Cluster quality analysis based on SVD, PCA-based k-means and NMF techniques: an online survey data

Hemangini Mohanty*

Centre for Data Science,
Institute of Technical Education and Research,
Siksha 'O' Anusandhan Deemed to be University,
Bhubaneswar-751030, Odisha, India
Email: hemangini.newindia@gmail.com
*Corresponding author

Santilata Champati, B.L. Padmasani Barik and Anita Panda

Department of Mathematics,
Institute of Technical Education and Research,
Siksha 'O' Anusandhan Deemed to be University,
Bhubaneswar-751030, Odisha, India
Email: santilatachampati@soa.ac.in
Email: barik.padmasani@gmail.com
Email: anitapanda@soa.ac.in

Abstract: With the increase in computerisation in every field, a huge amount of data is collected from everywhere. Therefore, extracting useful information has become a necessary task in the present era. Data mining helps to extract the information and uncover the relationship among the data. Clustering is an unsupervised technique used for partitioning objects into several groups and discover the hidden relationship among the data. There are many techniques used for clustering. In this article, a comparative study and analysis of three famous clustering techniques are done: principal component analysis (PCA), singular value decomposition (SVD) and non-negative matrix factorisation (NMF) for the clustering of a database. The database collected through a set of questionnaire surveys related to day-to-day activities. Then a comparison of their natural clustering ability is being done. Also, the use of normalised mutual information (NMI) and purity as two-cluster quality evaluation measures are explored. Then an attempt is made to show the amount of information from the original data matrix that the approximated data matrix contains. Next, to verify the accuracy of the variance covered by the approximated data matrix, the Frobenius norm is used. At last, the results are compared with the variance covered by using singular values, and a detailed analysis of each data matrix is explained.

Keywords: clustering; k-means; non-negative matrix factorisation; NMF; normalised mutual information; NMI; principal component analysis; PCA; purity; singular value decomposition; SVD.

Reference to this paper should be made as follows: Mohanty, H., Champati, S., Barik, B.L.P. and Panda, A. (2023) 'Cluster quality analysis based on SVD, PCA-based k-means and NMF techniques: an online survey data', *Int. J. Reasoning-based Intelligent Systems*, Vol. 15, No. 1, pp.86–96.

Biographical notes: Hemangini Mohanty received her MSc in Applied Mathematics and Computing from C.V Raman College of Engineering, Bhubaneswar, Odisha, India. She is currently doing her PhD at the Centee for Data Science, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. Her research interests include data mining, machine learning, artificial intelligence and fuzzy data mining.

Santilata Champati received her PhD, MPhil and MSc in Mathematics from the Berhampur University, Odisha, India. She is currently an Associate Professor at the Department of Mathematics, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. Her research interests include data mining, machine learning, application of linear algebra in data analysis and prediction of dynamic behaviour of data spaces using fractional difference analysis.

B.L. Padmasani Barik received her MPhil in Mathematics from Sambalpur University, Odisha and MSc in Mathematics from Utkal University, Odisha, India. She is currently working as a Lecturer at Nimapada Autonomous College, Odisha and pursuing her PhD at the Department of Mathematics, Institute of Technical Education and Research, Siksha 'O' Anusandhan Deemed to be University, Bhubaneswar, Odisha, India. Her research interests include rough set theory, data mining, machine learning and topological aspects in data mining.

Anita Panda received her PhD, MPhil, PGDCA and MSc in Mathematics from the Berhampur University, Odisha, India. She is currently a Professor at the Department of Mathematics, Institute of Technical Education and Research, Siksha 'O' Anusandhan (Deemed to be University), Bhubaneswar, Odisha, India. Her research interests include data mining, discrete dynamic system analysis, rough set theory, machine learning and application of mathematical analysis in big data analysis.

1 Introduction

In the present scenario, data is increasing daily on an enormous scale. Therefore, obtaining meaningful data free from noise is very much needed. To overcome this situation and getting meaningful information from a large database, many researchers choose data mining techniques. These techniques can be successfully implemented in every field where a large volume of data involved. It is a process of automatically discovering useful information from a tremendous amount of data (Pauca et al., 2006). Data mining approaches are widely used in the health sector to improve health care system, in the business sector to know the potential behaviour of customers, in e-commerce field to analyse their cross-selling, to visualise the gene expression, in education sector to improve the learning pattern and many more (Pujari, 2019; Tan et al., 2018). Depending on the problem statement, there are several processes used for data mining. Clustering is one of the most used techniques for partitioning objects into several homogenous groups (Pujari, 2019). Cluster analysis is an unsupervised learning technique, used for the discovery of uncovered relationships in the underlying data from the data distribution (Jafarzadegan et al., 2019; Tan et al., 2018). It helps to analyse the structure of datasets and to identify the behaviour of each cluster. It is used in many applications like information retrieval, to find patterns from data, decision-making, grouping customers based on their criteria, and for image segmentation (Jafarzadegan et al., 2019; Pauca et al., 2006).

