Artikel_ICAISD_Elly.pdf

by

Submission date: 20-Jul-2020 01:55PM (UTC+0700)
Submission ID: 1359841917
File name: Artikel_ICAISD_Elly.pdf (439.66K)
Word count: 2484
Character count: 13205

Comparative Analysis on Dimension Reduction Algorithm of Principal Component Analysis and Singular Value Decomposition for Clustering

Elly Muningsih[1], Hidayat Muhammad Nur[2], Fabriyan Fandi Dwi Imaniawan[3*], Saifudin[4], Vembria Rose Handayani[5], Feri Endiarto[6]

[1,2,5]Program Studi Sistem Informasi, Universitas Bina Sarana Informatika, Indonesia
[3]Program Studi Sistem Informasi, Sekolah Tinggi Manajemen Informatika dan Komputer Nusa Mandiri, Indonesia
[4]Program Studi Teknologi Komputer, Universitas Bina Sarana Informatika, Indonesia
[6]Program Studi Bahasa Inggris, Universitas Bina Sarana Informatika, Indonesia

E-mail: fabriyan.fbf@nusamandiri.ac.id

Abstract. Clustering is a method of dividing datasets into several groups that have similarity or the same characteristics. High-dimensional Datasets will influence the effectiveness of the grouping process. This study compares two dimension reduction algorithms, namely Principal Component Analysis (PCA) and Singular Value Decomposition (SVD) using K-Means clustering method to find out the best algorithm with the smallest Bouldin Davies Index evaluation. The dataset of this study involved public data from UCIMachine Learning which contains the number of weekly sales of a product. Data processing is performed by comparing the number of clusters from 3 to 10 and the dimension reduction from 2 to 10. From the data processing the RapidMiner tools, application with dimension reduction can provide better results than without dimension reduction. In particular, the PCA algorithm shows better results than the SVD, with which the best number of clusters is 5, and the number of dimensional reductions is 3 with a Bouldin Index of 0.376.

1. Introduction

Clustering or data grouping has long been studied and undeniably has provided benefits for many human activities such as science, business, machine learning, data mining, knowledge discovery and pattern recognition [7]. Clustering comprises partitioning a set of n objects in k$latex \leq$2 non-empty subsets (called clusters) in such a way that objects in a cluster share the same attributes, while the other clusters have different objects [7]. The purpose of clustering is to identify a set of unlabelled datasets by organizing data objectively into the homogeneous groups with minimizing similarity in object groups and maximizing differences between group objects [10]. In addition, clustering also aims to find homogeneous number of classes which are assumed to lie in low-dimensional data sub-spaces and generally grouping data is intended to visualize clusters in reduced dimensional space [2]. Clustering is useful to identify a distribution pattern in a dataset to facilitate the data analysis process [11]. A dataset which has a high dimensional space will influence the effectiveness of the grouping process [6].

Dimension reduction techniques are crucial parts in the clustering process because processing high-dimensional data is challenging. This technique is purposed to reduce dimensions by

altering existing features into a new low dimension space [6]. With the dimension reduction technique, the number of input variables can be reduced and the model dimension reduction can also be realized. Furthermore, information redundancy and computational complexity can significantly decrease [5]. Dimensional reduction techniques include PCA and SVD. Principal Component Analysis (PCA) is a technique used for collecting high dimensional data and subsequently using dependencies between variables to represent the data more systematically to form low dimensions without losing substantial information in the dataset [3]. PCA is the most common and the most widely used dimension reduction technique [5].

Meanwhile, Singular Value Decomposition (SVD) is a matrix decomposition algorithm and technique of feature transformation, where new features are generated from the original data [6]. SVD is a strong and reliable method for orthogonal matrix decomposition, in which the main property of SVD is its relationship with the rank of matrices and its ability to estimate the matrices from an assigned rank [9]. PCA is related to analysis of Principal Component, and is related to analysis of Principal Component and SVD which is seen as a more basic technique because it does not only provides a direct approach for calculation of main components (PC), but also reduces PCA in row and column spaces simultaneously [13].

