ADV MATH SCI JOURNAL
Advances in Mathematics: Scientific Journal **10** (2021), no.1, 535–542
ISSN: 1857-8365 (printed); 1857-8438 (electronic)
https://doi.org/10.37418/amsj.10.1.53 Spec. Iss. on ICIRPS-2020

ANALYSIS THE CLUSTER PERFORMANCE OF REAL DATASET USING SPSS TOOL WITH K-MEANS APPROACH VIA PCA

Muhammad Kalamuddin Ahamad[1] and Ajay Kumar Bharti

ABSTRACT. Partitioning problems are handled by the idea of cluster and this technique which plays the essential work in mining of data from the given dataset. The K-Means cluster is well accepted theory to apply on huge datasets, but has some drawbacks. The factual dataset is taken from the repository of data used for clustering. Furthermore, as getting the outcome of this procedure is essential to resolve the limitations and quality enhanced of cluster by apply the Principal Component Analysis (PCA) on the dataset. In paper we have demonstrate the results by experimental for factual datasets with dissimilarities. We have worked to validate the experimental significant for the clusters metric and component size minimized for different dataset during the processing on SPSS tool on the basis of eigenvalues. In this research paper we also discussed the comparative analysis of distance between initial centroid of wine and disease of heart dataset at the level of cluster k=2 and k=3.

1. INTRODUCTION

The emerging new trends, technology and growth of business through the internet services, another way said enriched the huge amount of resources of

[1]*corresponding author*

2020 *Mathematics Subject Classification.* 68T09, 94A16, 91C20.

Key words and phrases. K-Means, Principal component analysis, Dimensionality,Centroid, Eigen values.

Submitted: 18.11.2020; *Accepted:* 25.12.2020; *Published:* 22.01.2021.

the dataset such as storage of databases, audio, video, graphics, and images. In addition to the data consists the own characteristics like that consistent, structured, unstructured, uncertain, mixed, enormous, self-motivated and more complexes analyze and considerate of the people. Therefore, in the research field of data mining the more important to how investigate and self-mining of implicit, unidentified and more vital knowledge it can manage the support like in administrative behaviors. Therefore, its removal and detection of knowledge from the business database is helpful better quality of clustering. In the extraction and analyzed of data from the huge dataset to apply the statistical tool with employed the concepts of artificial intelligence. The data mining research field is very supportive in e-business and its applications to require a demo of functionality for industries [1], applications in business like as promoting, marketing, advertising [2]. There are author discussed the k-means clustering approaches and also proposed the performance in [3]. In this research paper the PCA techniques utilize on numerical attributes on the database the noisy feature reduces the dimensions of problem consider as the dataset but improve the cluster quality on the basis of the distance between initial centroids. This paper is arranged as follows manner. In Section 1- Introduction, Section 2- Literature Review, Section 3- Proposed Research Methodology, in the section-4 Brief the experimental effects. In the Section -5 discuss the conclusion and future scope.

1.1. **Mathematical illustration of coefficient Matrix.**

Definition 1.1. *Consider that the set of data value X is consisting nonempty set members of attribute m, extraction of data sample n then determine the mean and normalized of all existing members of attribute respectively. Illustrate in the mathematical form as equation-1 and normalized equation follow as equation-2*

$$\mu = 1/n(\sum_{i=1}^{n} x_i) = 0,$$

$$N = 1/(n-1)(\sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij})^2) = 1,$$

where represent the value of x_{ij} is normalized at j=1,2,3,. . . ,m.

Definition 1.2. *Consider the consisting of two dimensional datasets of nonempty attribute X_{ij}, where i= 1to n, and j=1 to m, further associated the covariance*

matrix Coefficient at the m attribute is

$$\begin{bmatrix} Cov_{11} & Cov_{12} & \cdots & Cov_{1m} \\ Cov_{21} & Cov_{22} & \cdots & Cov_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ Cov_{n1} & Cov_{n2} & \cdots & Cov_{nm} \end{bmatrix},$$

where the notation Cov_{ij} is the covariance coefficient between the X_i and X_j and also it's mentioned by C.