During the process of clustering, to visualise the data adequately, it is important to reduce the number of dimensions. There are many dimensionality reduction techniques applied according to the requirements. Mainly three matrix-based techniques viz. PCA, SVD, and NMF are used to compute a data matrix of reduced dimensionality from a huge dataset. These three techniques have been successfully applied to big data matrices for clustering (Atif et al., 2019; Gao and Zhang, 2005; Gu et al., 2020; Hoyer, 2004; Huang et al., 2017; Jain et al., 1999; Lee and Seung, 2001; Lee and Jun, 2013; Winck et al., 2012; Zadeh et al., 2006). PCA is a statistical method used to reduce the dimensionality of a datasets and for extracting the features. The goal of PCA is to reduce the size of the dataset by extracting hidden information from the datasets and then analysing the observations that simply explain the datasets. Also, using this method error rate is lower than other dimensionality reduction methods (Jafarzadegan et al., 2019) and thus it has been selected to combine the basic clustering. SVD is another approach that allows an exact representation of any matrix by eliminating the less important parts of that matrix with any desired number of dimensions (Deng et al., 2019; Gao and Zhang, 2005; Gu et al., 2020). It is also a popular method used for clustering since it has the ability to form natural clusters of the dataset. Both PCA and SVD are depending on the eigenvalues of the data matrices, but the steps they involved are quite different from each other (Tan et al., 2018). Nowadays, NMF is widely used as a clustering technique with the excellent performance (Hoyer, 2004; Long et al., 2014; Pauca et al., 2006). The aim is to find two non-negative matrices of lower dimensions with their product that provide a good approximation to the original data matrix (Lee and Seung, 2001). It minimises the error of the objective function to get the desired level of accuracy. After going through several research works related to matrix factorisation (MF) techniques applied, a curiosity to know about which technique gives better clusters for a large database is developed. Also, there is an issue of selecting an appropriate technique that is better for clustering in a particular real-life problem. In this article, an attempt is made to collect an independent dataset through a survey. After receiving all the responses, the dataset is pre-processed before different techniques applied to the original data matrix. Then we found the clusters and evaluated the quality of the clusters to compare different techniques of clustering and select the better one. At last, to verify the amount of information contained in the approximated data matrix, Frobenius norm is used and experimentally, it is being shown that the variance covered by the approximated data matrix is equal by the two verifications, i.e., singular values and Frobenius norm. This article consists of a detailed comparison of the information about the clusters by getting from different methods.

This article contains the following: in Section 2, a literature review of related works on the three different clustering techniques is described. In Section 3, the basic concepts of the involved techniques are described. The cluster quality evaluation process is described in Section 4. In Section 5, the data collection procedure and observations obtained are noted down. Finally, Section 6 concludes the article, and provides future directions.

2 Historical background

Several research works have been done in the field of data mining by using these three MF techniques. As a large dimensional datasets are involved, an appropriate method is required to reduce the dimensions of the original dataset without losing much information. PCA is one of the most popular method used for dimensionality reduction, and followed by applying k-means algorithm for clustering, it gives better clusters. Using PCA, in data mining enhances getting a lower error rate (Jafarzadegan et al., 2019). In Jafarzadegan et al. (2019), the authors have used some methods for combining hierarchical clustering approaches. They have shown that the clustering accuracy of PCA is more than other methods. In the health sector, Zhu et al. (2019) proposed a logistic regression model by combining PCA and k-means clustering for predicting a person is diabetic or not. For high dimensional data, Lee and Jun (2013) have shown that applying PCA into Gaussian mixture model gives better accuracy, increases the efficiency of clusters and it also eliminates the noisy data and controls the decision errors.

Many researchers have used these MF techniques separately in various fields. SVD is another mostly used technique of many researchers, since it gives natural clusters. The advantage of using SVD is it reduces the numerical error and it is easy to visualise the data into a geometrical structure (Deng et al., 2019; Gao and Zhang, 2005; Gu et al., 2020). In Gu et al. (2020), the authors have proposed a latent factor model using SVD for predicting ratings of products in a recommender system. They have assigned some weights to the initial points in the model to maintain the low-rank property and their model is best to anchorage the bad influences of noise data and provides a better prediction about the noise data. In Gao and Zhang (2005), the authors applied SVD for doing several smaller clustered structures for large inhomogeneous datasets to enhance the text retrieval accuracy and reduce the computing time also the storage costs. In Deng et al. (2019), the authors have shown that the SVD based tensor decomposition technique gives a better representation of high dimensional data. Zeng et al. (2019) developed a group-based k-SVD algorithm with some non-local self-similarity properties to extract the false diagnosis features and group the features that are more focused by using their algorithm.

NMF is now adopted by many researchers, as it is quite simpler than other techniques. It is widely used in research fields like image processing, hyperspectral image analysis, signal processing, data mining and document clustering, computer vision and computational biology (Atif et al., 2019; Belachew and Buono, 2020; Huang et al., 2017; Liu et al., 2018; Pauca et al., 2006). In Belachew and Buono (2020), the authors have developed a generalised hybrid projective NMF algorithm by combining alternating least-squares of the α-divergence which increases the clustering performance. They have proposed a projective NMF method which gives the highly orthogonal and sparse basis factors that enhance the quality of the clusters. In Huang et al. (2017), the authors have shown that clustering by NMF with input similarity data matrix gives better accuracy. In genetics, it is used to identify the differentially expressed genes and to cluster the gene samples (Liu et al., 2018). Pauca et al. (2006) used this concept for the analysis of a spectral dataset and they show that NMF gives better performance analysis for sparse datasets.

Some authors (Atif et al., 2019; Winck et al., 2012) have combined both SVD and NMF to see how the results interpreted. Lee and Jun (2013) have proposed a method using low-rank correlation SVD-based NMF to reduce the initialisation error for better clustering. In this article, the authors have shown that using SVD, before applying NMF to the data matrix decreases the initial error rate and sparsity of initial factors is approximated half of the original factors. Also, the computational process is cheaper by using their proposed method. Another research work has also been done using both the methods to create an intermediate surface to improve the surface generation time in a pin array device in Winck et al. (2012). Winck et al. (2012) have successfully implemented these two methods in command generation technique which eliminate the oscillations and improves the performance.

3 Basis concepts

3.1 SVD-based clustering

SVD is the most widely used MF technique used in big data matrices. The purpose of this method is to transform a large dataset to a lower dataset, which contains a large fraction of the information present in the original dataset (Tan et al., 2018). The key element of this method, it reduces the data matrix to an approximated data matrix, according to the rank of the original dataset. It decomposes any m by n matrix D into the product of three matrices viz.

$$D = U\Sigma V^T \tag{1}$$

where the columns of U (m by n) are normalised eigenvectors of DD^T, the columns of V (n by n) are normalised eigenvectors of D^TD (Strang, 2006). The r singular values on the diagonal of Σ (m by n) are the square roots of the non-zero eigenvalues of both DD^T and D^TD (Strang, 2006).