This study performs comparative analysis of PCA and SVD dimension reduction techniques with clustering methods to find the best number of clusters and dimensional reductions by evaluating the Davies-Bouldin Index (DBI). The clustering method used for comparison is the K-Means method which is a method of unsupervised data mining and partitioning data [11]. K-Means is frequently used because it can group large amounts of data in relatively fast and efficient computation time [1]. Meanwhile, Davies-Bouldin Index is a method to evaluate cluster validity in a clustering where the principal of DBI measurement is to maximize the distance between clusters and at the same time to minimize the distance between points in a cluster [4]. The smallest DBI value represents the best among the other DBI values.

The dataset used in this study is public data obtained through UCIMachine Learning, which consists of a weekly sales transaction of a product for 52 weeks with a total data point of 811. Data processing using RapidMiner tools is performed by cluster comparison from 3 to 10 and dimension reduction from 2 to 10. From data processing, it is expected that the implementation of dimension reduction can provide better outcomes than without implementation. Additionally, it can also be revealed which one is the better dimension reduction algorithm between PCA or SVD algorithm for data of product sale transaction.

2. Relevant Studies

A study [6] has compared three algorithms of dimension reduction for text clustering, namely Principal Component Analysis (PCA), Non-Negative Matrix Factorization (NMF), and Singular Decomposition Value (SVD). The experiment was carried out using two corpora linguistics of English and Arabic by analyzing the results based on the clustering quality. From the results of the data processing, the study has shown that PCA improves the quality of the clustering process and provides better results with shorter time of processing for Arabic and English documents. Meanwhile, another study [12] has proposed a new collaborative algorithm for data filtering recommendations based on dimension reduction and clustering techniques. The K-Means clustering and the Singular Value Decomposition (SVD) dimension reduction algorithm have been used to cluster the same users and to reduce dimensions. The study proposed and assessed two stages of the effective system recommendation that can yield a highly accurate and efficient recommendation. The result of the experiment has indicated that this new method significantly increases the performance of the recommendation system.

3. Method

The CRISP-DM (Cross-Industry Standard Process for Data Mining) method was used to build a model in this study. This method has several phases, namely data set collection, selection of relevant attributes, building clustering models without and with dimension reduction algorithms, and using clustering models for data clustering and model evaluation [8].

3.1. Dataset

The dataset used in this study is transaction data collected from UCIMachine Learning with a total of 811 initial data and 104 attributes. The dataset contains the number of product sales within a period of 52 weeks, the minimum and maximum data, and normal value data. More complete information on the transaction data involves:

- `Product_Code` P1, P2, P3, ..., P819. Some data of `product_code` is missing.
- Data of 52 weeks W0, W1, ..., W51.
- Data of minimum sale: MIN
- Data of maximum sale: MAX
- Normalised values of weekly data: Normalised 0, Normalised 1, ..., Normalised 51

3.2. Data Pre-processing

From the existing dataset, the attributes to be used were selected. The attribute used was the product code as an id for special attributes and attributes W0 up to W51 for regular attributes that were processed later. The attributes were an integer.

Product_Co...	W0	W1	W2	W3	W4	W5	W6	W7
P1	11	12	10	8	13	12	14	21
P2	7	6	3	2	7	1	6	3
P3	7	11	8	9	10	8	7	13
P4	12	8	13	5	9	6	9	13
P5	8	5	13	11	6	7	9	14
P6	3	3	2	7	6	3	8	6
P7	4	8	3	7	8	7	2	3
P8	8	6	10	9	6	8	7	5
P9	14	9	10	7	11	15	12	7
P10	22	19	19	29	20	16	26	20

ExampleSet (811 examples, 1 special attribute, 52 regular attributes)

Figure 1. Pre-processing of Dataset

4. Modeling and Evaluation

In this step, several steps are made for modeling using the RapidMiner tools, as illustrated in Figure 2 which shows the K-Means + PCA clustering model

