An Observation of Different Clustering Algorithms and Clustering Evaluation Criteria for a Feature Selection Based on Linear Discriminant Analysis

K. H. Tie, A. Senawi, and Z. L. Chuan

Abstract Linear discriminant analysis (LDA) is a very popular method for dimensionality reduction in machine learning. Yet, the LDA cannot be implemented directly on unsupervised data as it requires the presence of class labels to train the algorithm. Thus, a clustering algorithm is needed to predict the class labels before the LDA can be utilized. However, different clustering algorithms have different parameters that need to be specified. The objective of this paper is to investigate how the parameters behave with a measurement criterion for feature selection, that is, the total error reduction ratio (TERR). The *k*-means and the Gaussian mixture distribution were adopted as the clustering algorithms and each algorithm was tested on four datasets with four distinct clustering evaluation criteria: Calinski-Harabasz, Davies-Bouldin, Gap and Silhouette. Overall, the *k*-means outperforms the Gaussian mixture distribution in selecting smaller feature subsets. It was found that if a certain threshold value of the TERR is set and the *k*-means algorithm is applied, the Calinski-Harabasz, Davies-Bouldin, and Silhouette criteria yield the same number of selected features, less than the feature subset size given by the Gap criterion. When the Gaussian mixture distribution algorithm is adopted, none of the criteria can consistently select features with the least number. The higher the TERR threshold value is set, the more the feature subset size will be, regardless of the type of clustering algorithm and the clustering evaluation criterion are used. These results are essential for future work direction in designing a robust unsupervised feature selection based on LDA.

Keywords Unsupervised feature selection · Linear discriminant analysis · Clustering algorithm · Clustering evaluation criterion

K. H. Tie · A. Senawi (✉) · Z. L. Chuan
Centre for Mathematical Sciences, College of Computing and Applied Sciences, Universiti Malaysia Pahang, Lebuhraya Tun Razak, 26300 Gambang, Kuantan, Pahang, Malaysia
e-mail: azlyna@ump.edu.my

Z. L. Chuan
e-mail: chuanzl@ump.edu.my

I. M. Khairuddin et al. (eds.), *Enabling Industry 4.0 through Advances in Mechatronics*, Lecture Notes in Electrical Engineering 900, https://doi.org/10.1007/978-981-19-2095-0_42

1 Introduction

Data with large number of features or variables are known as high dimensional data. The number of features of high dimensional data can be two, tens, hundreds, thousands or even up to millions. The larger the number of features means the higher the dimension of the data. Data of high dimensionality are often associated with the problem of high complexities in modelling. In many cases, some of the features are highly correlated and redundant [11]. Therefore, a pre-processing step that transforms high dimensionality data into lower dimensionality data is necessary and thereby allow significant features to be identified.

Lots of work were devoted on dimensionality reduction using mutual information stated in [8] and [22], correlation-based criterion stated in [5] and [21], principal component analysis (PCA) stated in [1] and [23] and linear discriminant analysis (LDA) stated in [2] and [14]. The LDA-based method is one of the popular approaches for feature selection. The focus of LDA is to maximize the separation of multiple classes. It has high performance in classifying unknown dataset and works by finding the discriminating function that gives clear separation of the data samples.

LDA has shortcomings when the data is unsupervised since it was specifically designed for supervised data. In particular, the number of class labels of the data is required for the LDA to be utilized. Nevertheless, it still can be adopted for unsupervised data by performing some clustering onto the data so that the number of class labels can be predicted in advance [24].

It is worth to note that a combination of LDA and *k*-means clustering algorithm is believed to offer higher classification accuracy compared to a combination of PCA and *k*-means or just *k*-means alone [6]. In addition, a combination of LDA with Gaussian mixture clustering is expected to perform well in nonparametric regression [9].

This paper aims to present an analysis of how the number of selected features behave using LDA-based feature selection under different clustering algorithms and clustering evaluation criteria. Note that the feature selection method being used is an unsupervised approach. The fewer the number of features selected, the better the clustering algorithm and the clustering evaluation criterion.