3.2 PCA followed by k-means-based clustering

For high-dimensional data, finding different relationships among the attributes and to analyse them is difficult. To reduce the dimensionality of a large dataset, PCA is the mostly used method for it. Since, it gives the new set of dimensions (attributes) which captures as much as the variability of the data (Pujari, 2019; Tan et al., 2018) therefore, it is easy to discover the hidden relationships, and uncovering the outliers within the newly defined lower dimensions.

Here, our objective is to transform the dataset *D* of *p* dimension into a new sample set *Y* of lower dimension l with (l < *p*), where *Y* is the matrix of first few principal components (PCs) of *D*. We proceed as follows:

1 Standardised the dataset by finding the mean of the data using

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} \tag{2}$$

and subtracting the mean from each data points.

2 Computed the covariance matrix '*S*' with $S_{i,j}$ = *Covariance*(d_{*i}, d_{*j}), where $S_{i,j}$ is the covariance of the i^{th} and j^{th} attributes of the data.

3 Calculated eigenvalues and eigenvectors of the covariance matrix *S*. The eigenvector associated with the largest eigenvalue gives the direction of the data, which captures the most of the data variance.

4 Then, selected the eigenvectors of the larger eigenvalues so to eliminate the less significant data points and consider the PCs which gives a good approximation of the original data.

Next, we applied k-means algorithm for clustering the PCA decomposed data matrix.

3.2.1 k-means-based clustering

k-means is a prototype and partitioning-based technique for clustering, that is used to find a user-specified number of clusters, which are separated by their centroids (Tan et al., 2018). This method tries to make each data points belong to only one cluster, such that the centroids of each cluster are at minimum distance from the data points. Here, we set *k* = 3 since there are three eigenvalues of high significance. The steps of k-means-based clustering are described as follows:

1 first select the number of clusters k which you need

2 initialise the centroids and form k clusters by assigning each k data points to its closest centroids

3 recompute the centroid of each clusters and again assign data points to its closest centroid

4 continue the process, until there is no change in the cluster data points.

3.3 NMF-based clustering

It is a linear method used for dimensionality reduction, which is helpful in extracting hidden and intrinsic features of high dimensional datasets (Hoyer, 2004; Pauca et al., 2006). NMF factorises a given data matrix $D = [x_1, x_2, \ldots, x_n] \in R_+^{m\times n}$, where *n* are the data points and m are the data dimensions. The goal of NMF is to reduce the rank '*k*', into two low rank matrices $W \in R_+^{m\times k}$ and $\in R_+^{k\times n}$, that approximates *D* into a lower dimensionality form as:

$$D = WH \tag{3}$$

Which minimises the objective function as follows:

$$J_{NMF} = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(X_{i,j} - \left(WH_{i,j}^{T} \right) \right)^2 = \| D - WH^T \|_F^2 \quad \text{s.t. } W \geq 0, H \geq 0 \tag{4}$$

where $\| \,.\, \|_F$ is Frobenius norm and $X_{i,j}$ denotes the (*i*, *j*) element of *D* (Huang et al., 2017; Lee and Seung, 2001; Pauca et al., 2006).

4 Cluster quality evaluation

In this section, we have used two cluster quality evaluation measures, i.e., NMI and purity, to evaluate the quality of clusters formed by all the three techniques, since these two measures are external evaluation measures. We have classified the users into three classes based on their qualification. Here, class 1 consists of the post graduate students, class 2 consists of the teaching staff and class 3 consists of the undergraduate students.

4.1 Normalised mutual information

To make the score fit to the scale of [0, 1], normalisation of the mutual information is done. For calculating NMI, the following formula is used.

$$NMI(Y, C) = \frac{2 \times I(Y, C)}{H(Y) + H(C)} \tag{5}$$

where *C* is the cluster labels, *Y* is the class labels, *H*(*C*) = entropy of *C*, *H*(*Y*) = entropy of *Y*, and *I*(*Y*, *C*) = mutual information between *Y* and *C*.

4.2 Purity

A measure of a cluster is considered to be pure, if it contains all the labelled objects of only one single class.

$$Purity(C, Y) = \frac{1}{n} \sum_{c \in C} \max_{y \in Y} |c \cap y| \tag{6}$$

Here *c* represents a cluster and *y* represents a class.

5 Procedure

In the present work, an attempt is made to study the following:

1 clustering quality of SVD, PCA-based k-means, and NMF-based clustering techniques

2 how to create a database on some real-life issues

3 how to interpret the observations, so that decision making is facilitated.

To do so, the following steps are taken.

To collect an original dataset, first, we have prepared a set of ten questions having each question with five options. All these questions are related to day-to-day life problems. Then, we created a link for the questionnaire survey, so that it is convenient for giving the responses and all the responses were collected in one database. A total of 160 persons took interest in giving their choices. Next, we organise the data matrix by considering the frequency of 160 × 50 to 16 × 50 matrix by selecting the frequency of a data attribute for ten users. Then using MATLAB, we separately applied the above techniques to each question database and obtained the clusters. After getting the clusters, we have calculated NMI and purity.

Then, the observations studied to get some idea to decide or answer the above three questions presented here. While preparing the questionnaire, the following questions are included.

The questions are as follows.

1 You like to buy your clothes from

a a local store

b shopping mall

c online

d a wholesaler

e any of the above.

2 If you ever get a chance to buy clothes online, which of the following sites you would prefer?

a Flipkart

b Amazon

c Myntra

d Zabong

e Ajio.

3 You watch movies online using

a YouTube

b Hot star

c SonyLiv

d Amazon Prime

e Netflix.

4 How do you like to watch a movie?

a in cinema hall

b in TV at home

c online

d only during weekends using any of a, b, c

e do not like movies.

5 Which food you like the most?

a North Indian

b Tandoori and Mughlai

c Continental

d South Indian

e Chinese.

6 Which food type you like the most?

a Indian sweets

b chocolates

c ice cream

d biriyani

e chats and panipuri.

7 Which of the following Bollywood star is your favourite?

a Amitabh Bachhan

b Akshya Kumar

c Vicky Kaushal

d Rajkumar Rao

e others.