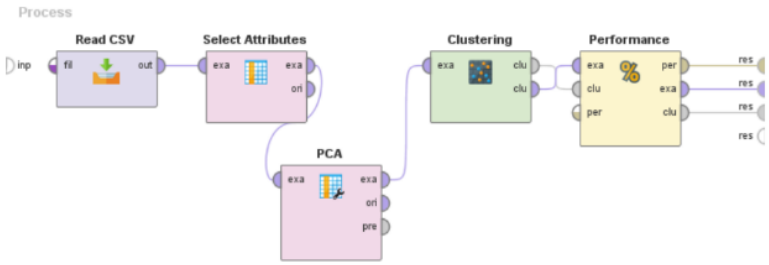

Figure 2. Process of K-Means+PCA clustering model

The steps in the modeling and evaluation process are as follows:

- Creating clustering model using the K-Means method, searching for cluster's DBI value from cluster 3 to 10.
- Followed by constructing a clustering model using the K-Means + PCA method where the number of clusters is 3 to 10 and the number of inserted dimensional reductions is 2 to 10. From each process, the DBI value is noted.
- The next process is to create the clustering model using the K-Means + SVD method where the number of clusters is also the same, 3 to 10, and the number of dimensional reductions is 2 to 10. From each process, the DBI value is also noted.
- Subsequently, a comparison between the model the K-Means + PCA method and the K-Means + SVD method. The smallest DBI value indicates the best results.

5. Results and Discussion

The following table 1 shows the DBI value for the K-Means clustering model with 3 to 10 clusters. Based on the table, the larger the cluster the greater the DBI value.

Table 1. K-Means Modelling cluster

Cluster	3	4	5	6	7	8	9	10
K-Means	0.626	0.864	0.777	1.988	1.939	2.204	2.342	2.178

From the data processing that has been done, it is known that the DBI values for the K-Means + PCA and K-Means + SVD clustering models. From the data, it appears that:

- In general and the overall DBI value of the K-Means clustering model is greater than the K-Means + PCA and K-Means + SVD modeling.
- For K-Means + PCA modeling, almost all of the DBI values are smaller than the DBI values of K-Means clustering model, except for cluster 5, where the DBI value of the K-Means clustering model is 0.777 while the DBI of K-Means + PCA clustering model is 0.780.

- In K-Means + SVD modeling, there is only one DBI value in the 3rd cluster which has smaller valeue; it is 2 reduction, with a DBI value of 0.461 compared to the K-Means clustering model value of 0.464. In the 4th cluster of K-Means + SVD modeling there are two smaller DBI values, for which 2 reduction is 0.607 and 3 reduction is 0.729. In the 5th cluster, there is only one DBI value of K-Means + SVD modeling, which is smaller than the value of the K-Means clustering model; it is reduction 2, with 0.582. In cluster 6 there are eight DBI values that are smaller than the K-Means clustering model value, of which DBI values for 2 to 9 reduction are 0.639 ; 0.854; 0.968; 1,130; 1,342; 1,660; 1,839 and 1,944 respectively. In cluster 7, 8, 9, and 10 DBI values of K-Means + SVD modeling are smaller than those of K-Means clustering model.
- The smallest DBI value among the models is obtained in cluster 5 and with 3 dimension reduction, with 0.376. Thus, it can be concluded that the best cluster for the data is cluster 5 with 2 dimension reduction. The DBI value of cluster 5 is shown in table 2, while Figure 3 shows a comparison of DBI values for 3 different models.

Table 2. Modelling Cluster 5

Cluster 5									
Dimension Reduction	2	3	4	5	6	7	8	9	10
K-Means + PCA	0.573	**0.376**	0.647	0.684	0.707	0.732	0.758	0.780	0.502
K-Means + SVD	0.582	0.810	0.928	1.170	1.481	1.679	1.864	1.966	2.101

