Comparison of dimensional reduction using the Singular Value Decomposition Algorithm and the Self Organizing Map Algorithm in clustering result of text documents

To cite this article: Muhammad Ihsan Jambak and Ahmad Ikrom Izzuddin Jambak 2019 IOP Conf. Ser.: Mater. Sci. Eng. 551 012046

View the article online for updates and enhancements.
Comparison of dimensional reduction using the Singular Value Decomposition Algorithm and the Self Organizing Map Algorithm in clustering result of text documents

Muhammad Ihsan Jambak*, and Ahmad Ikrom Izzuddin Jambak
Faculty of Computer Science, Sriwijaya University, Palembang, Indonesia.

E-mail: jambak@unsri.ac.id

Abstract. Dimension reduction has two methods namely feature selection and feature extraction. Dimension reduction using the feature selection method has a better influence than the feature extraction method on the cluster results. However, there is still a need for feature extraction methods to reduce dimensions. For this reason, an alternative algorithm is needed from the feature extraction method. Self Organizing Map (SOM) is one of the artificial neural network models that has a special nature that is effectively able to create spatial internal representations of input data, or in general to create smaller data dimensions. This research was examining the capability of SOM compared to Singular Value Decomposition (SVD) in reducing data dimension of text documents before there were clustered by k-Means. Results show that SVD still better than SOM in cluster quality index but SOM faster than SVD in computation times.

1. Introduction

High-dimensional data is data that has a large number of features. In high-dimensional data, it allows complex data to occur which is data that has noise, anomalies (outliers), missing meaning (missing value) and discontinuities between data [1, 2]. Some examples of data that have high dimensions are text document data, image data, and gene expression data. In the text document clustering process, each dimension or feature of the data has an important influence. The most frequently used algorithm for clustering is the k-Means algorithm that groups data into groups of k, with k being the desired number of groups [3]. The k-Means algorithm has a faster data processing time compared to other clustering algorithms.

Each data feature will affect the location of a data when grouped into a class. However, when a data has many features and is too diverse, clustering algorithms such as the k-Means algorithm are not able to find the closeness or similarity between data so that the data cannot be grouped properly. This is often called curse of dimensionality [4]. To get good results, high-dimensional data must go through the initial processing or pre-processing, namely dimension reduction. Dimension reduction is a process of reducing random variables or nonessential variables with certain considerations [5].

Dimension reduction has two methods namely feature selection and feature extraction. Dimension reduction using the feature selection method has a better influence than the feature extraction method on the cluster results using the k-Means algorithm [6]. The feature extraction algorithm that is often used is Principal Component Analysis (PCA). However, the PCA algorithm is often used as a dimensional reduction algorithm in several studies. This indicates that there is still a need for feature
extraction methods to reduce dimensions. For this reason, an alternative algorithm is needed from the feature extraction method.

Artificial Neural Network (ANN) is a calculation model that is inspired by the workings of nerves in the human body. The ANN does not stand as a specific algorithm, but rather a framework that forms the basis of various other calculation models [7]. Self Organizing Map (SOM) is one of the ANN models that are trained to use unsupervised data. In contrast to other ANN algorithms, SOM has a special nature that is effectively able to create spatial internal representations of input data [8]. In general, the purpose of mapping the SOM algorithm is to create smaller data dimensions.

[9] merged SOM with the Density-Based Spatial Clustering of Applications with Noise image segmentation algorithm (DBSCAN), in this study SOM was used as a reduction in the dimensions of image data which is high dimensional data. Others study were conducted by [10] and [11] who implemented SOM to reduce high-dimensional data, the results of this study showed the ability of SOM to project non-linear data into lower dimensions. Based on the description above, it is necessary to do research on the SOM algorithm as an alternative to the PCA algorithm in reducing feature extraction dimensions.