1.2. Evaluate the Covariance Matrix and Eigenvalues. Find the characteristic value of the characteristic equation $|\lambda I–C = 0|$and find the eigenvalues λ_j. This is sorted as $\lambda_1, \lambda_2, \ldots, \lambda_m$, and further find the orthogonal eigenvector. In this research paper simulated eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ related to pth principal component$(p \leq m)$.Therefore the contribution eigenvalue rate is higher than value 1.00 to pick from the considering dataset.

2. Literature review

The cluster is creating to reveal the more in sequence of co- linearity, multi-co-linearity, and regression and correlation between the attribute of the dataset. Finding the total consequence in statistical examination of data is to reflect several points to have common characteristics. In the consequences of large dimensional dataset to bothered the mining process of information and not well shape clusters. In this research paper, we illustrate with the lower dimensionality of data by principal component analysis.

The author used PCA concept is reduced dimensions for changing the original data for mining and classifies it's by using the k-means. Find out the consequences to illustrate the more accuracy of reduced dimensionality of large dataset for analysis [4]. Worldwide cause deaths are in among the woman with breast cancer, and prediction of its disease to possibility of treatment otherwise very risky for health author discussed in [5]. PCA analysis tool the summarized of giving regular set patterns; evaluate the deviation of different variables, covariance and performance of dataset [8].

In telecom business the characteristics of huge datasets for main motive to discover such as reduced of real dataset, minimize the users clustering, analysis [6]. The statistics of dataset dimensionality has set of attribute and such variety

of data used in the research study and mining concepts are employed in this field like as telecommunication industries for helping the administrative strategy [7]. The significant concept of network is represented in a lower dimension to protect the structure of network node explained in [9]. The cluster property is explained and also discussed removal of k-1 term of covariance matrix and PCA project the higher to lower dimensional space, data placement in lower space graph and applied the clustering algorithm k-means [10].

2.1. **K-Means Algorithm.** This algorithm to apply on a considering d-Dimensional dataset, Choose k-prototype centroid at random from data point, Create the early dividing of cluster by assigning the object to the closest Centroid, and finally create the cluster[4-5].

2.2. **Principal Component Analysis (PCA).** The Principal component analysis is can only extract a linear projection of the data. Consider the consisting of data like as $X = x_1, x_2, \ldots, x_M$ are M vectors authors explained in[3,10]. PCA is described consisting of few steps as follows.

Step 1: Initially determine mean of data the given data set as

$$\mu = 1/M(\sum_{i=1}^{M} x_i).$$

Step 2: In the second step find subtract of mean from each individual data element Subtraction, therefore represent in the mathematical term as

$$\bar{x} = \sum_{i=1}^{M} (x_i \breve{} \mu).$$

Step 3: Measure the matrix of covariance C as

$$C = 1/M(\sum_{i=1}^{M} (\bar{x_i})(\bar{x_i})^T).$$

Step 4: Compute the eigenvalue and eigenvector $CX = \lambda X$, where $\lambda = \lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_n$ are eigenvalues and C is covariance matrix.

Step 5: Reduced the dimension of dataset.

Steps 6: Return the reduced dataset for clustering process

3. Proposed research methodology

The dataset like disease of heart, and wine are generally available on UCI Irving repository of machine learning archive. These factual datasets are retrieve from path of repository is mentioned as https://archive.ics.uci.edu.ml/datasets. The dataset is contained the instances 297,178, number of attribute 14, 13 and multivariate type of data characteristics of heart disease and wine respectively. The proposed algorithm are discussed as following:

3.1. **Procedure of PCA on Tools.**

Statistical analyses of various datasets are iris, wine and heart disease. This data set is large the make cluster initially not good cluster quality. The PCA is good concept the reduction of dimensionality of the dataset. Examines of a dataset constructs to reduced dimensions. The adopted procedure in this research paper as following

Step 1: Initially set the name of variables or attributes, then after filling the data values into data view filed,