8 You won’t miss a movie of which of the following stars?

a Deepika Padukone

b Alia Bhatt

c Tapsee Punnu

d Kangana Ranaut

e Anuska Sharma.

9 If given a chance which of the following Asian countries you would like to visit?

a Thailand

b Nepal

c Sri Lanka

d China

e Saudi Arabia.

10 Who is your favourite sports star?

a Virat Kohli

b MS Dhoni

c Rohit Sharma

d Jasprit Bumrah

e Sikhar Dhawan

The following matrices are the matrix of each question wise data collected. Corresponding to each question, an independent data matrix is prepared and studied. For future study, we present the data matrices of all ten questions here:

$$\begin{pmatrix} 1&5&1&1&2\\ 0&4&3&0&3\\ 1&4&3&0&2\\ 1&1&3&1&4\\ 2&5&1&1&1\\ 2&3&1&0&4\\ 2&3&2&1&2\\ 0&5&1&0&4\\ 2&5&1&0&2\\ 0&5&1&1&3\\ 1&3&1&0&5\\ 1&4&2&0&3\\ 0&4&3&0&3\\ 0&3&0&2&5\\ 3&2&1&1&3\\ 2&5&0&0&3 \end{pmatrix} \begin{pmatrix} 2&2&6&0&0\\ 3&2&5&0&0\\ 3&4&3&0&0\\ 2&1&6&0&1\\ 4&3&3&0&0\\ 3&3&4&0&0\\ 1&4&4&0&1\\ 4&0&4&0&2\\ 3&1&6&0&0\\ 4&2&4&0&0\\ 3&1&6&0&0\\ 2&2&6&0&0\\ 4&2&4&0&0\\ 1&3&6&0&0\\ 3&1&5&0&1\\ 5&1&4&0&0 \end{pmatrix} \begin{pmatrix} 7&1&0&1&1\\ 7&1&0&2&0\\ 5&2&0&0&3\\ 3&3&0&1&3\\ 2&3&1&2&2\\ 5&0&0&3&2\\ 4&3&1&1&1\\ 4&3&0&3&0\\ 4&3&0&1&2\\ 7&0&0&0&3\\ 7&0&0&0&3\\ 4&1&0&2&3\\ 4&2&0&2&2\\ 2&3&0&3&2\\ 7&1&0&1&1\\ 3&6&0&0&1 \end{pmatrix} \begin{pmatrix} 3&2&3&2&0\\ 3&1&4&2&0\\ 6&1&0&3&0\\ 5&1&1&3&0\\ 3&3&1&3&0\\ 4&1&1&4&0\\ 4&4&0&2&0\\ 4&1&1&3&1\\ 4&1&2&3&0\\ 5&1&1&2&1\\ 1&2&3&4&0\\ 3&2&1&4&0\\ 5&2&1&1&1\\ 7&1&1&0&1\\ 2&1&2&5&0\\ 6&0&0&3&1 \end{pmatrix}$$

$$\begin{pmatrix} 2&4&0&2&2\\ 2&3&1&1&3\\ 0&3&2&4&1\\ 2&2&0&5&1\\ 2&4&0&2&2\\ 0&4&0&4&2\\ 2&3&0&3&2\\ 2&3&0&3&2\\ 1&5&1&2&1\\ 3&4&2&1&0\\ 3&3&0&2&2\\ 2&4&0&2&2\\ 3&3&0&0&4\\ 4&3&0&2&1\\ 2&3&1&1&2\\ 3&3&0&2&2 \end{pmatrix} \begin{pmatrix} 2&0&5&1&2\\ 1&0&0&7&2\\ 3&0&3&3&1\\ 1&1&0&3&5\\ 3&1&1&4&1\\ 3&0&1&2&4\\ 2&0&1&3&4\\ 4&1&0&3&2\\ 3&2&1&3&1\\ 3&0&2&4&1\\ 1&1&1&3&4\\ 0&0&0&8&2\\ 2&2&0&6&0\\ 1&0&0&6&3\\ 5&0&1&3&1\\ 3&0&1&3&3 \end{pmatrix} \begin{pmatrix} 2&4&3&1&0\\ 3&1&1&1&4\\ 0&1&2&1&6\\ 1&2&2&1&4\\ 2&4&1&0&3\\ 2&3&0&0&5\\ 3&3&1&0&3\\ 1&0&2&0&7\\ 2&2&1&0&5\\ 2&3&3&1&1\\ 1&1&1&2&5\\ 0&3&1&2&4\\ 2&0&1&0&7\\ 3&2&1&1&3\\ 2&3&3&0&2\\ 1&3&2&0&4 \end{pmatrix} \begin{pmatrix} 6&0&0&2&2\\ 3&3&1&2&1\\ 1&2&2&1&4\\ 7&0&1&1&1\\ 4&3&1&2&0\\ 4&4&0&1&1\\ 3&4&0&0&3\\ 4&1&3&1&1\\ 5&2&1&0&2\\ 4&2&2&0&2\\ 5&2&2&0&1\\ 6&2&1&1&0\\ 3&3&2&1&1\\ 5&2&2&0&1\\ 7&1&1&0&1\\ 4&3&0&1&2 \end{pmatrix}$$

$$\begin{pmatrix} 4&0&2&0&4\\ 7&1&1&1&0\\ 7&0&0&0&3\\ 6&2&1&0&1\\ 6&2&0&0&2\\ 5&1&0&2&2\\ 7&0&0&2&1\\ 5&1&1&1&2\\ 6&2&0&1&1\\ 7&1&0&1&1\\ 5&4&0&1&0\\ 7&2&0&0&1\\ 4&1&1&1&3\\ 8&1&0&0&1\\ 4&3&0&2&1\\ 4&1&0&2&3 \end{pmatrix} \begin{pmatrix} 4&6&0&0&0\\ 4&4&2&0&0\\ 4&5&0&0&1\\ 1&7&2&0&0\\ 1&9&0&0&0\\ 2&8&0&0&0\\ 5&4&0&1&0\\ 2&7&1&0&0\\ 0&7&3&0&0\\ 1&8&1&0&0\\ 1&6&1&1&1\\ 1&8&1&0&0\\ 2&7&0&0&1\\ 2&7&1&0&0\\ 1&9&0&0&0\\ 3&6&1&0&0 \end{pmatrix}$$

5.1 *Observation*

While analysing the results of each method for ten questions one by one, we observed that clustering by SVD in every question, it gives only one cluster of all the users. From this, one can say that all the users have the same kind of choices. However, it is not the reality. Due to huge differences between the largest and second-largest singular values, all the users correspond to one cluster.