2. Research Methodologies

2.1. Data and Pre-processing Data

The data test used English language journals that are downloaded through the several websites which is stored in the ASCII file (*.txt). With total 40 journals and text terms obtained from pre-processing results amounted to 35,000 words. The pre-processing stage is the conversion of data formats, a series of actions to convert input data from text to numerical so that the clustering process can be carried out. Stages carried out on are case folding, tokenizing, stop words removal, and stemming. Furthermore, the terms of the research test data are given weights using the tf-idf (term frequency-inverse document frequency).

2.2. Singular Value Decomposition

Singular Value Decomposition is a form of factor analysis on the matrix. In SVD matrix contains the frequency of occurrence of keywords decomposed into three matrix components [12]. The first matrix component (U) describes the line entity as an orthogonal matrix vector. The second matrix component (S) is a diagonal matrix containing the scalar value of the matrix. And the third component (V) is the column entity matrix as an orthogonal matrix vector.

The singular value decomposition of the weighted matrix will be used in the linear transformation approach. SVD is basically to estimate rank from the matrix. If matrix $H$ is known with dimensions $m \times n$, where $m \geq n$ and $\text{rank}(H) = r$, as follows SVD equation is:

$$H = U S V^T$$

where,$$
U^T U = V^T V = I_n\tag{2}
$$
and fulfill the conditions,$$
S = \text{diag}(\sigma_1, \ldots, \sigma_r)\tag{3}
$$
where,$$
\sigma_i > 0 \text{ for } 1 \leq i \leq r \text{ and } \sigma_j = 0 \text{ for } j \geq r + 1.
$$

2.3. Self-Organizing Map

Self-Organizing Map is one of the techniques in a neural network that uses unsupervised learning methods that use competitive learning methods. SOM aims to map high-dimensional data into the form of low-dimensional data by assuming a structured topology makes new class units based on data similarity [8, 13]. To project data into lower dimensions, SOM uses a weight vector, which is initially
randomly determined and then becomes a reference in grouping input vectors. Euclidean Distance is used to measure the distance between the input vector and the weight vector, with the input vector closest to the value with the weight vector will be selected as the winning neuron. The weight vector will then be updated based on the results of the first iteration using the alpha / learning rate multiplier which will always decrease if there is no change.

According to Haykin [14] there are three important components in SOM, namely: Competition, for each input pattern neurons calculate the value of each discriminant function that gives the basis for competition. Certain neurons with the smallest value of the discriminant function are declared winners; Cooperation, winning neurons determine the spatial location of the excited neuron topology environment to provide a basis for cooperation in a neuron environment; Synaptic Adaption, excited neurons decrease the value of the discriminant function associated with the input pattern through the corresponding weight adjustment so that the response of the winning neurons to the next application with the same input pattern will increase.

Following are the steps of SOM: (1). Determine the initial weight value randomly (0 to 1) a number of k dimensions; (2). Determine the value of learning rate and the number of iterations, in this study an experiment will be conducted first to determine the learning rate value and the best number of iterations; (3). Calculate the distance between input data and weights using the Euclidean Distance formula; (4). Determine the winning neurons, namely neurons whose weight values are closest to the input data; (5). Update the weight value; (6). Repeat steps using data to i + 1 until the iteration reaches the specified amount.

2.4. k-Means Clustering
k-Means is one of the frequently used partitioning methods due to its simplicity, this technique group objects into k groups. To do this clustering, the value of k must be determined first. The process that k-means is to choose k centres as a centroid for use in clustering data. The centroid point will determine success in the k-means algorithm. If the point of the centroid selected at the beginning is wrong, then the results obtained will not be optimal. In k-means the centroid point is often a problem that must be overcome because of random selection. The k-means algorithm classifies data with the number of clusters (k values) that have been predetermined; it is possible that the k value is the number of clusters desired or the value of k entered randomly. Entering the number of clusters that are not ideal in the data can cause poor clustering quality, allowing data not to be in the actual cluster.