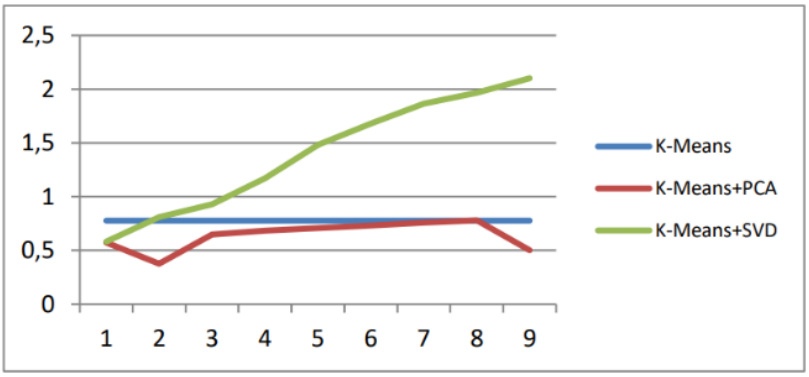

Figure 3. Comparison of DBI Value

After retaining the most optimal number of clusters, which is 5, the RapidMiner tool subsequently also revealed the results of the number of components for each cluster, as shown in table 3. Table 3 presents the number of components of each cluster, member id of each cluster, and information from each cluster.

Table 3. Information of Each Cluster

Cluster	Number of Components	Component ID	Information
1	482	2, 6, 7, 12, 23, 53, 77, 98,...	Cluster with lowest weekly sales
2	119	15, 16, 17, 18, 19, 24, 25, 27,...	Cluster with second-highest weekly sales
3	45	10, 51, 62, 107, 200, 202, 261, 263,...	Cluster with medium weekly sales
4	164	1, 3, 4, 5, 8, 9, 11, 13, 14, 20,...	Cluster with second-lowest weekly sales
5	1	407	Cluster with highest weekly sales

Figure 4 displays the centroid values of each cluster for PC1, PC2 and PC3 which are the result of dimension reduction using PCA.

Attribute	cluster_0	cluster_1	cluster_2	cluster_3	cluster_4
pc_1	-53.027	179.391	66.428	5.972	243.017
pc_2	-1.495	-5.634	16.072	3.641	70.870
pc_3	-0.149	-0.357	1.624	0.228	3.513

Figure 4. Values of centroid of clusters

6. Conclusion

From the data processing, it can be concluded that the PCA algorithm provides better results than SVD for dimension reduction of the data used in this study. Cluster 5 is the best number of clusters that can be used with the number of reduced dimension of 3. Due to time and energy constraints, the researchers are aware that the results of this study remain far from perfect. Therefore, subsequent comparative studies are needed using different and more diverse datasets or using another dimension reduction algorithm.

References

[1] Ahmar, A. S., Napitupulu, D., Rahim, R., Hidayat, R., Sonatha, Y., and Azmi, M, 2018 Using K-Means Clustering to Cluster Provinces in Indonesia. Journal of Physics: Conference Series, 1028(1)

[2] Allab, K., Labiod, L., and Nadif, M, 2017 A Semi-NMF-PCA Unified Framework for Data Clustering. IEEE Transactions on Knowledge and Data Engineering, 29(1), p. 2-16

[3] Dash, P., Nayak, M., and Prasad Das, G, 2014 Principal Component Analysis using Singular Value Decomposition for Image Compression. International Journal of Computer Applications, 93(9), p. 21–27

[4] Jumadi Dehotman Sitompul, B., Salim Sitompul, O., and Sihombing, P, 2019 Enhancement Clustering Evaluation Result of Davies-Bouldin Index with Determining Initial Centroid of K-Means Algorithm. Journal of Physics: Conference Series, 1235(1)

[5] Luo, S., Chen, T., and Jian, L, 2018 Using principal component analysis and least squares support vector machine to predict the silicon content in blast furnace system. International Journal of Online Engineering, 14(4), p. 149-162

[6] Mohamed, A. A, 2019 An effective dimension reduction algorithm for clustering Arabic text. Egyptian Informatics Journal, (xxxx), p. 0-4

[7] Pérez-Ortega, J., Almanza-Ortega, N. N., and Romero, D, 2018 Balancing effort and benefit of K-means clustering algorithms in Big Data realms. PLoS ONE, 13(9), p. 1–19