For a given data matrix D, the SVD of $D = U\Sigma V^T$ is obtained. The data matrix D is approximated to,

$$\tilde{D} = \sum_{i=1}^{k} \sigma_i u_i v_i^T \tag{7}$$

Here, k is chosen in such a way that the desired level of similarity is achieved, where

$$similarity = \frac{\|\tilde{D}\|_F^2}{\|D\|_F^2} \approx \frac{\sum_{i=1}^{k} |\sigma_i|^2}{\sum_{i=1}^{n} |\sigma_i|^2} \tag{8}$$

The Frobenius norm is used for calculating square of the norm of the original matrix D and approximated matrix $\tilde{D}$. Then to get a better approximation of the given data matrix in the projected space, the formula $\tilde{D} = \sum_{i=1}^{k} \sigma_i u_i v_i^T$ is used for different values of k till the desired accuracy is achieved.

To know the role of singular value associated in SVD process, consider the data matrix D. As DD^T and D^TD are the symmetric matrices, their corresponding normalised eigenvectors form orthonormal sets. Therefore, we can write $D = U\Sigma V^T \Rightarrow DV = U\Sigma$, where U and V are orthogonal matrices having normalised eigenvectors of DD^T and D^TD as its columns. Hence for any column v_i of V, there exists u_i of U and σ_i of Σ such that

$$Dv_i = \sigma_i u_i$$
$$\Rightarrow \|Dv_i\| = \|\sigma_i u_i\|$$
$$\Rightarrow \|Dv_i\| = |\sigma_i|\,\|u_i\|$$
$$\Rightarrow \|Dv_i\| = |\sigma_i| \text{ as } \|u_i\| = 1.$$

From the above expression, it is observed that the stretchability of the vector v_i by the transform D is $|\sigma_i|$. Also, it is being found that $\sigma_i u_i v_i^T$ represents an approximation of the matrix D. The variance of the data matrix covered by the approximated matrix is given by $\frac{\sigma_i^2}{\sum_{i=1}^{n} \sigma_i^2}$. Here, this result is illustrated with the following verification.

For the first data matrix,

$$\sum_{i=1}^{5} \sigma_i^2 = 519.997, \frac{\sigma_1^2}{\sum_{i=1}^{5} \sigma_i^2} = 0.8611,$$

$$\frac{\sum_{i=1}^{2} \sigma_i^2}{\sum_{i=1}^{5} \sigma_i^2} = 0.92, \frac{\sum_{i=1}^{3} \sigma_i^2}{\sum_{i=1}^{5} \sigma_i^2} = 0.95764.$$

Together the three singular values contain 95.77% information of the original data matrix. By Frobenius norm, the following calculation can be done.

$$\|A\|_F^2 = 519.999, \frac{\|\sigma_1 u_1 v_1^T\|_F^2}{\|A\|_F^2} = 0.86123$$

$$\frac{\left\|\sum_{i=1}^{2} \sigma_i u_i v_i^T\right\|_F^2}{\|A\|_F^2} = 0.92016, \frac{\left\|\sum_{i=1}^{3} \sigma_i u_i v_i^T\right\|_F^2}{\|A\|_F^2} = 0.9578.$$

Also, the approximated matrix has 86.12% similarity of that data matrix, where similarity is $\frac{\|\tilde{D}\|_F^2}{\|D\|_F^2}$. Similarity, it is verified for remaining all the nine matrices.

For each data matrix, user clustering and attribute clustering are formed using different clustering techniques and values of NMI and Purity of clusters are calculated. To simplify the observation and get better and quicker conclusions, this information is represented on a tabular form given in Table 1.

Table 1 shows NMI and Purity comparison of all the three methods.

Table 1 NMI and purity of clusters

Question no.	*SVD*	*PCA-based k-means*	*NMF*

1	C_1 = {1, 2, …, 16}
	NMI = 0.0014
	Purity = 100%
		C_1 = {2, 3, 7, 12, 13}
		C_2 = {1, 5, 8, 9, 10, 16}
		C_3 = {4, 6, 11, 14, 15}
		NMI = 0.093
		Purity = 62.5%
			C_1 = {1, 5, 7, 8, 9, 10, 12, 16}
			C_2 = {6, 11, 14, 15}
			C_3 = {2, 3, 4, 13}
			NMI = 0.218
			Purity = 68.75%

2	C_1 = {1, 2, …, 16}
	NMI = 0.0014
	Purity = 100%
		C_1 = {2, 4, 9, 11, 12, 14, 15}
		C_2 = {3, 5, 6, 7}
		C_3 = {8, 10, 13, 16}
		NMI = 0.446
		Purity = 68.75%
			C_1 = {1, 2, 4, 9, 11, 12, 14, 15}
			C_2 = {8, 10, 13, 16}
			C_3 = {3, 5, 6, 7}
			NMI = 0.446
			Purity = 68.75%

3	C_1 = {1, 2, …, 16}
	NMI = 0.0014
	Purity = 100%
		C_1 = {16}
		C_2 = {1, 2, 6, 10, 10, 11, 15}
		C_3 = {3, 4, 5, 7, 8, 9, 12, 13, 14}
		NMI = 0.049
		Purity = 68.75%
			C_1 = {1, 3, 9, 10, 11, 12, 15}
			C_2 = {2, 6, 8, 13, 14}
			C_3 = {4, 5, 7, 16}
			NMI = 0.128
			Purity = 62.5%