The next section of this paper discusses the clustering algorithms and clustering evaluation criteria that were used to group the samples of the data sets before the LDA-based feature selection can be carried out. The discussions are given in Sect. 2.1 and Sect. 2.2, respectively. The feature selection method applied for the observation is explained in brief in Sect. 2.3, while Sect. 3 describes how the experiment was performed. Section 4 is reserved for presenting the results and discussing the experimental findings. Finally, Sect. 5 delivers some concluding remarks of the study.

2 Related Works

2.1 Clustering Algorithms

Clustering is a process of creating a number of clusters by dividing the data based on their characteristics [19]. Clustering algorithms are used to solve classification problems involving data that do not have target or dependent variables. They will put together the data samples which are similar to each other in the same cluster while separating them as different as possible from other clusters. In this paper, two clustering algorithms are considered: *k*-means and Gaussian mixture distribution. These two clustering algorithms were chosen for the analysis as they have distinct clustering approaches. The *k*-means is a hard-clustering method [12, 17]. On the other hand, Gaussian mixture distribution is a soft clustering method [10, 25]. Hard clustering methods assign the data points only to a cluster only while soft clustering methods may put data into different clusters.

k-means. The *k*-means is a clustering algorithm that makes centroid a base cluster and minimizes the sum of distances between the data samples and their respective cluster centroid. There are many ways of measuring the distance, but Euclidean distance is the most commonly used [20]. The *k*-means has been proven easy to use, consume less memory and has high computation efficiency when compared to other clustering algorithms [18].

Gaussian Mixture Distribution. Gaussian mixture distribution is a clustering algorithm based on the superposition of multiple Gaussian distributions which is also known as Gaussian mixture model (GMM). The advantage of using the Gaussian mixture distribution is that it can create a model-based framework where the number of clusters and contribution of each feature in the clustering process can be shown [16].

2.2 Clustering Evaluation Criterions

Clustering evaluation criterion is a core component in finding different groups in data. Basically, an evaluation criterion is a distance value that quantifies which group a sample belongs. There are many criteria that can be used to determine the best division with an optimal number of clusters for a dataset. Four of them are discussed below: Calinski-Harabasz index, Davies-Bouldin index, Silhouette index and Gap statistic.

Calinski-Harabasz. The Calinski-Harabasz focuses on the ratio of the sum of cluster scattering within a cluster and cluster scattering between the clusters. Higher ratio indicates that the clusters are stable and have a good division between different clusters. This criterion is a good choice when quick results are desired as it is simple and thus can be computed within a short time.

Davies-Bouldin. The Davies-Bouldin evaluates the cluster by comparing the distance between clusters and the size of each cluster. The lower the index means, the better the separation between the clusters. It only needs a few controlled parameters to determine the number of clusters [26]. However, this criterion is limited to Euclidean distance only.

Silhouette. The Silhouette calculates the difference between the distance of points of different clusters and the adjacency of points within the cluster. The difference is then compared to the maximum distance of points of different clusters or the maximum adjacency of points within the cluster. The higher the index value means the better the separation of the clusters. This criterion has its benefit as it can determine whether the data lying in the correct cluster, incorrect cluster or overlapping cluster. It also has been proven to give higher performance efficiency in predicting the number of clusters than the Calinski-Harabasz and Davies-Bouldin [3]. However, it has greater complexity when compared to the Calinski-Harabasz and Davies-Bouldin.

Gap Statistic. The gap statistic compares the changes of dispersion within a cluster with the expected error for the same number of clusters is under a null reference distribution. However, the gap statistic is likely to fail if underestimation and overestimation of cluster occurred. Its performance was found to be relatively higher when combined with k-means compared to basic k-means [7].