Step 2: Create the structured data set in 2D,

Step 3: Go to analyze the 2D structure data set until return the eigenvalues,

i. Select the dimension reduction tool factor,

ii. Select the coefficient of component from descriptive field,

iii. Select the Extraction (Method) generate the variance matrix based on eigenvalues, eigenvalues set >1 and maximum iteration of convergence set at 25

iv. Press OK,

Step 4: Return the eigenvalues,

Step 5: Stop the procedure

3.2. **KMWPCA Algorithm.**

Step 1: Consider the arranged dataset, and after applied the reduction tools from SPSS,

Step 2: Extraction of some component by using the Step 1,

Step 3: Consider the element of variance and initial eigenvalues,

Step 4: To projecting data in lower dimensional subspace, getting reduced the dimension of real datasets,

Step 5: After reduction component of dataset to analyze K-Mean to achieve the distance between initial centroid for best clustering,

Step 6: Stop the process.

4. Discuss the Experimental Results

In research studies a concept to analyzed PCA implementation on SPSS Statistics 17.0 tools. Measure the eigenvalues of heart disease dataset by PCA and it's implemented on this given tools shown the results in table-1 and also analysis the simulating result is illustrated of same dataset in figure-1.

Table 1: Measure the eigenvalues at cluster k=3

Total Variance Explained at K=3

	Initial Eigenvalues(λ)			Extraction Sums of Squared Loadings		
Component	Total	% of Variance(σ^2)	Cumulative %	Total	% of Variance	Cumulative %
1	3.097	23.823	23.823	3.097	23.823	23.823
2	1.578	12.139	35.962	1.578	12.139	35.962
3	1.261	9.702	45.664	1.261	9.702	45.664
4	1.108	8.524	54.188	1.108	8.524	54.188
5	1.005	7.728	61.916	1.005	7.728	61.916
6	.877	6.750	68.666	-	-	-
7	.837	6.438	75.104	-	-	-
8	.752	5.784	80.888	-	-	-
9	.683	5.253	86.140	-	-	-
10	.559	4.303	90.443	-	-	-
11	.464	3.566	94.010	--	-	-
12	.417	3.210	97.219	-	-	-
13	.361	2.781	97.778ss			
14	.311	2.222	100	-	-	-

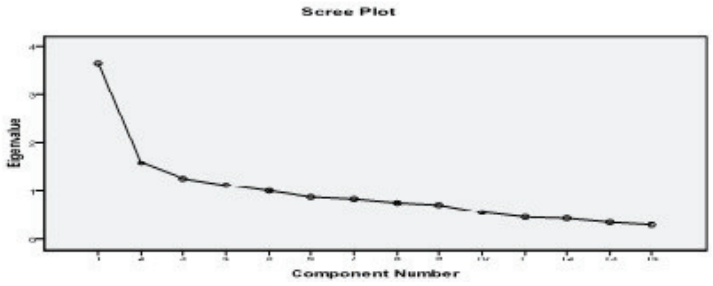
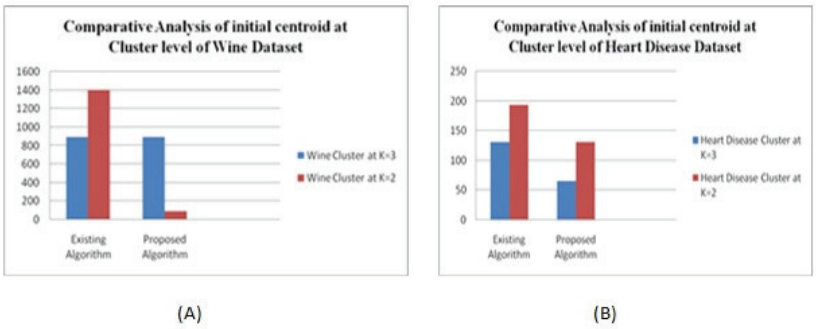


Figure 1. Analysis the extraction component from Heart Disease Dataset with eigenvalues

To measure the distance between initial centroid of given two dataset wine and heart disease respectively shown the result in below table-2. In figure-2, illustrate the comparatively analysis of existing(k-means algorithm) and proposed algorithm(k-means with PCA).