4	C_1 = {1, 2, …, 16}
	NMI = 0.0014
	Purity = 100%
		C_1 = {3, 4, 8, 10, 13, 14, 16}
		C_2 = {1, 2, 6, 9, 11, 12, 15}
		C_3 = {5, 7}
		NMI = 0.188
		Purity = 62.5%
			C_1 = {3, 4, …, 14, 16}
			C_2 = {11, 15}
			C_3 = {1, 2}
			NMI = 0.27
			Purity = 68.75%

Table 1 NMI and purity of clusters (continued)

Question no.	*SVD*	*PCA-based k-means*	*NMF*
5	C_1 = {1, 2, …, 16}	C_1 = {2, 13}	C_1 = {1, 2, 5, 7, 8, 9, 12, 13, 15}
	NMI = 0.0014	C_2 = {3, 4, 6}	C_2 = {3, 4, 6}
	Purity = 100%	C_3 = {1, 5, 7, 8, 9, 10, 11, 12, 14, 15, 16}	C_3 = {10, 11, 14, 16}
		NMI = 0.247	NMI = 0.312
		Purity = 68.75%	Purity = 68.75%
6	C_1 = {1, 2, …, 16}	C_1 = {4, 6, 7, 11, 16}	C_1 = {1, 2, 5, 10, 12, 13, 14}
	NMI = 0.0014	C_2 = {1, 2, 12, 13, 14}	C_2 = {4, 6, 7, 11}
	Purity = 100%	C_3 = {3, 5, 8, 9, 10, 15}	C_3 = {3, 8, 9, 15, 16}
		NMI = 0.166	NMI = 0.231
		Purity = 56.25%	Purity = 62.5%
7	C_1 = {1, 2, …, 16}	C_1 = {2, 3, 7, 12, 13}	C_1 = {2, 6, 7, 9, 13, 14}
	NMI = 0.0014	C_2 = {1, 5, 8, 9, 10, 16}	C_2 = {3, 4, 8, 11, 12, 16}
	Purity = 100%	C_3 = {4, 6, 11, 14, 15}	C_3 = {1, 5, 10, 15}
		NMI = 0.093	NMI = 0.154
		Purity = 62.5%	Purity = 56.25%
8	C_1 = {1, 2, …, 16}	C_1 = {2, 3, 5, 6, 7, 13, 16}	C_1 = {1, 4, 8, 9, 10, 11, 12, 14, 15}
	NMI = 0.0014	C_2 = {1, 4, 12, 15}	C_2 = {2, 5, 6, 7, 13, 16}
	Purity = 100%	C_3 = {8, 9, 10, 11, 14}	C_3 = {3}
		NMI = 0.3	NMI = 0.263
		Purity = 62.5%	Purity = 62.5%
9	C_1 = {1, 2, …, 16}	C_1 = {2, 3, 7, 10, 12, 14}	C_1 = {2, 3, …,, 12, 14}
	NMI = 0.0014	C_2 = {1, 6, 8, 13, 16}	C_2 = {1, 13, 16}
	Purity = 100%	C_3 = {4, 5, 9, 11, 15}	C_3 = {15}
		NMI = 0.099	NMI = 0.085
		Purity = 56.25%	Purity = 56.25%
10	C_1 = {1, 2, …, 16}	C_1 = {1, 2, 3, 7, 16}	C_1 = {4, 5, 6, 8, 9, …, 15}
	NMI = 0.0014	C_2 = {4, 9, 11}	C_2 = {1, 2, 3, 7, 16}
	Purity = 100%	C_3 = {5, 6, 8, 10, 12, 13, 14, 15}	NMI = 0.164
		NMI = 0.254	Purity = 68.7.5%
		Purity = 62.5%	

It is observed that SVD-based clustering is natural clustering, but in case of PCA and NMF, one has to mention the number of clusters before clustering, which is a limitation of the methods. Here in the present case, we have selected $k = 3$ as three number of significant eigenvalues are found in all the questions and most of the cluster members are same in both which can be observed from Table 1.

In this work, we have applied SVD, PCA-based k-means and NMF-based clustering techniques to the dataset. From the above clustering techniques, we found that SVD is better from other clustering techniques. It has some special features which are not present in other two clustering techniques. By applying this method, one can get the clusters and from the singular value, one can get the information about the significance of the clusters. However, in case of k-means and NMF-based clustering, the significance of the cluster cannot be visualised easily. As the singular value of the associated cluster inform regarding the significance of a cluster, it facilitates better decision making by using SVD-based clustering. On the other hand, one can observe the features of attributes and the behaviour of the users. We can know the similarities and difference between the attributes and users. Hence, we can observe and get more hidden information from the attribute clusters as well as from the user clusters.

Then, the calculation of the NMI and purity of clusters of each data matrix is done. It is observed that in SVD-based clustering, all user rows correspond to one cluster. Because of which the purity is found to be 100% with normalised mutual information 0.0014. The NMI of clusters by PCA-based k-means clustering and NMF-based clustering are 0.1936 and 0.2271, respectively. Also, the purities of PCA and NMF-based clustering are 62.5% and 64.375%, respectively. From this, it is being observed that the clusters by PCA-based k-means clustering have sharing minimum MI with each other. Hence, it is not better than the NMF-based clustering according to NMI point of view.

Similarly, the purity of clusters of NMF-based clustering is slightly higher than PCA-based k-means clustering.

For the present data matrix, SVD-based clustering is found to be better, followed by NMF-based clustering and PCA-based k-means clustering are the least.

5.2 Analysis

In this present work, our main focus is to know which clustering method gives better clusters, how to collect data and interpret the observation to get useful information for enhancing better decision-making capacity. From the observation, it is found that SVD-based clustering gives better clusters and also reveals the hidden relationships in the dataset. The other two clustering techniques give the clusters of the dataset. They don't give any information about the users or attributes. An effort is made to collect the dataset and prepare the data before applying the clustering techniques. After getting the clusters of each dataset, analysis is done on the basis of the information about the clusters. These analyses help to improve decision making capacity.

