):1962-1979.

[9] Doretti, M., Geneletti, S. & Stanghellini, E. (2018). Missing data: A unified taxonomy guided by conditional independence." *International Statistical Review*, 86(2):189–204.

[10] Singh, S. & Horn, S. (2000). Compromised imputation in survey sampling, *Metrika,* 51: 266–276.

[11] Khondoker M.R. (2018). Big data clustering, Wiley StatsRef: Statistics Reference Online. John Wiley & Sons Ltd, Chichester, 1:1–10.

[12] Cao F, Liang J, Jiang G. (2009). An initialization method for the k-means algorithm using neighborhood model, *Comput Math with Appl,* 58:474–483.

[13] Madhu, G. & Nagachandrika G. (2016). A New Paradigm for Development of Data Imputation Approach for Missing Value Estimation. *International Journal of Electrical and Computer Engineering*, *6*:3222-3228.

[14] Biessmann & Felix (2019). DataWig: Missing Value Imputation for Tables, *J. Mach. Learn. Res.* 20:1751-1756.

[15] Mahmud M.S., Huang J.Z. and Salloum S. (2020). A survey of data partitioning and sampling methods to support big data analysis, *Big Data Mining Analytic*s, 3:85–101.

[16] Zahra S., Ghazanfar M.A., Khalid A. (2015). Novel centroid selection approaches for k-means-clustering based recommender systems. Inf Sci (Ny), 320:156–189.

[17] Jain A.K. (2010). Data clustering: 50 years beyond k-means. *Pattern Recognition Lett*er, 31:651–666.

[18] Zhu Z. & Liu N. (2021). Early Warning of Financial Risk Based on K-Means Clustering. Algorithm. *Complex,* 55:16831-12.

[19] Woźnica, K., & Biecek, P. (2020). Does imputation matter? Benchmark for predictive models. ArXiv. /abs/pp. 2007.02837.

[20] I.D. Borlea, R.E. Precup, F. Dragan & A.B. Borlea (2017). Centroid update approach to K-means clustering. *Advances in Electrical and Computer Engineering*, 17(4):3–10.

[21] Aggarwal, C.C., Reddy, C.K. (eds.): Data Clustering: Algorithms and Applications (2013).

[22] Xu J., Xu B. & Zhang W. (2009). Stable initialization scheme for k-means clustering. *Wuhan Univ J Nat Sci*, 14:24–28.

[23] C. Diks, C. Hommes & J. Wang (2019). Critical slowing down as an early warning signal for financial crises?. *Empirical Economics*, 57(4):1201–1228.

[24] Fahad A, Alshatri N, Tari Z. (2014). A survey of clustering algorithms for big data: taxonomy and empirical analysis. *IEEE Trans Emerg Top Computer,* 2(3):267–279.

[25] Fränti P., Sieranoja S. (2019). How much can k-means be improved by using better initialization and repeats?. *Pattern Recognition*, 93:95– 112.

[26] Schafer, J. L. & J. W. Graham (2002). Missing data: our view of the state of the art. *Psychol Methods* 7 (2):147-177.

[27] Molenberghs G., C. Beunckens, C. Sotto, & M. G. Kenward (2008). Every missingness not at random model has a missingness at random counterpart with equal fit. *J. R. Stat. Soc. Ser. B. Stat. Methodology*, 70 (2):371–388.

[28] M. Z. Hossain, M. N. Akhtar, R. B. Ahmad and M. Rahman (2019). A dynamic K-means clustering for data mining. *Indonesian Journal of Electrical Engineering and Computer Science*, 13(2): 521–526.

[29] Mohan, K. & J. Pearl (2014b). On the testability of models with missing data. *In AISTATS*,1:643–650.

[30] Pearl, J. & K. Mohan (2013). Recoverability and testability of missing data: Introduction and summary of results, Technical Report R-417. *University of California, Los Angeles.*

[31] Aggarwal C.C., Reddy C.K. (2014), Data clustering algorithms and applications. *CRC Press, United States*,1:589–601.

[32] Bhaskaran, K., Smeeth, L., (2014), What is the difference between missing completely at random and missing at random? *International Journal of Epidemiology*, 43(4):1336–1339.

[33] UCI Repository for dataset: *https://archive.ics.uci.edu*

[34] Kaggle for dataset: *https://www.kaggle.com/datasets*

[35] Scikit Learn library: *https://scikit-learn.org*