Table 2: Measure the distance between initial centroids of cluster

Datasets	Level of Cluster	Minimum distance between initial centroids	
		Existing Algorithm	Proposed Algorithm
Wine	K=3	895.845	895.844
	K=2	1402.192	92.143
Heart Disease	K=3	130.603	65.023
	K=2	193.811	131.031



(A) (B)

FIGURE 2. (A)Comparative analysis of centroids between existing and proposed algorithm(Wine dataset), (B)Comparative analysis of centroids between existing and proposed alogrithm(Heart disease dataset

In the figure-2, show the minimum distance of initial centriods at level of cluster k=3,wine, and heart disease datasets are smaller as compared than at level of cluster k=2.In this research paper, the proposed algorithm(k-means with PCA) is better than the existing algorithm(k-means) on the basis of minimum distance between the initial centroids and obtain good cluster.

5. CONCLUSION AND FUTURE SCOPE

We discuss and address in this study work lessening the component reduction and it's performance by the principal component analysis measured using SPSS statistics 17.0 tools. The dataset of heart disease are simulated on the reduction component tool. It is Lessening the dimension of the dataset by threshold value on the estimated eigenvalues. Furthermore, the simulation method is more significant and used successfully for partitioning the huge dimensionality for factual existing dataset and helpful for validating of clustering performance. In addition this paper shows the study of comparative study for existing algorithm with proposed algorithm at cluster level. If the cluster level increases then

the small distance between initial centeroid deceases and gets a well-defined cluster. This concept is to identify the appropriate causes for disease and support treatment of heart ailment, and also relevant information extract of wine dataset.

REFERENCES

[1] D.L. OLSON: *Data Mining in Business Services*, Service Business, **1**(3) (2007), 181-193.

[2] S. KUDYBA, R. HOPTROFF: *Data Mining and Business Intelligence: A Guided to Productivity*, IGI Global, 2001.

[3] C.F. TSAI, C.W. TSAI, H.C. WU, T. YANG: *ACODF: A Novel Data Custer Approach for Data Mining in Large Databases*, Journal of Systems and Software, **73**(1) (2004), 133-145.

[4] N. ZHANG, K. LEATHAM, J. XIONG, J. ZHONG: *PCA K-Means Based Clustering Algorithm for High Dimensional and Overlapping Spectra Signals*, In 2018 Ninth International Conference on Intelligent Control and Information Processing(ICICIP), (2008), 349-345.

[5] A. JAMAL, A. HANDAYANI, A.A. SEPTIANDRI, E. RIPMIATAIN, Y. EFFENDI: *Dimensionality Reduction using PCA and K-Means Clustering for the Breast Cancer Prediction*, Lontar Komputer: Jurnal Ilmiah Teknologi Informasi, (2018), 192-201.

[6] M. ALKHAVRAT, M. ALJNIDI, K. ALJOUMAA: *A Comparative Dimensionality Reduction Study in Telecom Customer Segmentation using Deep Learning and PCA*, Journal of Big Data, **7**(1) (2020), art.id.9.

[7] I.M. AL-ZUBAI, A. JAFA, K. ALJOUMAA: *Predicting Customer's Gender and Age Depending on Mobile Phone Data*, Journal of Big Data, **6**(1) (2019), art.no.18.

[8] P.R. PERES-NETO, D. JACHSON, K.M. SOMERS: *How Many Principal Components? Stopping Rules for Determining the Number for Non -trivial Axes Revisited*, Computational Statistics and Data Analysis, **49**(4) (2005), 974-997.

[9] D. WANG, P. CUI, W. ZH: *Structural Deep Network Embedding*, In Proceeding of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, (2016), 1225-1234.

[10] C. DING, X. HE: *K-Means Clustering via Principal Component Analysis*, In Proceedings of the 21st International Conference on Machine Learning, (2004), art.no.9.

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MAHARISHI UNIVERSITY OF INFORMATION TECHNOLOGY & INTEGRAL UNIVERSITY, LUCKNOW, UTTAR PRADESH, INDIA.

Email address: ahamad_kalam@rediffmail.com

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, MAHARISHI UNIVERSITY OF INFORMATION TECHNOLOGY LUCKNOW, UTTAR PRADESH, INDIA.

Email address: ajay_bharti@hotmail.com