[8] Sastry, S. H., and Babu, P. M. S. P, 2013 Implementation of CRISP Methodology for ERP Systems. 2(05), 203–217. Retrieved from http://arxiv.org/abs/1312.2065

[9] Swathi, H. R., Sohini, S., Surbhi, and Gopichand, G, 2017 Image compression using singular value decomposition. IOP Conference Series: Materials Science and Engineering, 263(4), p. 5-8

[10] Warren Liao, T, 2005 Clustering of time series data - A survey. Pattern Recognition, 38(11), 1857-1874

[11] Widiyaningtyas, T., Prabowo, M. I. W., and Pratama, M. A. M, 2017 Implementation of k-means clustering method to distribution of high school teachers. International Conference on Electrical Engineering, Computer Science and Informatics (EECSI), 2017-December (September), p. 19-21

[12] Zarzour, H., Al-Sharif, Z., Al-Ayyoub, M., and Jararweh, Y, 2018 A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques. 2018 9th International Conference on Information and Communication Systems, ICICS 2018, 2018-Janua, p. 102-106

[13] Zhang, L., Marron, J. S., Shen, H., and Zhu, Z. (2007). Singular value decomposition and its visualization. Journal of Computational and Graphical Statistics, 16(4), p. 833-854

Artikel_ICAISD_Elly.pdf

ORIGINALITY REPORT

13%	6%	9%	6%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	journals.plos.org Internet Source	2%
2	Submitted to University of Liverpool Student Paper	1%
3	A Supriyatna, I Carolina, W Widiati, C Nuraeni. "Rice Productivity Analysis by Province Using K-Means Cluster Algorithm", IOP Conference Series: Materials Science and Engineering, 2020 Publication	1%
4	Selma Benkessirat, Narhimène Boustia, Nachida Rezoug. "Chapter 33 Overview of Recommendation Systems", Springer Science and Business Media LLC, 2019 Publication	1%
5	Triyanna Widiyaningtyas, Martin Indra Wisnu Prabowo, M. Ardhika Mulya Pratama. "Implementation of K-means clustering method to distribution of high school teachers", 2017 4th International Conference on Electrical	1%

	Engineering, Computer Science and Informatics (EECSI), 2017 Publication	
6	icaisd.info Internet Source	1%
7	Submitted to University of Bridgeport Student Paper	1%
8	www.ijcaonline.org Internet Source	1%
9	Shihua Luo, Tianxin Chen, Ling Jian. "Using Principal Component Analysis and Least Squares Support Vector Machine to Predict the Silicon Content in Blast Furnace System", International Journal of Online Engineering (iJOE), 2018 Publication	1%
10	Tao Li, Yongzhen Ren, Yongjun Ren, Lina Wang, Lingyun Wang, Lei Wang. "NMF-Based Privacy-Preserving Collaborative Filtering on Cloud Computing", 2019 International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData), 2019 Publication	1%

11	scimag.com Internet Source	1%
12	lume.ufrgs.br Internet Source	1%
13	Submitted to Universiti Kebangsaan Malaysia Student Paper	<1%
14	worldwidescience.org Internet Source	<1%
15	tuprints.ulb.tu-darmstadt.de Internet Source	<1%
16	Submitted to University of Nottingham Student Paper	<1%
17	Tao Li, Yanqing Wang, Yongjun Ren, Yongzhen Ren, Qi Qian, Xi Gong. "Nonnegative matrix factorization-based privacy-preserving collaborative filtering on cloud computing", Transactions on Emerging Telecommunications Technologies, 2020 Publication	<1%
18	eprints.utm.my Internet Source	<1%
19	epdf.pub Internet Source	<1%
20	N R Radliya, M R Fachrizal, A R Rabbi.	<1%

	"Monitoring Application for Clean Water Access and Clustering using K-Means Algorithm", IOP Conference Series: Materials Science and Engineering, 2019 Publication	
21	Submitted to University of Central Florida Student Paper	<1%
22	eprints.qut.edu.au Internet Source	<1%

Exclude quotes	On	Exclude matches	< 5 words
Exclude bibliography	On		