The steps in the k-Means algorithm are as follow [2, 5, 15, 16]: (1) Determine the number of clusters (k) as the number of clusters to be formed; (2) Generate k centroids (starting point) randomly; (3) For each data, locate the nearest cluster centre, so each cluster centre has a subset of datasets; (4) For each cluster k, find the centre of the cluster area, update the location of each cluster centre to the new value of the centre area; (5) Repeat steps 3 and 5 to the data in each cluster becomes centralized or completed.

2.5. Davies Bouldin Index
For the success of the clustering process then one of the important issues is cluster validation. Although in general clustering validation is categorized into external clustering validation and internal clustering validation; but we realize that the purpose of clustering is to group objects in the same or most similar cluster and different objects in different clusters. For this reason, internal evaluation is more suitable.

The internal clustering validation step is based on 2 criteria, the first is Compactness, the members of each cluster must be close to each other. The typical measure of compactness is the variant. The second is Separateness which measures how different or separate clusters are from other clusters [17, 18]. One calculation of internal clustering validation is the Davies Bouldin Index (DBI) with the smaller the index value, the better the clustering results. DBI is based on the similarity of the Rij size of the cluster. Rij is a cluster with Si as dispersion size and dij as distance.
where,

\[ d_{ij} = d(x_i, v_j) \]  

(5)

\[ S_i = \frac{1}{|R_i|} \sum_{x \in R_i} d(x, v_i) \]  

(6)

Therefore, the Davies Bouldin Index (DBI) is:

\[ DBI = \frac{1}{nc} \sum_{i=1}^{nc} R_i \]  

with \( n_c \) is amount of cluster, then:

\[ R_i = \max_{j \neq i} \frac{d_i}{d_j} \]  

(7)

As a comparison reference, the original data is first clustered with k-Means by conducting various experiments to get the most optimum k value. The measurement to determine the most optimum k value is to calculate the DBI value and find the lowest value. Next, the original data is reduced using either SVD or SOM. Each data that has been reduced is clustered with k-Means and each DBI value is measured.

In addition to DBI values that describe quality rather than clusters, computational time was also observed and evaluated for each test both in the original data clustering by k-Means and in clustering data reduced by SVD and SOM. Furthermore, all clustering performance both without dimension reduction and dimension reduction by SVD and SOM were compared with statistical tests, Analysis of Variance (ANOVA).

3. Results

Testing is done with the help of RapidMiner Studio Version 9.0 software. The first test is testing the original data to obtain the most optimum k value. There have been 12 parameter variations were used where for each variations was tested 10 times; and obtained the optimum k value at k = 3. The variation of the k-Means parameters were done by changing the values of determined or not good start value; use or not local random seed value; and determining random seed value. Furthermore, the value of k = 3 is used in subsequent tests for data that have been reduced by SVD and SOM, with variations in the same k-Means parameter settings.

| Parameter Variations - # | k-Means | SVD + k-Means | SOM + k-Means |
|---------------------------|---------|---------------|---------------|
|                           | Average DBI | Average Time (sec) | Average DBI | Average Time (sec) | Average DBI | Average Time (sec) |
| 1                         | 4.344    | 18            | 0.395        | 43            | 0.762        | 32            |
| 2                         | 4.489    | 16            | 0.395        | 42            | 0.762        | 31            |
| 3                         | 4.391    | 15            | 0.395        | 45            | 0.762        | 28            |
| 4                         | 4.542    | 15            | 0.395        | 57            | 0.750        | 30            |
| 5                         | 4.725    | 20            | 0.395        | 66            | 0.762        | 33            |
| 6                         | 4.491    | 14            | 0.395        | 43            | 0.762        | 28            |
| 7                         | 4.613    | 14            | 0.395        | 42            | 0.762        | 31            |
| 8                         | 4.275    | 14            | 0.395        | 45            | 0.762        | 30            |
| 9                         | 4.524    | 15            | 0.395        | 44            | 0.762        | 29            |
| 10                        | 4.602    | 14            | 0.395        | 43            | 0.762        | 30            |
| 11                        | 4.590    | 21            | 0.395        | 43            | 0.762        | 28            |
| 12                        | 4.522    | 17            | 0.395        | 48            | 0.762        | 28            |

But the dimension reduction process adds the total computing time significantly as well, which originally was 16.08 seconds to 46.75 seconds by SVD and becomes 29.83 seconds by SOM. Or the dimensional reduction process by SVD has an impact on increasing computing time almost 3 times
longer; while dimension reduction by SOM is faster than SVD because it only adds 1.8 times longer computing time.