2.3 MSOLS Feature Selection Guided by LDA

Multiple sequential orthogonal least squares (MSOLS) algorithm is a feature selection and ranking method to that rank significant features by using principal component analysis (PCA) to guide the selection [4]. In the method, the principal components are treated as dependent variables while the original features are taken as independent variables in the multiple regression models. The measurement criterion used to rank features importance in the MSOLS method is called as total error reduction ratio (TERR). In this paper, a similar approach proposed by [4] is adopted. However, instead of PCA, this paper utilizes the LDA as the dependent variables so as the characteristics of the LDA is taken into account to guide the feature selection and ranking. Detail steps to perform the LDA can be found in [15].

3 Methodology

Four public datasets were used to analyze how the number of selected features using the LDA-based feature selection described earlier behave under different clustering algorithms and clustering evaluation criteria. The clustering algorithms and clustering evaluation criteria discussed in Sect. 2.1 and 2.2, respectively, were applied in the

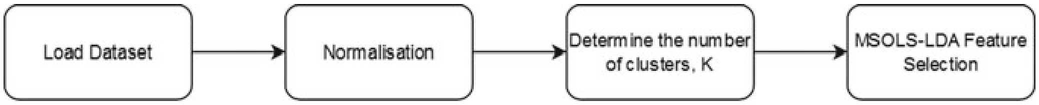
Load Dataset
Normalisation
Determine the number of clusters, K
MSOLS-LDA Feature Selection

Fig. 1 Overall flowchart

experiment. In addition, three different threshold values of the TERR criterion were used to perform the observation: 80, 90 and 95%. One unsupervised dataset (Alate Adelges) and three supervised datasets (Iris, Wine, and Breast Cancer) with neglected class label were employed. Hence, 24 different tests were conducted for each dataset based on the two clustering algorithms, four evaluation criteria and three threshold values. Note that all features of the datasets are numeric values. The number of selected features was recorded for every test.

The Alate Adelges dataset consists of 19 features which were collected based on 40 winged aphids. The full dataset can be found from [13]. The Iris Data consists of 4 features, collected from 3 classes of Iris plants where 50 observations were recorded for each class. The Wine datasets consists of 13 features collected based on 178 observations from 3 types of wine. Meanwhile, the Breast Cancer dataset consists of 10 features which were collected based on 699 observations. The Iris, Wine, and Breast Cancer datasets can be obtained from the UCI Machine Learning Repository. The overall flowchart is shown in Fig. 1.

4 Results and Discussions

The results of the experiments on the four datasets are given in Table 1, Table 2, Table 3, Table 4, Table 5, Table 6, Table 7 and Table 8. It can be observed that the *k*-means algorithm shows either lower or same average number of selected features than the Gaussian mixture distribution algorithm for all threshold values being considered except for the case 95% threshold value when tested with the Alate Adelges dataset. This proves that the feature selection method selects a smaller number of features when coupled with the *k*-means algorithm. The Calinski-Harabasz, Silhouette and

Table 1 Number of selected features of Iris dataset at different threshold values of TERR and clustering evaluation criteria under the *k*-means algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	2	2	2	2	2
90	2	2	3	2	2
95	2	2	3	2	2
Average	2	2	2	2	

Table 2 Number of selected features of Iris dataset at different threshold values of TERR and clustering evaluation criteria under the Gaussian mixture distribution algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	2	2	2	2	2
90	2	2	3	2	2.25
95	2	2	3	2	2.25
Average	2	2	2.67	2	

Table 3 Number of selected features of Alate Adelges dataset at different threshold values of TERR and clustering evaluation criteria under the *k*-means algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	3	3	3	3	3
90	6	6	7	6	6.25
95	13	13	14	13	13.25
Average	7.33	7.33	8	7.33	

Table 4 Number of selected features of Alate Adelges dataset at different threshold values of TERR and clustering evaluation criteria under the Gaussian mixture distribution algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	4	4	3	3	3.5
90	7	7	6	7	6.75
95	12	12	13	10	11.75
Average	7.67	7.67	7.33	6.67	