There are three clusters found by the other two clustering methods. In first data matrix, the NMI of the clusters of PCA-based k-means clustering and NMF-based clustering are 0.093 and 0.218, respectively. The purities of the clusters by PCA and NMF are 62.5% and 68.75%, respectively. Therefore, for the first data matrix NMF-based clustering is better than PCA-based k-means clustering since it has high NMI and purity. Again, for the second data matrix, both the methods give same cluster member in each cluster and the NMI and purity are 0.446 and 68.75%, respectively. Similarly, by comparing all the data matrices, we can observe that these two types of clustering techniques are almost same. Hence, further study is needed to concluded which technique is better.

By considering these two cluster quality measures, SVD-based clustering is found to be better clustering method for the present data matrix. Here, the analysis of each question is explained below according to SVD-based clustering.

1 While analysing the first question, it is found that approximately 38% of user prefer to buy clothes from a shopping mall. However, from the clusters formed by SVD, it was seen that no user row corresponds to it, as the concept strength of it is 5.5361 which is very small in comparison to 21.1611 of the concepts of cluster-1. Hence all users in this data matrix prefer to buy clothes from a local store, online, and a wholesaler corresponds to cluster 1. However, in reality, people buy clothes from a local store is very less in number even though it has a strong base in the market. On the other hand, only 15% customers prefer to buy from online still it is accepted as its strength is high. Then there are 30% of person who have no particular choice from where they buy clothes with a low concept strength, i.e., 4.4243 Hence to dominate the market these fluctuating customers are to be targeted

2 In the second question, the three e-commerce sites Flipkart, Amazon, and Zabong belong to cluster 1 and have dominated the market as their concept strength is 23.8429 which is the highest strength. The attribute Jabong is associated with the first concept covering 92% information. Hence it is not rejected even if it is not the preference of any one of the users. The attribute Myntra is associated with a concept of less strength 5.3364. Hence it is not accepted even if 48% of users preferred it. The attribute Ajio which is associated with a low strength concept and preferred by only 3% users are not accepted.

3 In this question, almost 47% of users like watching movies online from YouTube. This percentage is comparatively higher than all other platforms. Together with YouTube, Hot star and Amazon Prime corresponds to cluster 1 which has 22.6853 strength of this concept. However, the user's choice of watching movies using Amazon Prime and Hot Star is 20% and 13.35%, respectively. The attributes Netflix and SonyLiv opted by 18% and 1.25% with a concept strength 8.0058 and 4.6969 are not associated with any of the user rows.

5 In the fourth question, the person who like watching movies in the cinema hall, in the TV at home and the person who does not likes movies are preferred by 40%, 15% and 3.125% respectively with concept strength 21.4197. However, all the users correspond to it as the concept strength is dominating. On the other hand, the person who likes watching movies only during weekends and from online with concept strength 7.3467 and 4.3484 preferred by 28% and 16% of user are not associated with any user row.

5 In this question, North Indian, South Indian, Tandoori and Mughlai, and continental food correspond to cluster 1 which is associated with the high concept strength 19.8151. Therefore, even if only 4.375% person like continental food it is accepted since the concept strength associated with it contains higher information. The attribute Chinese food is preferred by 18% of the user associated with a lower concept strength 5.8381 so that it is not considered.

6 In the sixth question, Indian sweets, chocolate, ice cream and chats, and panipuri have corresponded to cluster 1 which is associated with the high concept strength 21.5296. Therefore, chocolates and ice-cream are also accepted even if only 13% of the total users preferred to them. On the other hand, the attribute Biriyani is preferred by approximately 41% user which is majority, but associating with a low concept strength, i.e., only 7.3571 it corresponds to no user row.

7 In the case of seven, all the attributes form a single cluster which may lead to any-one of the following conclusions. All attributes are similar or the viewers are confused.

8 In this case, the viewers of Bollywood actress Deepika Padukone and Alia Bhatt belong to cluster 1 which has the concept strength 21.5636 and approximately 45% and 21.25% of user like both of them respectively. Almost 34% user like Tapsee Punnu, Kangana Ranaut, and Anuska Sharma whose corresponding concept strength is 6.6353, 4.2617 and 3.7027, respectively. All these three attributes are not accepted as their corresponding concept strength is low.

9 In this question, all the user rows correspond to Thailand and Sri Lanka which are belong to cluster 1 since their concept strength is highest, i.e., 25.1227. Although only 3.75% user likes to visit Sri Lanka it is accepted as its first component contains maximum information. The other attributes have lower concept strength so they are not considered.

10 At last, it is the analysis of favourite sports star. While analysing the responses it is found that around 68% of persons like MS Dhoni and 21% like Virat Kohli. After applying SVD to the database, it can be seen that the fans of MS Dhoni, Virat Kohli, and Jasprit Bumrah associated with cluster 1 which has the highest concept strength of 28.8350. Therefore, in this case, no user likes Jasprit Bumrah but due to the high concept strength, he has a fan following. Again, Rohit Sharma and Sikhar Dhavan liked by 10% user is rejected as their concept strength is relatively lower, i.e., 6.8779 and 3.6330.

6 Conclusions

For a real-life problem, depending on the data type, there are many clustering techniques available. In the present work, an experimental study of the three clustering techniques is done. To conclude which clustering method is best for clustering, a comparison is made by evaluating NMI and Purity as two cluster quality measures. As SVD-based clustering gives natural clustering informing about the significance and hidden relations among the elements of a cluster. Hence, it helps in better decision making. The other two methods are seemed to be equally likely in this present data matrix, since their NMI and Purity are almost same in every data matrix.

In this work, a dataset is formed based on the most suitable choice. The selected option is considered as ‘1’ and all other options are considered as ‘0’, which is not the reality. Every option may have some level of preference. In this present dataset, all the data points form a single cluster. Due to the high difference in the largest and second-largest singular values, all user rows correspond to largest singular value. Hence, rating matrix may provide more scope to express the priorities of a user.