4. Conclusion
SVD reduction techniques work well reduce the data by processing the entire data matrix (using input term weighting), by reducing the multiplication of the matrix solutions to a minimum, as a result SVD produces a good representation of cluster input data. Although the SOM reduction technique is also able to provide significant results but not as good as the SVD, but SOM has an advantage over SVD, which is significantly faster computation time.

Because in this study the focus was on high-dimensional data originating from text documents, then feature extraction reduction techniques could be considered less suitable to process text data dimension reduction. This is because the extraction technique only takes dataset with the highest variance. While the text dataset presented with weights word has a different variance values in each document, so that the input data used do not represent all data values.

References
[1] Chen T C et al 2009 World Congress on Computer Science and Information Engineering pp. 670-673.
[2] Jambak M I et al 2018 The Impacts of Singular Value Decomposition Algorithm Toward Indonesian Language Text Documents Clustering International Conference of Reliable Information and Communication Technology 173-183.
[3] Jin X and Han J 2017 K-means Clustering Encyclopedia of Machine Learning and Data Mining ed: Springer 695-697.
[4] Keogh E and Mueen A 2017 Curse of dimensionality Encyclopedia of Machine Learning and Data Mining, ed: Springer 314-315.
[5] Han J et al 2011 Data mining: concepts and techniques: Elsevier.
[6] Hasanah SIR et al 2018 Comparison of dimensional reduction using singular value decomposition and principal Component analysis for clustering results of Indonesian language text documents The 2nd International Conference on Applied Sciences, Mathematics, and Informatics (ICASMI), Bandar Lampung
[7] van Gerven M and Bohte S 2017 Editorial: Artificial Neural Networks as models of neural information processing Frontiers in Computational Neuroscience 11.
[8] Kohonen T 1990 The self-organizing map Proceedings of the IEEE 78 1464-1480.
[9] Chen Q et al 2017 Towards a Hybrid Approach of Self-Organizing map and density-Based spatial clustering of applications with noise for image segmentation Developments in eSystems Engineering (DeSE) 10th International Conference 238-241.
[10] Campoy P 2009 Dimensionality reduction by self organizing maps that preserve distances in output space Neural Networks IJCNN 2009. International Joint Conference 432-438.
[11] Cong T et al 2013 Applications of ANNs in flow and heat transfer problems in nuclear engineering: a review work Progress in Nuclear Energy 62 54-71.
[12] Leskovec J et al 2014 Mining of massive datasets: Cambridge university press.
[13] Kohonen T 2013 Essentials of the self-organizing map Neural Networks 37 52-65.
[14] Haykin S 1999 Multilayer perceptrons Neural networks: a comprehensive foundation 2 156-255.
[15] Madhulatha T S 2011 Comparison between k-means and k-medoids clustering algorithms Advances in Computing and Information Technology, ed: Springer 472-481.
[16] Nur’aini K et al 2015 Combination of singular value decomposition and K-means clustering methods for topic detection on Twitter Advanced Computer Science and Information Systems (ICACSIIS), 2015 International Conference 123-128.
[17] Ristevski B et al 2008 A comparison of validation indices for evaluation of clustering results of DNA microarray data Bioinformatics and Biomedical Engineering ICBBE 2008. The 2nd International Conference 587-591.

[18] Wiwie C et al 2015 Comparing the performance of biomedical clustering methods Nature methods 12, 1033.