Table 5 Number of selected features of Wine dataset at different threshold values of TERR and clustering evaluation criteria under the *k*-means algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	7	7	8	7	7.25
90	10	10	10	10	10
95	11	11	11	11	11
Average	7.33	7.33	8	7.33	

Table 6 Number of selected features of Wine dataset at different threshold values of TERR and clustering evaluation criteria under the Gaussian mixture distribution algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	7	7	7	8	7.25
90	10	10	10	11	10.25
95	11	11	11	12	11.25
Average	7.67	7.67	7.33	6.67	

Table 7 Number of selected features of Breast Cancer dataset at different threshold values of TERR and clustering evaluation criteria under the *k*-means algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	5	5	4	5	4.75
90	8	8	7	8	7.75
95	9	9	8	9	8.75
Average	7.33	7.33	6.33	7.33	

Table 8 Number of selected features of Breast Cancer dataset at different threshold values of TERR and clustering evaluation criteria under the Gaussian mixture distribution algorithm

TERR threshold (%)	Clustering evaluation criterion				Average
	Calinski-Harabasz	Silhouette	Gap	Davies-Bouldin	
	Number of selected features				
80	5	5	5	5	5
90	8	8	8	8	8
95	9	9	9	9	9
Average	7.33	7.33	7.33	7.33	

Davies-Bouldin clustering evaluation criteria shows the same average number of selected features for each dataset under the *k*-means algorithm. This can be seen from Table 1, Table 3, Table 5, and Table 7. In the meantime, the Gap statistic fail to give a consistent pattern if compared to the other three criteria under the same clustering algorithm.

If the Gaussian mixture distribution is considered, only the Calinski-Harabasz and Silhouette evaluation criteria seem to show the same number of selected features when the same threshold values are considered through all four datasets. However, none of the criteria able to consistently select features with the least number when the Gaussian mixture distribution is applied.

It can also be observed that the number of selected features can increase significantly when the threshold value increases for all evaluation criteria, under both clustering algorithms as depicted in Table 3 and Table 4. Thus, the higher the TERR threshold value is set, the more the number of features may be selected, regardless which type clustering algorithm as well as the clustering evaluation criterion.

5 Conclusion

This paper analyses how different combinations of clustering algorithms and clustering evaluation criteria behave with an LDA-based feature selection. Through the experiment, it can be inferred that the feature selection method selects a smaller number of features when coupled with the *k*-means algorithm than the Gaussian mixture distribution algorithm. The Calinski-Harabasz, Silhouette, and Davies-Bouldin select the same number of features with the *k*-means algorithm, which is better than the Gap clustering criterion. As for the TERR threshold value, the higher the value is set, the more the number of features will be selected. Thus, one can anticipate that a higher TERR value will lead to a better classification result.

Acknowledgements The authors would like to thank the International Islamic University Malaysia (IIUM), Universiti Malaysia Pahang (UMP) and the Universiti Teknologi MARA (UiTM) for providing financial support under the IIUM-UMP-UiTM Sustainable Research Collaboration Grant 2020 (Vote Number: RDU200722).

References

1. Adegbola OA, Adeyemo IA, Semire FA, Popoola SI, Atayero AA (2020) A principal component analysis-based feature dimensionality reduction scheme for content-based image retrieval system. Telkomnika 18(4):1892–1896
2. Alharbi AS, Li Y, Xu Y (2017) Integrating LDA with clustering technique for relevance feature selection. In: Peng W, Alahakoon D, Li X (eds) Advances in Artificial Intelligence: 30th Australasian Joint Conference. Springer, Melbourne, pp 274–286
3. Baarsch J, Celebi ME (2012) Investigation of internal validity measures for *k*-means clustering. In: Proceedings of the International MultiConference of Engineers and Computer Scientists, pp 471–476. Newswood Limited, Hong Kong
4. Billings SA, Wei HL (2005) A multiple sequential orthogonal least squares algorithm for feature ranking and subset selection. ACSE Research Report (908). University of Sheffield
5. Chormunge S, Jena S (2018) Correlation based feature selection with clustering for high dimensional data. J Electr Syst Inf Technol 5(3): 542–549
6. Ding C, Li T (2007) Adaptive dimension reduction using discriminant analysis and *K*- means clustering. In: Ghahramani Z (ed) ACM International Conference Proceeding Series, vol 227. Association for Computing Machinery, New York, pp 521–528
7. El-Mandouh AM, Mahmoud HA, Abd-Elmegid LA, Haggag MH (2019) Optimized *K*-means clustering model based on gap statistic. Int J Adv Comput Sci Appl (IJACSA) 10(1):183–188
8. Gao W, Hu L, Zhang P (2020) Feature redundancy term variation for mutual information-based feature selection. Appl Intell 50(4):1272–1288