In accordance with some other researchers (Vinh et al., 2010), in the present work, it is considered that a cluster having more Purity and NMI value is a good cluster. However, this may not be true in every situation that is in some situation less MI may give better clustering. Hence, further work can be done to study on the role of NMI and Purity, while evaluating the cluster quality. Next, we will explore some other external quality measures to finalise better clustering techniques. Using different aspects of recent developments in Mathematical research a detailed discussion on the functioning of a technique can be done. At last, a humble effort is made to understand the basic issues that a beginner needs to address before working on real-life problems.

References

Atif, S.M., Qazi, S. and Gillis, N. (2019) ‘Improved SVD-based initialization for nonnegative matrix factorization using low-rank correction’, *Pattern Recognition Letters*, Vol. 122, pp.53–59, ISSN 0167-8655, DOI: 10.1016/j.patrec.2019.02.018.

Belachew, M.T. and Buono, N.D. (2020) ‘Hybrid projective nonnegative matrix factorization based on divergence and the alternating least squares algorithm’, *Applied Mathematics and Computation*, Vol. 369, p.124825, ISSN 0096-3003, DOI: 10.1016/j.amc.2019.124825.

Deng, C., Yin, M., Liu, X.Y., Wang, X. and Yuan, B. (2019) ‘High-performance hardware architecture for tensor singular value decomposition’, *2019 IEEE/ACM International Conference on Computer-Aided Design (ICCAD)*, IEEE, pp.1–6, DOI: 10.1109/ICCAD45719.2019.8942082.

Gao, J. and Zhang, J. (2005) ‘Clustered SVD strategies in latent semantic indexing’, *Information Processing and Management*, Vol. 41, pp.1051–1063, ISSN; 0306-4573, DOI: 10.1016/j.ipm.2004.10.005.

Gu, Y., Yang, X. and Peng, M. (2020) ‘Robust weighted SVD-type latent factor models for rating prediction’, *Expert Systems with Applications*, Vol. 141, p.12885, ISSN: 0957-4174, DOI: 10.1016/j.eswa.2019.112885.

Hoyer, P.O. (2004) ‘Non-negative matrix factorization with sparseness constraints’, *The Journal of Machine Learning Research*, 12 January, Vol. 5, pp.1457–1469.

Huang, S., Xu, Z. and Wang, F. (2017) ‘Nonnegative matrix factorization with adaptive neighbors’, *2017 International Joint Conference on Neural Networks (IJCNN)*, IEEE, pp.486–493, DOI: 10.1109/IJCNN.2017.7965893.

Jafarzadegan, M., Safi-Esfahani, F. and Beheshti, Z. (2019) ‘Combining hierarchical clustering approaches using the PCA method’, *Expert Systems with Application*, Vol. 137, pp.1–10, ISSN: 0957-4174, DOI: 10.1016/j.eswa.2019.06.064.

Jain, A.K., Murty, M.N. and Flynn, P.J. (1999) ‘Data clustering: a review’, *ACM Computing Surveys*, Vol. 31, No. 3, pp.264–323, DOI:10.1145/331499,331504.

Lee, D.D. and Seung, H.S. (2001) ‘Algorithms for non-negative matrix factorization’, *Advances in Neural Information Processing Systems*, Vol. 13, pp.556–562.

Lee, J. and Jun, C. (2013) ‘PCA-based high-dimensional noisy data clustering via control of decision errors’, *Knowledge-Based Systems*, Vol. 37, pp.338–345, ISSN: 0950-7051, DOI: 10.1016/j.knosys.2012.08.013.

Liu, J., Wang, D., Gao, Y., Zheng, C., Xu, Y. and Yu, J. (2018) ‘Regularized non-negative matrix factorization for identifying differentially expressed genes and clustering samples: a survey’, *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, Vol. 15, No. 3, DOI: 10.1109/TCBB.2017.2665557.

Long, M., Wang, J., Ding, G., Shen, D. and Yang, Q. (2014) 'Transfer learning with graph co-regularization', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 26, No. 7, pp.1805–1818.

Pauca, V.P., Piper, J. and Plemmons, R.J. (2006) 'Nonnegative matrix factorization for spectral data analysis', *Linear Algebra and Its Applications*, Vol. 416, No. 1, pp.29–47.

Pujari, A.K. (2019) *Data Mining Techniques*, 4th ed., Universities Press, Private Limited, India.

Strang, G. (2006) *Linear Algebra and Its Applications*, 4th ed., Cengage Learning, India Private Limited.

Tan, P., Steinbach, M. and Kumar, V. (2018) *Introduction to Data Mining*, 6th Impression, India Education Services Pvt. Ltd., Addison Wesley, Pearson.

Vinh, N.X., Epps, J. and Bailey, J. (2010) 'Information theoretic measures for clustering comparison: variants, properties, normalization and correction for chance', *Journal of Machine Learning Research*, Vol. 11, pp.2837–2854.

Winck, R.C., Kim, J., Book, J.W. and Park, H. (2012) 'Command generation techniques for an array using the SVD and the SNMF', *10th IFAC Symposium on Robot Control*.

Zadeh, M.B., Jutten, C. and Mansour, A. (2006) 'Sparse ICA via cluster-wise PCA', *Neurocomputing*, Vol. 69, pp.1458–1466, ISSN: 0925-2312, DOI: 10.1016/j.neucom.2005.12.022.

Zeng, M., Zhang, W. and Chen, Z. (2019) 'Group based k-SVD denoising for bearing fault diagnosis', *IEEE Sensors Journal*, Vol. 19, No. 15, pp.6335–6343, DOI: 10.1109/JSEN.2019.2910868.

Zhu, C., Idemudia, C.U. and Feng, W. (2019) 'Improved logistic regression model for diabetes prediction by integrating PCA and k-means techniques', *Informatics in Medicine Unlocked*, ISSN: 2352-9148, DOI: 10.1016/j.imu.2019.100179.