9. Hastie T, Tibshirani R (1996) Discriminant analysis by Gaussian mixtures. J Roy Stat Soc Ser B (Methodol) 58(1):155–176
10. He C, Fu H, Guo C, Luk W, Yang G (2017) A fully-pipelined hardware design for Gaussian mixture models. IEEE Trans Comput 66(11):1837–1850
11. Houari R, Bounceur A, Kechadi MT, Tari AK, Euler R (2016) Dimensionality reduction in data mining. Expert Syst Appl Int J 64(C): 247–260
12. Kamper H, Livescu K, Goldwater S (2017) An embedded segmental *K*-means model for unsupervised segmentation and clustering of speech. In: IEEE Automatic Speech Recognition and Understanding Workshop (ASRU),pp 719–726
13. Krzanowski WJ (2018) Attribute selection in correspondence analysis of incidence matrices. J Roy Stat Soc: Ser C (Appl Stat) 42(3):529–541
14. Kumar, BS, Ravi V (2017) LDA based feature selection for document clustering. In: Proceedings of the 10th Annual ACM India Compute Conference, pp. 125–130. Association for Computing Machinery, New York
15. Lu J, Plataniotis KN, Venetsanopoulos AN (2003) Face recognition using LDA-based algorithms. IEEE Trans Neural Netw 14(1):195–200
16. Maugis C, Celeux G, Martin-Magniette ML (2009) Variable selection for clustering with Gaussian mixture models. Biometrics 65(3):701–709
17. Mohd MRS, Herman SH, Sharif Z (2017) Application of *K*-Means clustering in hot spot detection for thermal infrared images. In: IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE), pp 107–110
18. Morissette L, Chartier S (2013) The *k*-means clustering technique: general considerations and implementation in Mathematica. Tutor Quant Methods Psychol 9(1)
19. Nazari Z, Kang D, Asharif MR, Sung Y, Ogawa S (2016) A new hierarchical clustering algorithm. In: ICIIBMS 2015–International Conference on Intelligent Informatics and Biomedical Sciences, pp 148–152
20. Duda O, Peter E, Hart DGS (eds) (2000) Pattern Classification. 2nd edn. Wiley, United States
21. Senawi A, Wei HL, Billings SA (2017) A new maximum relevance-minimum multicollinearity (MRmMC) method for feature selection and ranking. Pattern Recognit. 67: 47–61
22. Sharmin S, Shoyaib M, Ali AA, Khan MAH, Chae O (2019) Simultaneous featureselection and discretization based on mutual information. Pattern Recogn 91:162–174
23. Uddin, MP, Mamun, MA, Hossain, MA (2020) PCA-based feature reduction for hyperspectral remote sensing image classification. IETE Techn Rev 1–21
24. Ünlü R, Xanthopoulos P (2019) Estimating the number of clusters in a dataset via consensus clustering. Expert Syst Appl 125:33–39
25. Vashishth V, Chhabra A (2019) GMMR: a Gaussian mixture model based unsupervised machine learning approach for optimal routing in opportunistic IoT networks. Comput Commun 134:138–148
26. Xiao J, Lu J, Li X (2017) Davies bouldin index based hierarchical initialization *K*-means. Intell Data Anal 21(6):1327–1338